Statistical Learning from Shifting, Indirect, or Unseen Data: Efficient Algorithms and Theoretical Guarantees

Ronak Mehta

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Zaid Harchaoui, Chair
Armeen Taeb
Alex Luedtke

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical Learning from Shifting, Indirect, or Unseen Data: Efficient Algorithms and Theoretical Guarantees

Ronak Mehta

Chair of the Supervisory Committee:
Zaid Harchaoui
Department of Statistics

A fascinating phenomenon underlying statistical machine learning and artificial intelligence is "out-of-distribution" (or OOD) generalization. Data can (and in some settings, must) be used to draw inferences regarding probability distributions other than the one from which they were sampled. Understanding this mystery gives promise to statistical analyses that exhibit a degree of universality, such as clinical trials whose conclusions reflect many subpopulations or pre-defined image/text encodings that can be used to solve many classification tasks simultaneously. This dissertation tackles the theoretical and algorithmic challenges of designing methods that exhibit these modern notions of generalization.

Chapter 2 studies a learning framework called distributionally robust optimization (DRO), which promotes OOD by training models to optimize the worst-case expected loss achievable within a collection of possible training distributions. These maximum-type objectives present challenges for designing stochastic learning algorithms, as unbiased estimates of the gradient are not easily computed. We design an estimator equipped with a progressive bias (and variance) reduction scheme, for which the resulting algorithm is shown to have a linear convergence guarantee. Although our optimization results apply more generally to DRO problems, we focus attention on a subclass of objectives called spectral risk measures, which have appealing statistical and computational properties previously unexplored in machine

learning. We provide theoretical and practical guidance on selecting the various problem parameters, such as the collection of distributions over which to maximize. Finally, we present (among others) extensions to group DRO, a popular extension of the framework amenable to training neural network models.

Chapter 3 takes insights from the DRO application and pursues stochastic algorithms for a more general class of optimization problems, dubbed semilinear min-max problems. These objectives interpolate between the well-understood class of bilinear and relatively less-understood nonbilinear min-max problems, and have applications to problem classes such as convex minimization with functional constraints as special cases. We present the first complexity guarantees for this problem class, using a randomized algorithm with components inspired by the simulation literature (such as adaptive sampling of new data and adaptive averaging of historical data). We prove convergence guarantees in both convex and strongly convex settings with a fine-grained dependence on individual problem constants. The results yield complexity improvements in even specific cases, such as bilinearly coupled problems. We also provide a lower complexity bound on the performance of deterministic algorithms applied to the semilinear problem class.

Chapter 4 shifts focus from the implementation of large-scale learning algorithms to their output. We investigate predictive models that learn via a pre-training procedure with unlabeled data and can then make predictions for downstream classification tasks (without having seen any directly labeled training data from that task). This capability, known as zero-shot prediction, is made possible by three ingredients: 1) massive, carefully curated pre-training datasets, 2) "self-supervised" labels that allow models to learn universal features of structured data (e.g., images/text), and 3) the translation of downstream data into the format seen during pre-training using a technique called prompting. We analyze all three ingredients theoretically by establishing both the sample complexity and the limits of prompting in terms of simple distributional conditions. Inspired by this theory, we explore

variants on the pre-training objective and prompting strategies that show practical benefits such as improved zero-shot classification accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

Throughout my Ph.D., I have been surrounded by the best of the best. I often describe the experience as traveling from dojo to dojo, each time being exposed to a new field and donning the title of "beginner" again and again. Be it my mentors, collaborators, or friends, I could not have asked for a better community in which to learn and grow.

From my research advisor, Zaid Harchaoui, I observed and admired a relentless pursuit of excellence. Thank you for believing that I, too, could join this pursuit. I am now a person who finds joy, perhaps even comfort, in chasing the impossible. Zaid is cultivated in a large number of fields, and is possibly the only one in the world who can use "Rao-Blackwellization" and "large language models" in the same sentence. It is this eclectic blend of classical and cutting-edge disciplines that I originally gravitated toward. Taking pride in one's work, never cutting corners, and helping others unconditionally are only a few of the lifelong lessons I will carry with me from Zaid's guidance.

I experienced my first deep dive into the field of optimization with Vincent Roulet. Although we both share an interest in the theoretical aspects of the field, it was from him that I learned to be an experimentalist. From using mathematics to debug code to maximizing information gain with each new experiment, the art and science of numerical methods were extremely challenging for me when I started this journey. Thank you, Vincent, for patiently fostering what I now consider to be my most valuable skill. Beyond this, I also appreciate your perspective on the personal journey, which constitutes the hardest part of a Ph.D.

Having done research in optimization for some time, while I never dared to call myself an expert, I did not realize how far I was until I met Jelena Diakonikolas. When writing our journal paper together, I witnessed her willingness to leave no technical stone unturned,

while still making complicated ideas crystal clear for the reader. In trying to emulate these qualities during late nights in Madison, I finally started to develop my own intuition (and rejoiced at the sight of negative quadratic terms). Thank you, Jelena, for your patience and inspiration.

By working with Lang Liu, my understanding of statistical theory was completely transformed. The two skills of his that I admired the most are the ability to formalize the chaos of applied AI into compelling theoretical problems and the mathematical creativity to invent new tools to solve them. While I attempt to display the first skill in this dissertation, I am not sure I (or anyone, for that matter) can perform the second skill quite like Lang. Thank you, Lang, for spending your personal time even after your graduation to invest in my growth.

At the expense of sounding cliché, as I reflect on my papers over the years, the most successful ones were those that told the best stories. If the technical storytelling of this thesis is to be appreciated, it is really Krishna Pillutla who deserves the credit. For every paper, presentation, rebuttal, or poster, I always struggled with the subtleties and constant decision-making inherent to scientific communication. How should I invite my audience into my problem setting without overwhelming them? Should I surrender to a reviewer's interpretation of my work, or should I firmly push back on their comments? Luckily, life is made easier by simply imagining what Krishna would do in these situations. I must also thank Krishna for staying up until the morning hours during my first paper submissions.

The intellectual community at the University of Washington was a gift that kept on giving. I thank John Thickstun and Alec Greaves-Tunnell for inspiring me during my first year. I am only now realizing the magnitude of the time investments they made during their last year, such as John's generative models course or Alec's contribution to the machine learning course we taught with Zaid. These experiences drew me into the research that became this dissertation. I would also like to recognize and thank my committee members, Alex Luedtke

and Armeen Taeb, as well as my co-authors and collaborators: Azadeh Yazdan-Shahmorad, Noah Stannis, Eric Shea-Brown, Ali Shojaie, Soumik Pal, Ludwig Schmidt, Briana Abrahms, Kasim Rafiq, and Medha Agarwal. To my research peers, Alex Bank, Konstantin Golobokov, Mateus Piovezan Otto, and Facheng Yu, thank you for inspiring me with your enthusiasm and determination. Your futures are beyond bright, and I am looking forward to seeing what they bring.

Essential to the research in this document were two funding initiatives from the National Science Foundation: the Institute for Foundations of Data Science (IFDS) and the Institute for the Foundations of Machine Learning (IFML). Much of my work, which emphasizes theoretical understanding, would not be possible if it were not for these foundations-focused initiatives.

I express my gratitude for the administrative and technical staff of our department: Ellen Reynolds, Vickie Graybeal, Tracy Pham, Kristine Chan, Veronica Bae, and Asa Sourdiffe. I never took for granted your professionalism, efficiency, and care for the students, and have been saved by your emails more times than I should admit.

Looking a bit further into the past, this journey would not have even begun if not for my professors at Johns Hopkins University. I knew that applied mathematics and statistics would become a lifelong endeavor after taking Avanti Athreya's course. Donniell Fishkind took a chance on me and orchestrated my first-ever internship in data science. I inherited a sincere love of research from working with my first advisor, Joshua Vogelstein. Finally, from Carey Priebe, I learned that there is always an exciting lesson to be learned from even the simplest of statistical problems (independent Bernoulli trials would be his case in point).

On a more personal note, I was very fortunate to live in a home that was always filled with loving friends (whether they officially lived there or not). Certain periods of my Ph.D. could not have been traversed without a certain degree of obsession. Yet, I never had to earn this obsession at the price of solitude. Instead, my goals, lifestyle, humor, successes, and

failures were shared with others. Medha, Saksham, Jeff, and Brianna, thank you for your companionship. There is no me without you.

I am grateful for the friends made in my cohort. Shreya and James, thank you for accepting me as I was in the earliest days of the program. Jillian, thank you for teaching me what it means to be an adult and to stand up for myself. Apara, thank you for your thoughtfulness and reliability over the years. Alana, thank you for sharing with me the interests that no one else does. Finally, I thank Kayla for improving my confidence as a teacher and mentor for others. To you all, I look forward to the next time we have the chance to meet, wherever and whenever it may be.

My final acknowledgement is dedicated to my family. Firstly, to my cousins Aayush, Aditya, Abhishek, Akash, and Priya for traveling across the globe to attend my defense. Secondly, to my brother Kunal for not only supporting me, but also managing our trips, business, and many other responsibilities when I was too overwhelmed to do so myself. I will happily pay back this effort with interest in the coming years. Lastly, to my parents, Deepak and Nisha Mehta. They often like to quip that there must have been a mix-up at the hospital, as they wonder how this Ph.D. kid came from them. I, on the other hand, never saw a reason to wonder. Being kind, hardworking, and loyal like my father, and being disciplined and thoughtful like my mother, is all I ever really tried to do.

DEDICATION

To my parents

Chapter 1
INTRODUCTION

There has never been a more exciting time to work at the intersection of statistical machine learning (ML) and artificial intelligence (AI), as developments in modern AI challenge and redefine the conventional view of "generalization" in statistics. To understand this phenomenon, consider the example of a scientist pursuing knowledge discovery or an ML engineer building a new product. While their goals differ, their processes bear significant similarity; each will frame their goal mathematically using a quantitative *parameter* within a statistical model, and may consider collecting data for the express purpose of learning or inferring this parameter. This reflects Ronald Fisher's famous Rothamsted experiment in agriculture [Parolini, 2015] or the introduction of CIFAR-10 to study compositional features in image classification [Krizhevsky, 2009]. In both cases, the practitioner seeks the ability to *generalize*, or to use their sample or training set, respectively, to achieve an understanding of possibly unseen data. While this *modus operandi* has been tremendously successful for designing the theory and methods of contemporary statistics, it is increasingly pressing for scientists, engineers, and researchers to use data to learn about populations (or data-generating distributions) *other than* the one from which they were drawn.

Why would such a need for *out-of-distribution (OOD) generalization* arise?

- **Accountability:** Randomized control trials (RCTs) have been the gold standard experimental design for decades. However, when certain interventions are unethical (e.g., exposure to carcinogens) or conducting such a trial is prohibitively expensive, observational data may be used as an alternative option. Managing the discrepancy between interventional and observational data led to the development of the propensity score [Rosenbaum and Rubin, 1983], a mainstay of causal inference.

- **Safety:** AI models are now deployed in critical domain applications such as energy planning [Guigues and Sagastizábal, 2013], materials engineering [Yeh, 2006], and financial regulation [He et al., 2022]. When the deployment environment differs drastically from the training environment, brittle models may lead to catastrophic outcomes such as misdiagnoses, bankruptcies, and mechanical failures. A natural question is whether such shifts in distribution can be simulated or accounted for during the training phase itself.

- **Fairness:** A well-documented phenomenon is the tendency of AI models, such as facial recognition systems, to perform well on majority subgroups in the evaluation data (such as male or lighter-skinned individuals) and exhibit social biases on minority subgroups [Buolamwini and Gebru, 2018]. It is of clear interest whether models can be developed to promote equitable behavior, even if minority subgroups are scarce in the training population, which constitutes one popular notion of algorithmic fairness. As we elaborate on in Chapter 2, fairness by this definition can also be framed as a distribution shift problem, in which minority subgroups appear with higher probability in the evaluation data.

- **Efficiency:** Finally, high-performance models such as large language models (LLMs) may be trained on trillions of tokens and thousands of GPUs in order to be viable for applications such as AI chat assistants [Hoffmann et al., 2022]. Such a magnitude of resources is not accessible to the average individual, motivating research into the reuse of institutionally-trained models with little to no training data from the target distribution provided by the scientist. For this setting in particular, not only may the evaluation data change, but the task itself may change as well, introducing additional complexity.

We argue that in statistics and machine learning, out-of-distribution generalization is the rule rather than the exception. This dissertation contributes to the theory and methods

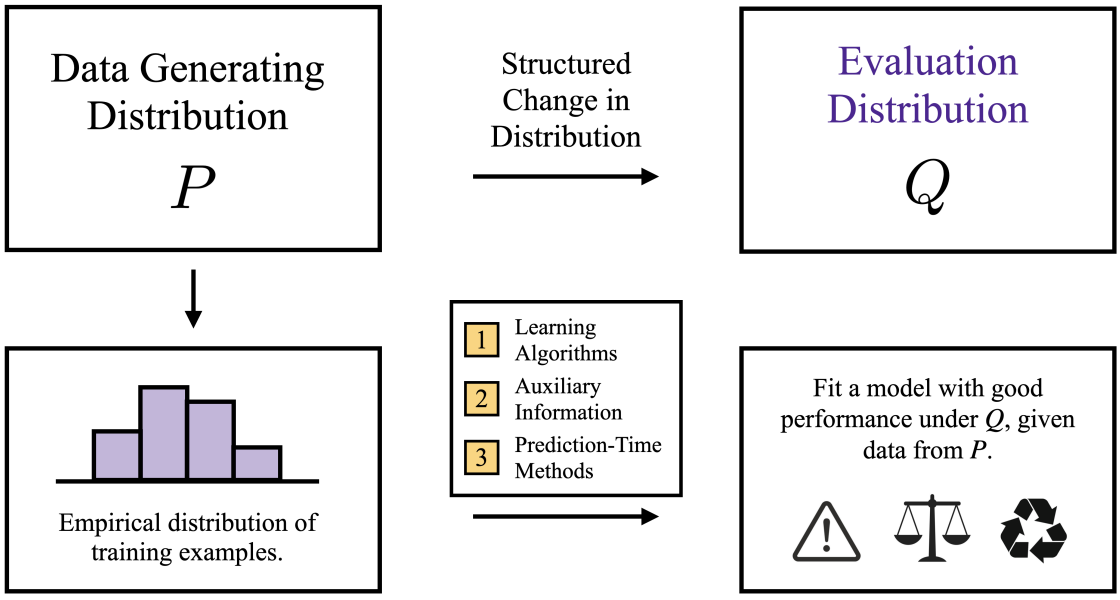

Figure 1.1: **Illustration of Out-of-Distribution Generalization.** While traditional statistical learning paradigms would deal only with the data-generating distribution P during evaluation/deployment, this dissertation considers cases in which the evaluation distribution Q may be different. Often, the change/shift from P to Q exhibits additional structure (discussed, e.g., in Section 2.8). Techniques for handling this shift fall into alternative learning algorithms (Chapter 2 and Chapter 3), usage of auxiliary/side information Section 4.3, or methods of querying pre-trained models (Section 4.4).

of OOD generalization by 1) retrospectively analyzing existing learning algorithms and 2) prospectively developing new ones (see Figure 1.1 for an illustration). In the remainder of this chapter, Section 1.1 introduces existing lines of research at a high-level, Section 1.2 collects technical details necessary to understand the main contributions, and Section 1.3 provides an overview of these contributions and an outline of the document.

1.1 Context and Motivation

1.1.1 Alternative Learning Objectives

Parameters are fit to data by optimizing some performance measure called a *loss function*, and while many losses (such as the squared, hinge, or cross-entropy loss) have been explored

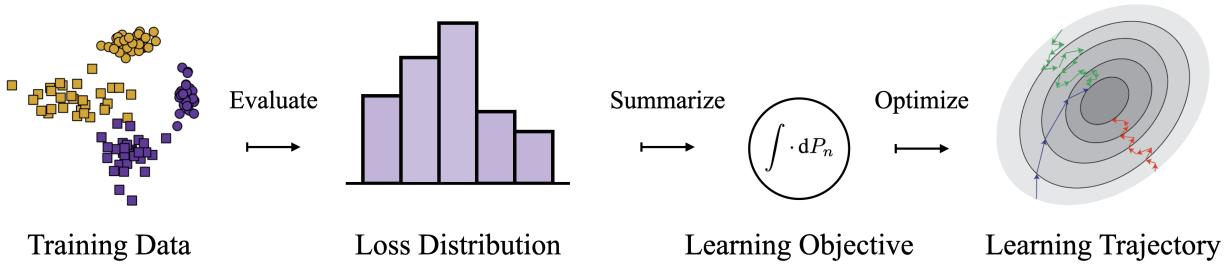

Figure 1.2: **Illustration of Learning Pipeline.** In the illustration above, a predictive model is evaluated on n training examples to produce the histogram of errors (or loss distribution). This loss function is summarized by a single quantity, perhaps by integrating over the empirical distribution P_n, after which this quantity is minimized by invoking an optimization algorithm.

in the literature, one aspect of these objectives has remained consistent: that the loss is averaged over the training data to produce the final summary. Upon reflection, the usage of the simple average is a choice in and of itself, and we may consider other ways of aggregating the loss into a univarite quantity to be optimized (see Figure 1.2). In applications such as finance, alternative *risk measures*, or summaries of the loss distribution, have been explored for decades [He et al., 2022]. Two highly related questions of increasing interest are whether these risk measures can be used to learn models that perform well on (1) the tails of the loss distribution and (2) on other data-generating distributions without including any additional training data.

1.1.2 Data Selection over Model Selection

Statistics pedagogy often describes data in an oracle framework, that is, the analysis starts after a number of independent and identically distributed (i.i.d.) data points are drawn from a probability distribution and supplied to the researcher. In both scientific and industrial applications, the researcher may often have the agency to both influence the data-generating mechanism *and* conduct the analysis simultaneously. While organizations such as Google, Meta, and OpenAI have access to essentially the entire Internet, it is now recognized that

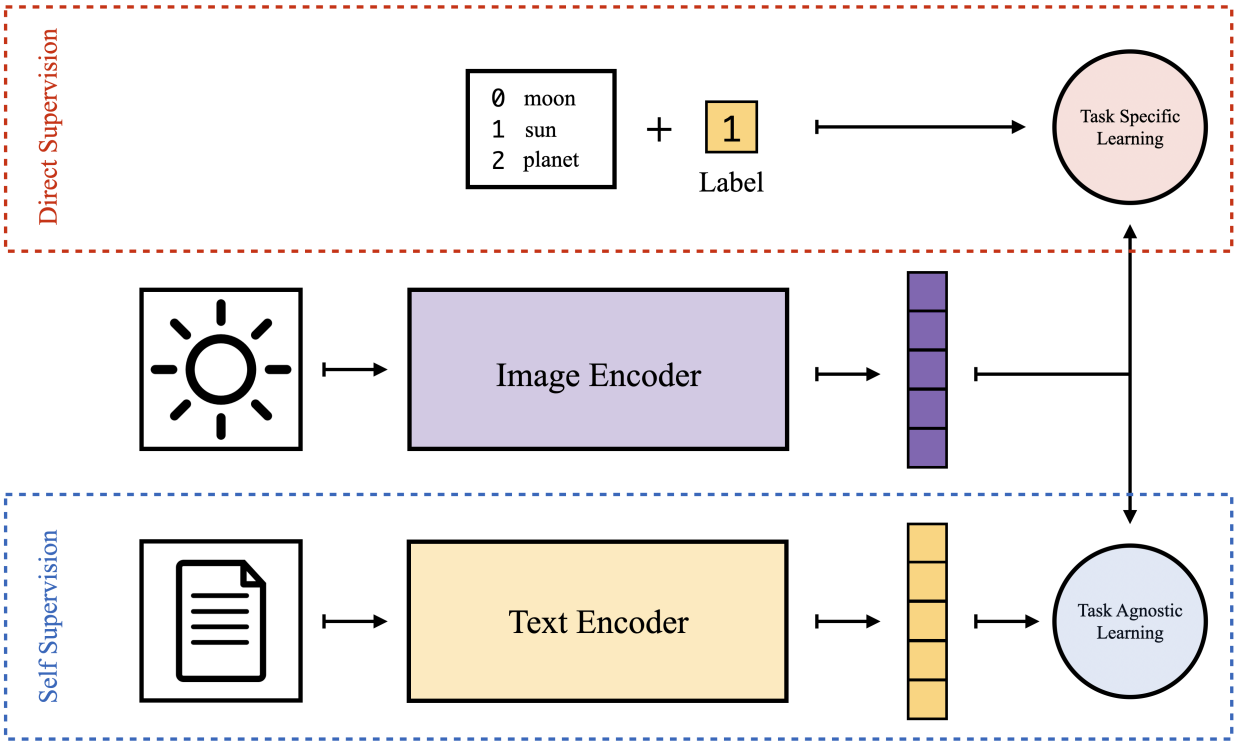

Figure 1.3: **Foundation Model Pre-Training versus Direct Supervision.** The purple encoder is optimized to produce an informative representation of input images. In the traditional supervision framework (top), labeled training data is provided so that the encoder produces representations that are informative for the particular task of predicting the given label. Labeled data, often produced using high-quality annotators, is relatively scarce compared to the number of images that are unlabeled. Because unlabeled images may still have accompanying captions, in modern frameworks such as self-supervision, the image encoder may be (pre-)trained to be predictive of its caption (and vice versa).

a large quantity of data does not produce models that generalize well if much of it is low-quality [Gadre et al., 2023, Li et al., 2024a]. In other words, one may also ask which *data* promotes OOD generalization in addition to which *models*.

1.1.3 Universal Representations of Structured Data

An emerging phenomenon is the advent of *foundation models*, or pre-trained neural networks whose internal layers can be used as fixed or tunable feature mappings for structured data

such as images, text, or audio (see Devlin et al. [2019b], Chen et al. [2020], Bardes et al. [2022], Oquab et al. [2024] and references therein). These reusable feature representations have had a major impact in settings such as data-scarce image or text classification, as they may simplify these problems to simple linear or logistic regression on pre-trained features. Even more surprisingly, they can be used for zero-shot prediction, or generating classifiers without *any* additional training data. This capability has evolved the notion of generalization even further: instead of generalizing to different distributions over the same sample space, these models can now generalize to different input and output spaces entirely. The pre-training process for an image embedding foundation model is shown in Figure 1.3 and compared to the classical notion of (direct) supervised learning. We describe the usage of these models for zero-shot prediction in the upcoming Section 1.2.3. While this technique has forged ahead from the applied perspective [Radford et al., 2021, Pratt et al., 2023, Xu et al., 2024], we also hope to achieve an improved theoretical understanding of it in Chapter 4 of this dissertation.

1.2 Technical Overview

This dissertation investigates several statistical and computational aspects of these recent perspectives on generalization, from linear models for tabular data to foundation models for images and natural language. Our study is unified by developing and understanding stochastic algorithms for large-scale optimization problems that invariably arise across settings.

1.2.1 Empirical Risk Minimization

We center ourselves in the eminent framework of *empirical risk minimization (ERM)*. Consider a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $(\Omega, \mathcal{F})$ is a measurable space and $\mathbb{P}$ is a probability measure. Consider a statistical model $(\Xi, \mathcal{P})$, where $\mathcal{P}$ contains probability measures over Ξ, the space of observable data. Then, let $\xi : \Omega \to \Xi$ be a random variable that is governed by an unknown probability measure $P \in \mathcal{P}$. Given a parameter space Θ

and *risk functional* $\mathcal{R} : \Theta \times \mathcal{P} \to \mathbb{R}$, we wish to determine a parameter of the form

$$\boldsymbol{\theta}_0 \in \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}(\boldsymbol{\theta}, P), \tag{1.1}$$

assuming such a minimizer exists. Although P is unknown, we assume access to i.i.d. samples $\xi_1, \ldots, \xi_n \sim P$, often referred to as the *training set*, with associated empirical measure $P_n := \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$. The ERM principle defines an estimator $\hat{\boldsymbol{\theta}}_n$ via

$$\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}(\boldsymbol{\theta}, P_n). \tag{1.2}$$

A familiar example of (1.2) is maximum likelihood estimation, wherein Θ parametrizes a class of probability measures $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, and $\mathcal{R}(\boldsymbol{\theta}, P_n)$ is the empirical Kullback-Liebler risk between $P_{\boldsymbol{\theta}}$ and P_n. More generally, the user may define a measurable *loss* $\ell : \Theta \times \Xi \to \mathbb{R}$ and define

$$\mathcal{R}(\boldsymbol{\theta}, P) := \mathbb{E}_{\xi \sim P} \left[\ell(\boldsymbol{\theta}, \xi) \right] \implies \mathcal{R}(\boldsymbol{\theta}, P_n) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, \xi_i), \tag{1.3}$$

where $\mathbb{E}[\cdot]$ denotes the expectation functional. While many works may consider the "average loss" format of (1.3) to be synonymous with empirical risk minimization, we consider more general functionals throughout this thesis. The first of two highly relevant examples throughout the thesis is the functional

$$\mathcal{R}(\boldsymbol{\theta}, P) = \max_{Q \in \mathcal{Q}(P)} \mathbb{E}_Q \left[\ell(\boldsymbol{\theta}, \xi) \right] \tag{1.4}$$

where $\mathcal{Q}(P)$ is a to-be-specified set of probability measures on Ξ related to P, called the *uncertainty set*. By computing the worst-case expectation over many probability measures, this objective naturally promotes aversion to risk and upweights the "harder" examples in the training set. Letting $P^{\otimes b}$ denote the probability measure governing b i.i.d. draws from P, the second prominent example is

$$\mathcal{R}(\boldsymbol{\theta}, P) = \mathbb{E}_{P^{\otimes b}} \left[\ell_b(\boldsymbol{\theta}, \xi_1, \ldots, \xi_b) \right], \tag{1.5}$$

where b denotes a batch size and $\ell_b : \Theta \times \Xi^b \to \mathbb{R}$ is a loss function (dependent on this batch size). This type of objective is often applied in unsupervised problems, where structure within and between data points (similarity, etc.) acts as a replacement for labels. Notably, both (1.4) and (1.5) involve relationships between data points and are not expressible simply as average losses across the observed sample. This is a challenge from both statistical and optimization perspectives, and will guide much of the work in forthcoming sections. We give a brief overview of our topics of study and defer specific details to the individual chapters.

1.2.2 Distribution Shift and Distributional Robustness

A natural application of the risk functional (1.4) is when the practitioner observes training data from P, but expects the estimated parameter to be evaluated on test data from another probability measure Q. For instance, a training and test set of images may differ in distribution due to heterogeneous lighting conditions (a natural shift) or corruption of the test images through blurring (a synthetic shift). While this phenomenon—commonly known as *distribution shift* [Quiñonero-Candela et al., 2022] in the literature—is often treated as a specialized problem setting, we emphasize that virtually all machine learning models are deployed on data that are not distributed identically to the training data. In the empirical setting, by carefully designing the uncertainty set $\mathcal{Q}(P_n)$, one hopes that the parameter is *distributionally robust*, or will have approximately uniform performance across many probability measures that are sufficiently close to P. This framework, aptly named *distributionally robust optimization (DRO)*, differs from classical work in distribution shift [Huang et al., 2006, Sugiyama et al., 2007, 2008], in that the user has no information about the particular distribution Q on which the model will be evaluated (such as unlabeled covariates). Thus, $\mathcal{Q}(\cdot)$ is often defined using qualitative conditions such as existence and boundedness of a likelihood ratio $\frac{\mathrm{d}Q}{\mathrm{d}P}$ (see Chapter 2).

Under various assumptions on $\mathcal{Q}$, the maximization problem over probability measures

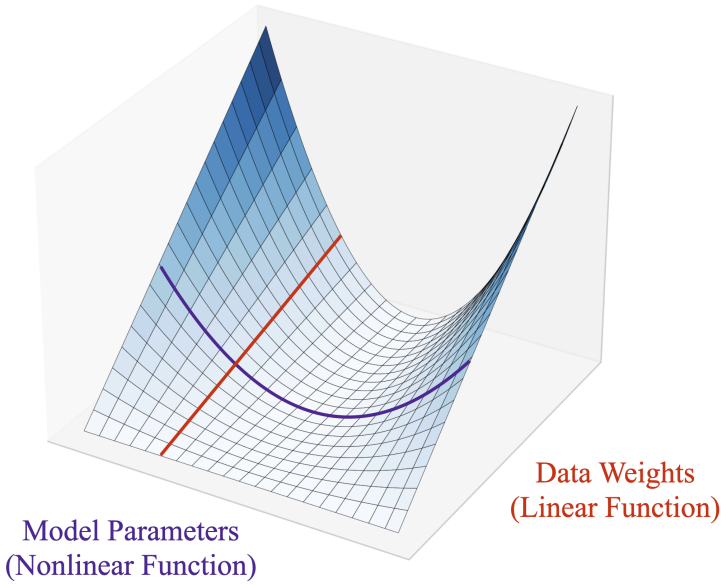

Figure 1.4: **Linear-Nonlinearly Coupled Objective.** Illustration of the objective (1.6), which is possibly nonlinear in $\boldsymbol{\theta}$ (the model parameters) but linear in the data weights $\boldsymbol{q}$.

in $\mathcal{Q}(P_n)$ can often be made into a finite-dimensional program of the form

$$\max_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_Q \left[\ell(\boldsymbol{\theta}, \xi) \right] = \max_{\boldsymbol{q} \in \mathcal{Q}} \sum_{i=1}^{n} q_i \ell(\boldsymbol{\theta}, \xi_i). \tag{1.6}$$

where $\mathcal{Q} \subseteq \mathbb{R}^n$ contains weight vectors $\boldsymbol{q} = (q_1, \ldots, q_n)$ on each training example. However, as we discuss in Chapter 2, this optimization problem will often be intractible for large values of n, and typical stochastic gradient-type algorithms will face challenges that are unseen when minimizing sample averages. Statistical questions include how to construct (sub)gradient estimators for these maximum-type objectives with low bias and variance, and how such objectives behave as n grows (noting that the dimension of the set $\mathcal{Q}$ grows linearly in n). Furthermore, (1.6) generates a min-max problem in which the objective is nonlinear in the minimizing (or primal) variables $\boldsymbol{\theta}$ but is in fact linear in the maximizing (or dual) variables $\boldsymbol{q}$ (see Figure 1.4). We study this "dual-linear" coupling extensively in Chapter 3.

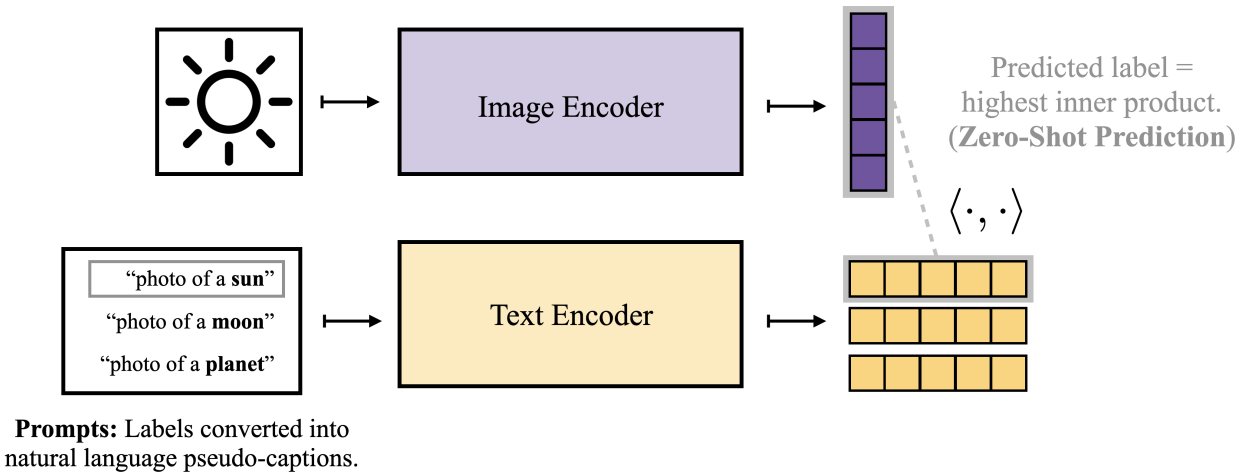

Figure 1.5: **Prompting of Foundation Models.** Illustration of the indirect predictor (1.7), which relies on image and text encoders pre-trained via self-supervision (see Figure 1.3)

1.2.3 Universal Representations and Zero-Shot Prediction

The risk functional (1.5) is based on the objective used to train the CLIP series of neural network models [Radford et al., 2021] and its predecessors [Chen et al., 2020]. Here, we have that $\Xi = \mathcal{X} \times \mathcal{Z}$, where $\mathcal{X}$ and $\mathcal{Z}$ denote observation spaces of two data modalities (most commonly images and text). Then, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, a pair of functions $\boldsymbol{\alpha} : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{\beta} : \mathcal{Z} \to \mathbb{R}^d$ which are optimized so that $\boldsymbol{\alpha}(\boldsymbol{x})$ and $\boldsymbol{\beta}(\boldsymbol{z})$ are close in $\mathbb{R}^d$ if and only if $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$ are semantically similar. Remarkably, these encoders can be used to perform downstream prediction tasks without *any direct labeled data*—generalization of this kind is known as *zero-shot prediction*.

To understand zero-shot prediction, we first contrast it with the related setting of few-shot learning. Let $\boldsymbol{x} \in \mathcal{X}$ be an input that accompanies a label $\boldsymbol{y} \in \mathcal{Y}$, which could be a class label or even a structured object such as a parse tree. Common to both zero-shot prediction and few-shot learning is a pre-training procedure in which a large unlabeled dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ is used to produce $\boldsymbol{\alpha}$. Pre-training typically occurs through the process of *self-supervised learning (SSL)*, using a pretext task (quantified by an objective

like (1.5)) that can be solved with only instances of $\boldsymbol{x}$ (e.g., filling in a blank image patch). In few-shot learning, the user may then access a labeled dataset $(\boldsymbol{x}_1^{\text{lab}}, \boldsymbol{y}_1^{\text{lab}}), \ldots, (\boldsymbol{x}_{n^{\text{lab}}}^{\text{lab}}, \boldsymbol{y}_{n^{\text{lab}}}^{\text{lab}})$ from which a predictor can be trained inexpensively. This often takes the form of a linear classifier $\boldsymbol{x} \mapsto \mathbf{W}\boldsymbol{\alpha}(\boldsymbol{x}) + \boldsymbol{b}$ for $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{Y}|}$. Because such data is not available for zero-shot prediction, the ingenuity of practitioners has yielded the following solution: if (1) each pre-training example $\boldsymbol{x}_i$ (a web image, say) is paired with another *view* $\boldsymbol{z}_i \in \mathcal{Z}$ (e.g., a caption in natural language) and (2) if each label $\boldsymbol{y} \in \mathcal{Y}$ can intelligently be embedded into $\mathcal{Z}$, then the relationship between each $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ could provide the means to perform prediction. Concretely, one learns the complementary encoder $\boldsymbol{\beta}$ during pre-training and designs a number of *prompts* $\boldsymbol{z}_j^{\boldsymbol{y}}$ for $\boldsymbol{y} \in \mathcal{Y}$ and $j = 1, \ldots, M$. Then, the function

$$\boldsymbol{x} \mapsto \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \frac{1}{M} \sum_{j=1}^{M} \langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{z}_j^{\boldsymbol{y}}) \rangle \tag{1.7}$$

is employed for prediction. An example of a prompt is the template text "photo of a ___.", where the blank can be filled by the textual representation of the class (e.g., "cat" or "dog"). The entire procedure is depicted in Figure 1.5.

The zero-shot prediction pipeline, from pre-training to prompt selection, is clearly a wild departure from what is explained by statistical learning theory. Moreover, while some components of these systems have been studied in the context of few-shot learning (such as the reasons why various pre-training objectives result in encoders that provably accelerate learning), unique aspects of zero-shot prediction, such as the role of prompting and the cost of "translating" modalities, have not yet received theoretical treatment. This is the subject of Chapter 4.

1.3 Contributions and Outline

We study these problems of interest across the next three chapters of the dissertation. We conclude in the final chapter with a summary and discussion. Before stating the contributions, we briefly comment on some work that is not included in this dissertation: 1) a

collaborative work on multivariate time series prediction with graph neural networks (featured at the KDD 2023 Workshop on Mining and Learning with Graphs) conducted during an industry internship [Yang et al., 2023], 2) a collaborative work on using black-box prediction sets to classify animal behavior from accelerometry data with uncertainty quantification [Agarwal et al., 2024], 3) a collaborative work studying the relationship of neuropsychiatric symptoms such as depression with Alzheimer's disease (under review at *Alzheimer's & Dementia*), and 4) a work for which the dissertation author is the primary contributor that develops stochastic algorithms for solving independent component analysis problems with supervision (under review at ICASSP 2026).

Practical Algorithms for Distributionally Robust Optimization In Chapter 2, we provide scalable, stochastic algorithms for solving distributionally robust optimization (DRO) problems. We prove the first theoretical linear convergence rate for stochastic algorithms on regularized DRO problems and handle several practical questions of interest. These include duality results over the probability simplex that not only allow for the efficient implementation of the algorithm but also give guidance on choosing hyperparameters such as the uncertainty set. We also specialize our results to specific objectives (studied previously in quantitative finance) known as spectral risk measures (SRMs), which we prove can be represented as DRO objectives. We derive several appealing properties of SRMs, including bias bounds for sample estimators under general conditions and computational properties that allow for simpler implementations than previous DRO approaches. The subject of the chapter is not only the theoretical aspects, but also extensive experimentation across tabular, image, and natural language examples. In particular, we test the proposed algorithms in group distribution shift/fairness scenarios and demonstrate performance benefits in terms of worst-case group-wise error.

This chapter is joint work with Vincent Roulet, Krishna Pillutla, Lang Liu, and Zaid Harchaoui. The results span two papers, Mehta et al. [2023] published at AISTATS 2023 and Mehta et al. [2024b] published at ICLR 2024. The latter was accepted as a Spotlight

(top 5% of submissions).

Complexity Guarantees for Semilinear Min-Max Optimization In Chapter 3, we consider general saddle point optimization problems (a.k.a. min-max problems) where the coupled term in the objective is linear in at least one of either the primal or dual variables. This "semilinear" min-max problem includes not only distributionally robust optimization as a special case but also other classes such as convex minimization with functional constraints. We provide a constructive upper bound on the complexity guarantees for controlling the primal-dual gap criterion, in that the algorithm is derived by way of the analysis (contrary to the experimentally-driven methods of Chapter 2). The method improves upon classical approaches to nonbilinearly-coupled min-max problems such as extragradient [Korpelevich, 1976, Nemirovski, 2004], Popov's method [Popov., 1980], or dual extrapolation [Nesterov, 2007a] by delicately using the non-uniformity Lipschitz and smoothness constants of the nonlinear components of the objective. We also present specialized algorithms for when the linear variable has a rectangular feasible set, i.e., it can be optimized along its coordinates, which yields improved complexities in even the bilinear setting.

This chapter is joint work with Jelena Diakonikolas and Zaid Harchaoui. A portion of the ideas in this work were published in our NeurIPS 2024 paper Mehta et al. [2024a]. They are generalized and improved significantly in the chapter, and a journal manuscript is under review [Mehta et al., 2025].

Generalization Bounds for Data Curation and Zero-Shot Prediction Chapter 4 targets questions such as model reuse and self-supervised learning mentioned in Section 1.2.3. We study the estimation of a joint probability distribution over multimodal data (or a linear functional thereof) when given knowledge of the true marginal distributions of each modality. This problem is motivated by the prominent practice in large-scale machine learning to rebalance a pre-training dataset in order to achieve a particular marginal distribution over subgroups of variables (e.g., text data with equal representation across languages). Con-

sidering the estimation of a linear functional of the data-generating distributions, we prove a non-asymptotic bound on the mean squared error when incorporating the marginals into the estimator. Specifically, a balancing procedure known in statistical contexts as raking ratio estimation, iterative or bi-proportional fitting, or the Sinkhorn algorithm is used to alter the empirical measure into one that satisfies the marginal constraints. The bound is then proven using a recursive formula for iterations of the procedure, which furnishes both a first-order term (which improves upon the variance of the empirical mean estimator) and explicit higher-order terms. Furthermore, when the iteration number is scaled appropriately, we recover the efficient asymptotic variance that was derived in Bickel et al. [1991] using tools of (asymptotic) semiparametric efficiency theory. Our approach, based on the recursive formula, yields a new closed-form expression for this first-order term (which was previously stated variationally via projections) in terms of the conditional mean operator between the modalities. The second part of Chapter 4 casts a zero-shot prediction procedure that is used ubiquitously in modern practice (from pre-training to prompting) as a formal estimator of a function defined on the population distribution. We establish performance limits of zero-shot prediction as compared to the Bayes optimal predictor on downstream tasks. Both theoretical analyses are demonstrated with experiments on language-image pre-training and zero-shot image classification.

This chapter is joint work with Lang Liu and Zaid Harchaoui. It contains some theoretical and experimental results on data curation from our NeurIPS 2024 paper [Liu et al., 2024] and results on prompting and downstream prediction from our ICML 2025 paper [Mehta and Harchaoui, 2025]. The latter was accepted as an Oral (top 1% of submissions). Part of this work was conceptualized and developed at the Simons Institute for the Theory of Computing as part of the Modern Paradigms in Generalization (2024) program. The content of previous chapters also benefited greatly from the fruitful scientific interactions at this program.

Software The papers mentioned above accompany software packages to reproduce the code and experiments. These include `lerm` [Mehta et al., 2023], `prospect` [Mehta et al.,

2024b], `drago` [Mehta et al., 2024a], `balancing` [Liu et al., 2024], and `zeroshot` [Mehta and Harchaoui, 2025]. Pip-installable packages called `deshift` and `drlearn` collect versions of the DRO algorithms developed in this dissertation for PyTorch and scikit-learn workflows, respectively. All packages are linked at https://ronakdm.github.io/software.

Chapter 2

LEARNING UNDER DISTRIBUTION SHIFT WITH LIKELIHOOD RATIOS

2.1 Introduction

In Chapter 1, we introduced alternate notions of risk functionals by which we may define parameters of interest. One such risk, which was (1.4), takes the maximum expected loss achievable over an uncertainty of distributions. When evaluated at an empirical measure P_n, this yields a variant of empirical risk minimization known as *distributionally robust optimization (DRO)*. This chapter is dedicated to studying the properties of such objectives and designing scalable algorithms for optimizing them in practice.

While particular cases of the uncertainty set $\mathcal{Q}(P_n)$ (e.g., a closed ball in f-divergence or Wasserstein distance) have been studied extensively in DRO [Namkoong and Duchi, 2016, Shapiro, 2017, Kuhn et al., 2019, Duchi and Namkoong, 2021], we begin at a broader starting point: placing assumptions on $\mathcal{Q}(\cdot)$ that ensure that the maximization over probability measures can be reformulated into a finite-dimensional program of the form

$$\max_{\boldsymbol{q} \in \mathcal{Q}} \sum_{i=1}^{n} q_i \ell(\boldsymbol{\theta}, \xi_i), \tag{2.1}$$

where, as in (1.6), $\mathcal{Q} \subseteq \Delta^{n-1} := \{(p_1, \ldots, p_n) : \sum_{i=1}^n p_i = 1, p_i \geq 0 \ \forall i\}$ contains probability mass vectors. Here, we will call $\boldsymbol{\theta}$ the primal variable and $\boldsymbol{q}$ the dual variable. Empirical risk minimization (ERM) is recovered by (2.1) when $\mathcal{Q} = \{\mathbf{1}/n\}$ for $\mathbf{1} = (1, \ldots, 1)$ (i.e., there is only one feasible dual variable). We call the minimization of (2.1) a *likelihood ratio-based distributionally robust optimization (LR-DRO)* problem, as the finiteness will result from considering $\mathcal{Q}(P_n)$ to contain only distributions that are absolutely continuous with respect to P_n (see Section 2.2 for details). A canonical example of an LR-DRO problem is f-DRO,

in which we define $\mathcal{Q}(P_n)$ as a closed ball about P_n in f-divergence. There is also another class of risk functionals known as *spectral risk measures (SRMs)*, which are lesser-known as DRO objectives in machine learning.

Despite the increasing adoption of objectives of the form (2.1), optimization approaches have relied on using the full-batch or stochastic subgradient method out-of-the-box [Fan et al., 2017, Kawaguchi and Lu, 2020, Laguel et al., 2020, Levy et al., 2020], both enduring considerable limitations. The per-iteration complexity of full-batch methods (those that invoke every first-order oracle $\{\ell(\cdot, \xi_i), \nabla\ell(\cdot, \xi_i)\}_{i=1}^n$) is $\tilde{O}(n)$ function/gradient evaluations. For stochastic[1] variants, unbiased estimates of any subgradient, while needing only $O(1)$ gradient evaluations, still need $O(n)$ function calls to compute the vector of losses $(\ell(\boldsymbol{\theta}, \xi_1), \ldots, \ell(\boldsymbol{\theta}, \xi_n))$, as we describe in Section 2.4. This yields the same per-iteration complexity as the full-batch method in automatic differentiation frameworks. A number of methods abandon convergence to the minimal risk altogether and resort to $O(1)$-time stochastic subgradient updates, but are biased [Kawaguchi and Lu, 2020, Levy et al., 2020], i.e., do *not* converge to the optimal value. In this chapter, we present a class of stochastic algorithms that enjoy a theoretical *linear* convergence guarantee to the solution of a regularized LR-DRO problem while needing only a constant number of calls to the first-order oracles per iteration and have excellent empirical performance compared to baselines. We also provide theoretical and practical guidance on selecting the uncertainty set and other parameters defining the objective. Despite the min-max structure, we adopt a single-hyperparameter primal-only viewpoint to emphasize the importance of practicality, in the spirit of variance-reduced stochastic algorithms used for finite sum objectives [Johnson and Zhang, 2013, Defazio et al., 2014].

The remainder of the chapter is outlined as follows. Section 2.2 introduces likelihood ratio-based DRO and spectral risk measures. Section 2.3 establishes statistical properties such as an upper bound on the bias incurred by stochastic estimates of SRMs. For vanilla

[1]We use the term "stochastic" to include both *streaming* algorithms in which fresh samples from the data-generating distribution are provided at each iterate, and *incremental* algorithms, in which multiple passes are made over a fixed dataset.

stochastic gradient-type algorithms that may only converge to a ball about the optimum, these bias bounds allow us to determine the worst-case radius of this ball. In Section 2.4 and Section 2.5, we present the details of our modified form of the objective (2.5) and the proposed algorithm, with its convergence analysis given in Section 2.6. In Section 2.7 we answer questions of practical performance, such as how to select the uncertainty set and how to efficiently solve the associated maximization problem. Important connections to related fields such as empirical likelihood [Owen, 1990] are given in Section 2.8. Numerical benchmarks on regression and classification problems are given in Section 2.9. Finally, we consider several extensions of the algorithms and analyses in Section 2.11 with concluding remarks in Section 2.12.

2.2 Preliminaries

We first characterize LR-DRO as categorization of various DRO objectives, and then introduce the two example objectives used extensively throughout the entire chapter: 1) closed balls in f-divergence (i.e., f-DRO) and 2) spectral risk measures (SRMs). On terminology, we sometimes identify DRO objectives with their uncertainty set, as we have done with f-DRO above. Some results (such as the duality relations in Section 2.7) will be established separately for f-DRO and SRMs. Furthermore, because SRMs are relatively new to machine learning, we dedicate Section 2.3 to some novel results deriving their properties.

In the LR-DRO class, the primary assumption on $\mathcal{Q}(P)$ for any probability measure P over Ξ is that for every $Q \in \mathcal{Q}(P)$, it holds that $Q \ll P$ (i.e., Q is absolutely continuous with respect to P). By narrowing the options for the perturbed distributions in this fashion, any $Q \in \mathcal{Q}(P)$ is exactly specified by the *likelihood ratio*, or Radon-Nikodym derivative $\frac{\mathrm{d}Q}{\mathrm{d}P} : \Xi \to [0, \infty)$, that is,

$$\mathcal{R}(\boldsymbol{\theta}, P) = \max_{Q \in \mathcal{Q}(P)} \mathbb{E}_Q \left[\ell(\boldsymbol{\theta}, \xi)\right] = \max_{\beta \in \mathcal{B}(P)} \mathbb{E}_P \left[\beta(\xi)\ell(\boldsymbol{\theta}, \xi)\right], \tag{2.2}$$

where $\mathcal{B}(P) := \left\{\frac{\mathrm{d}Q}{\mathrm{d}P} : Q \in \mathcal{Q}(P)\right\}$. The set $\mathcal{B}(P)$ can often be specified by explicitly stating conditions on the likelihood ratio β instead of the probability measure Q, as seen in

the following examples. Below, let $\mathcal{B}_0(P)$ denote the collection of valid likelihood ratios for probability measures under P, or the non-negative measurable functions $\beta(\cdot)$ satisfying $\mathbb{E}_P[\beta(\xi)] = 1$.

Example 2.2.1 (f-DRO)**.** In the case of f-DRO, we have that

$$\mathcal{Q}(P) := \{Q : Q \ll P \text{ and } D_f(Q\|P) \le \rho\},$$

where $D_f(Q\|P) = \int_\Xi f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q$ is an f-divergence generated by f and ρ is a radius parameter (see Appendix A.1.2 for a review of f-divergences). Then, associated with the f-ball uncertainty set is the feasible set

$$\mathcal{B}(P) = \{\beta \in \mathcal{B}_0(P) : \mathbb{E}_P[f(\beta(\xi))] \le \rho\}. \tag{2.3}$$

In words, f-DRO places a moment condition on β, which is much weaker than the assumptions placed on spectral risk measures.

Example 2.2.2 (Spectral Risk Measures)**.** For a bounded, measurable, non-decreasing function $s : (0,1) \to [0,\infty)$ satisfying $\int_0^1 s(t)\,\mathrm{d}t = 1$ (called the *spectrum*) and a real-valued random variable X with cumulative distribution function (CDF) F, we define the functional $\mathbb{L}_s$ as

$$\mathbb{L}_s[X] = \int_0^1 s(t)F^{-1}(t)\,\mathrm{d}t$$

where $F^{-1}(p) := \inf\{x : f(x) \ge p\}$ is the right generalized inverse CDF or *quantile function* of X. By Hölder's inequality, the quantity above is well-defined when $\mathbb{E}_F|X| < \infty$. We call the function $\mathbb{L}_s$ the *spectral risk measure (SRM)* with spectrum s. In the notation of (2.2), the learning objective associated with the SRM is then

$$\mathcal{R}(\boldsymbol{\theta}, P) := \mathbb{L}_s[\ell(\boldsymbol{\theta}, \xi)] \text{ for } \xi \sim P.$$

It remains to show that there exists an uncertainty set $\mathcal{Q}$ such that

$$\mathbb{L}_s[\ell(\boldsymbol{\theta}, \xi)] = \max_{Q \in \mathcal{Q}(P)} \mathbb{E}_Q \left[\ell(\boldsymbol{\theta}, \xi)\right]$$

holds. This is established in the upcoming Proposition 2.3.1 from Section 2.3, where we prove that $\mathcal{Q}(P)$ corresponds to all all Q such that $\frac{\mathrm{d}Q}{\mathrm{d}P}$ follows the same distribution (under P) as $s(U)$ for $U \sim \mathrm{Unif}(0,1)$. Thus, we immediately have that

$$\mathcal{B}(P) = \left\{\beta \in \mathcal{B}_0(P) : \beta(\xi) \overset{\mathrm{d}}{=} s(U). \text{ for } \xi \sim P, U \sim \mathrm{Unif}(0,1)\right\}. \tag{2.4}$$

Thus, the spectral risk measure uncertainty set constrains the *entire* marginal distribution of the likelihood ratio, as opposed to the moment conditions used in f-DRO.

Given (2.2), in the empirical setting, the learning problem of interest reduces to

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{\beta \in \mathcal{B}(P_n)} \left\{\mathbb{E}_{P_n}\left[\beta(\xi)\ell(\boldsymbol{\theta}, \xi)\right] = \frac{1}{n}\sum_{i=1}^{n} \beta(\xi_i)\ell(\boldsymbol{\theta}, \xi_i)\right\}. \tag{2.5}$$

Notice that the maximization in (2.5) only depends on the n numbers $(\beta(\xi_1)/n, \ldots, \beta(\xi_n)/n)$, which can naturally be interpreted as importance weights for each training example of the objective, as $\frac{1}{n}\sum_{i=1}^{n} \beta(\xi_i) = \mathbb{E}_{\xi \sim P_n}\left[\beta(\xi)\right] = 1$. Even though distributions in the uncertainty set reduce to reweightings of the given training examples, we emphasize that the objective in (2.5) is viewed as a consistent estimator of its population counterpart, which contains distributions that need only be absolutely continuous with respect to P. We proceed to the specific properties of SRMs in Section 2.3 and then discuss algorithms to optimize (2.1) in Section 2.4.

2.3 Properties of Spectral Risk Measures

SRMs have been studied extensively in quantitative finance [Artzner et al., 1999, Föllmer and Schied, 2002, Rockafellar and Uryasev, 2013, Acerbi and Tasche, 2002, Pflug and Ruszczyński, 2005, Kuhn et al., 2019] and convex analysis Rockafellar and Royset [2014], Ben-Tal and Teboulle [2007]. Despite being relatively underexplored in machine learning, one particu-

lar SRM called the *superquantile* or *conditional value-at-risk (CVaR)* has recently received careful attention in the learning setting [Curi et al., 2020, Levy et al., 2020, Laguel et al., 2020, 2021], though it has not been unified with DRO as shown in this chapter. As we will see in Section 2.7, they also have particular computational properties that make them an appealing choice for practitioners.

Because SRMs are really summaries of univariate real-valued random variables, we will derive some relevant properties by only considering a real-valued random variable X representing the loss at a particular parameter $\boldsymbol{\theta}$, i.e., $X = \ell(\boldsymbol{\theta}, \xi)$. The randomness is induced by $\xi \sim P$. We first provide a novel variational form of this functional that is expressed in terms of likelihood ratios, which gives (2.4). We then state a number of examples. In the empirical setting, the maximum can be expressed in terms of an even simpler feasible set, which we discuss alongside optimization details in Section 2.4.

2.3.1 Spectral Risk Measures and the Likelihood Ratio

Before stating the result, note for context that the uncertainty sets in f-DRO (see (2.3)) are based on relatively coarse attributes of β. For example, the χ^2-divergence is given by $f(x) = x^2 - 1$, indicating that any distribution such that the second moment is bounded as $\mathbb{E}_P[\beta^2(\xi)] \leq \rho + 1$ is feasible. As we will see below, SRMs constrain the entire marginal distribution of β, which is a much stronger condition.

Proposition 2.3.1. *Let s be left-continuous (in addition to being bounded, measurable, non-negative, and non-decreasing). Then, there exists a unique CDF $G_s(v) := \sup\{t : s(t) \leq v\}$ such that*

$$\mathbb{L}_s[\ell(\boldsymbol{\theta}, \xi)] = \max_{\beta \in \mathcal{B}(P)} \mathbb{E}_{\xi \sim P}[\beta(\xi)\ell(\boldsymbol{\theta}, \xi)] \tag{2.6}$$

for

$$\mathcal{B}(P) = \{\beta \in \mathcal{B}_0(P) : \mathbb{P}[\beta(\xi) \leq v] = G_s(v)\}. \tag{2.7}$$

Furthermore, $\beta_\star(\xi)$ maximizing (2.6) can be written as a measurable function of $\ell(\boldsymbol{\theta}, \xi)$.

Proof. Let F be the CDF of $\ell(\boldsymbol{\theta}, \xi)$. Because s is left-continuous and non-decreasing, we have by Bobkov and Ledoux [2019, Proposition A.2] that there exists a unique CDF G_s given by the expression in the statement such that the right generalized inverse of G_s is s. Thus, we have that

$$\begin{aligned}\mathbb{L}_s[\ell(\boldsymbol{\theta}, \xi)] &= \int_0^1 s(t)F^{-1}(t)\,\mathrm{d}t = \int_0^1 G_s^{-1}(t)F^{-1}(t)\,\mathrm{d}t \\ &= \max\left\{\mathbb{E}\left[V\ell(\boldsymbol{\theta}, \xi)\right] : V : \Omega \to \mathbb{R} \text{ measurable and } \mathbb{P}\left[V \le v\right] = G_s(v)\right\} \qquad (2.8) \\ &\ge \max\left\{\mathbb{E}\left[\beta(\xi)\ell(\boldsymbol{\theta}, \xi)\right] : \beta \in \mathcal{B}(P)\right\}.\end{aligned}$$

To make the inequality above into an equality, we must prove that for $V_\star$ maximizing (2.8), there exists a $\beta_\star$ in the feasible set above such that $V_\star = \beta_\star(\xi)$. Accordingly, define,

$$\beta_\star(\xi) := G_s^{-1}(F(\ell(\boldsymbol{\theta}, \xi)) = s(F(\ell(\boldsymbol{\theta}, \xi)),$$

which is non-negative because s is non-negative. Measurability follows from the measurability of $\ell(\boldsymbol{\theta}, \cdot)$, F, and s. To show that $\beta_\star$ is a likelihood ratio, compute the expectation:

$$\mathbb{E}_P\left[\beta_\star(\xi)\right] = \mathbb{E}_P\left[G_s^{-1}(F(\ell(\boldsymbol{\theta}, \xi)))\right] = \int_0^1 G_s^{-1}(t)\,\mathrm{d}t = 1,$$

because $F(\ell(\boldsymbol{\theta}, \xi)) \sim \mathrm{Unif}(0, 1)$ and $\int_0^1 s(t)\,\mathrm{d}t = 1$. This completes the proof. □

While Proposition 2.3.1 expresses a measure of tail risk as a distributionally robust objective, the reverse has been done for f-divergences, in that their dual form (see Shapiro [2017, Section 3.2]) is interpreted as a tail error. To understand the CDF G_s, or rather the random variable it corresponds to, we consider two examples of SRMs. Here, we present their continuous spectra $t \mapsto s(t)$ as opposed to their discretizations in (2.13) and (2.14).

- **Superquantile:** For $\tau \in (0, 1]$, the *τ-superquantile* [Rockafellar and Royset, 2013],

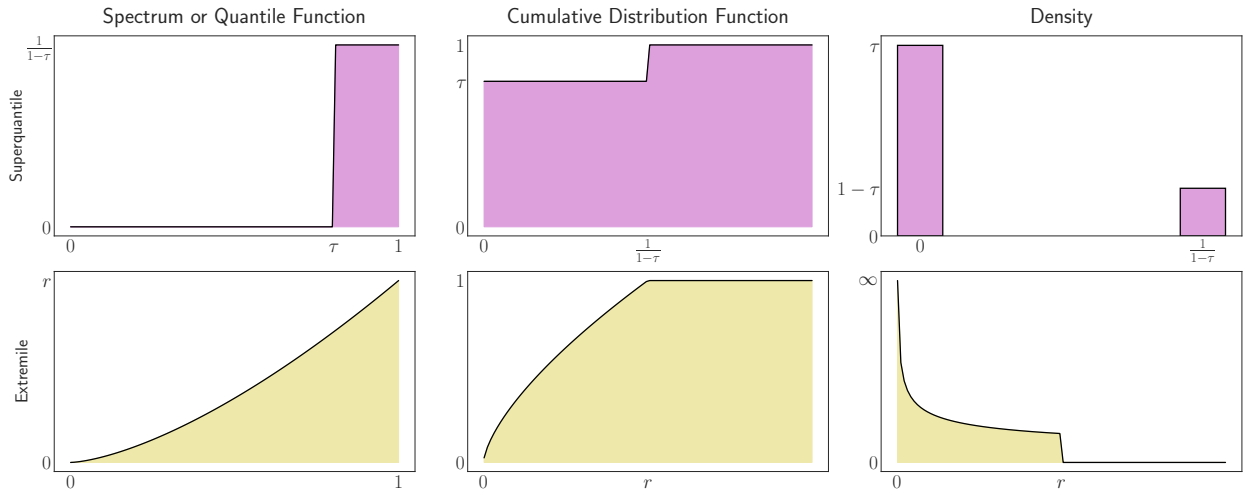

Figure 2.1: **Example Spectra and Induced Random Variables.** Visualization of the superquantile and extremile spectra and the induced distribution function G_s of the likelihood ratio.

a.k.a. conditional value-at-risk, is specified by $s(t) = \frac{1}{1-\tau}\mathbb{1}\{\tau \le t \le 1\}$. Thus, we have

$$G_s(v) = \tau\mathbb{1}\{v \ge 0\} + (1-\tau)\mathbb{1}\{v \ge 1/(1-\tau)\}.$$

In other words, it is the CDF of a Bernoulli-like random variable that is 0 with probability $1-\tau$ and $1/\tau$ with probability τ.

- **Extremile:** For $r > 1$, the *r-extremile* [Daouia et al., 2019] is specified by $s(t) = rt^{r-1}$. Thus, we can compute by direct inversion

$$G_s(v) = \left(\frac{v}{r}\right)^{\frac{1}{r-1}} \mathbb{1}\{v \in [0, r]\} + \mathbb{1}\{v > r\}.$$

These examples are visualized alongside their quantile function, CDF, and densities in Figure 2.1.

2.3.2 Bias Bounds for Spectral Risk Measures

Bias bounds that are uniform over $\boldsymbol{\theta} \in \Theta$ help establish the performance guarantees of stochastic gradient descent-type algorithms, as done in Levy et al. [2020], for instance. They are also helpful for establishing uniform convergence of $\mathcal{R}(\cdot, P_n)$ process to $\mathcal{R}(\cdot, P)$, as one may perform the decomposition

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathcal{R}(\boldsymbol{\theta}, P_n) - \mathcal{R}(\boldsymbol{\theta}, P)| \leq$$
$$\underbrace{\sup_{\boldsymbol{\theta}\in\Theta} |\mathcal{R}(\boldsymbol{\theta}, P_n) - \mathbb{E}_P\left[\mathcal{R}(\boldsymbol{\theta}, P_n)\right]|}_{\text{concentration}} + \underbrace{\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{E}_P\left[\mathcal{R}(\boldsymbol{\theta}, P_n)\right] - \mathcal{R}(\boldsymbol{\theta}, P)|}_{\text{bias}},$$

and control the first term using standard technical tools, namely, concentration inequalities. As before, we will first consider a real-valued random variable X (which we may think of as $X = \ell(\boldsymbol{\theta}, \xi)$) with CDF F, and n i.i.d. copies denoted $X_1, \ldots, X_n \sim F$. Assume in addition that $\mathbb{E}_F\left[|X_1|\right] < \infty$. As before, we let F^{-1} denote the quantile function and let s be a spectrum for the spectral risk measure $\mathbb{L}_s[\cdot]$. Similarly, let F_n and F_n^{-1} denote the empirical CDF and empirical quantile function, respectively. We use the notation $\mathbb{L}_s[F]$ below to indicate $\mathbb{L}_s[X]$ for X with CDF F. Under very general settings, namely the existence of greater-than-two moments of X, the bias can be shown to decay at an $n^{-1/2}$ rate. Later in this section, we make stronger assumptions to produce a bound that decays at rate n^{-1}. We will make use of the following theorem.

Theorem 2.3.1 (Theorem 2.10 of Bobkov and Ledoux [2019])**.** *Consider two two probability distributions on $\mathbb{R}$ with associated CDFs F and G, respectively, and quantile functions $F^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) \geq t\}$ and $G^{-1}(t) := \inf\{x \in \mathbb{R} : G(x) \geq t\}$. Given that the random variables associated with F and G are absolutely integrable, we have that*

$$\int_{-\infty}^{\infty} |F(x) - G(x)| \, dx = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right| \, dt,$$

where both the left and right-hand sides are finite.

We proceed to the first bias bound.

Proposition 2.3.2. *Assume that for real-valued random variable X with CDF X, it holds that $\mathbb{E}_F |X|^p < \infty$ for $p > 2$. Then, it holds that*

$$|\mathbb{E}_F[\mathbb{L}_s[F_n]] - \mathbb{L}_s[F]| \leq \sqrt{\frac{2}{n}} \left(\frac{p}{p-2} \right) \sup_{t \in (0,1)]} |s(t) - 1| \cdot \|X\|_{\mathbf{L}^p(\mathbb{P})},$$

where $\|X\|^p_{\mathbf{L}^p(\mathbb{P})} := \int_\Omega X^p(\omega)\, d\mathbb{P}(\omega)$. We interpret the supremum above as the essential supremum according to the uniform measure on $(0, 1)$.

Proof. In the notation below, we will sometimes write $F_n(\cdot; \omega)$ to indicate a realixation of $F_n(\cdot)$ at outcome $\omega \in \Omega$, and similarly for $F_n^{-1}(\cdot; \omega)$ and $F_n^{-1}(\cdot)$. We first observe that

$$\mathbb{E}_F \left[\int_0^1 F_n^{-1}(t)\, \mathrm{d}t \right] = \mathbb{E}_F \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \mathbb{E}_F [X] = \int_0^1 F^{-1}(t)\, \mathrm{d}t.$$

Thus,

$$\begin{aligned} |\mathbb{E}_F[\mathbb{L}_s[F_n]] - \mathbb{L}_s[F]| &= \left| \mathbb{E}_F \left[\int_0^1 (s(t) - 1)(F_n^{-1}(t) - F^{-1}(t))\, \mathrm{d}t \right] \right| \\ &\leq \sup_{t \in [0,1]} |s(t) - 1| \cdot \mathbb{E}_F \left[\int_0^1 \left| F_n^{-1}(t) - F^{-1}(t) \right| \mathrm{d}t \right] \\ &\leq \sup_{t \in [0,1]} |s(t) - 1| \cdot \sqrt{\mathbb{E}_F \left[\left(\int_0^1 |F_n^{-1}(t) - F^{-1}(t)|\, \mathrm{d}t \right)^2 \right]}, \end{aligned}$$

where the second inequality is an application of Jensen's inequality. By Theorem 2.3.1, we also have that $\int_0^1 |F_n^{-1}(t; \omega) - F^{-1}(t)|\, \mathrm{d}t = \int_{-\infty}^\infty |F_n(x; \omega) - F(x)|\, \mathrm{d}x$, indicating that the random variable

$$\omega \mapsto \int_{-\infty}^\infty |F_n(x; \omega) - F(x)|\, \mathrm{d}x \in \mathbf{L}^2(\mathbb{P}).$$

By the triangle inequality on $\mathbf{L}^2(\mathbb{P})$, we have that

$$\begin{aligned}\sqrt{\mathbb{E}\left[\left(\int_0^1 |F_n^{-1}(t) - F^{-1}(t)|\,\mathrm{d}t\right)^2\right]} &= \sqrt{\mathbb{E}\left[\left(\int_{-\infty}^{\infty} |F_n(x) - F(x)|\,\mathrm{d}x\right)^2\right]} \\ &= \left\|\int_{-\infty}^{\infty} |F_n(x) - F(x)|\,\mathrm{d}x\right\|_{\mathbf{L}^2(\mathbb{P})} \\ &\leq \int_{-\infty}^{\infty} \|F_n(x) - F(x)\|_{\mathbf{L}^2(\mathbb{P})}\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \sqrt{\mathbb{E}\left[|F_n(x) - F(x)|^2\right]}\,\mathrm{d}x.\end{aligned}$$

Next, notice that for fixed $x \in \mathbb{R}$, $nF_n(x) \sim \mathrm{Binom}(n, F(x))$, so that

$$\mathbb{E}\left[|F_n(x) - F(x)|^2\right] = \mathbb{V}\mathrm{ar}\, F_n(x) = \frac{F(x)\,(1 - F(x))}{n}.$$

Appling Lemma A.1.1 gives

$$\sqrt{\mathbb{E}_F\left[\left(\int_0^1 |F_n^{-1}(t) - F^{-1}(t)|\,\mathrm{d}t\right)^2\right]} \leq \sqrt{\frac{2}{n}}\left(\frac{p}{p-2}\right)\|X\|_{\mathbf{L}^p(\mathbb{P})},$$

which achieves the desired claim. □

For context, Proposition 2.3.2 operates in general conditions that are of particular importance in optimization. To put this in context, a number of works provide non-asymptotic uniform learning bounds on spectral (and related) risks [Maurer et al., 2021, Khim et al., 2020, Lee et al., 2020]. However, these approaches require boundedness of the random variable of interest, which eliminates any potential application to heavy-tailed losses. Asymptotic approaches proceed by assuming Lipschitz continuity of the spectrum s [Shao, 1989], the trimming of s (i.e., $s(t) = 0$ for all $t \in [0, \alpha) \cup (1 - \alpha, 1]$ with $0 < \alpha \leq 1$) [Shorack, 2017, Shao, 1989], or bounded derivatives of the population quantile function F^{-1} [Xiang, 1995]. The τ-superquantile does not even have a continuous spectrum, whereas the spectrum of the r-extremile is not Lipschitz for $1 \leq r < 2$. Because s must be non-decreasing to achieve convexity (as we discuss in the upcoming Section 2.4), trimming the upper tail of s

is not reflective of practice. Finally, because losses such as the square loss or logistic loss can grow to infinity for unbounded inputs, the derivative $F^{-1}(t)$ as $t \to \infty$ cannot be assumed to be bounded. Proposition 2.3.2 only requires that the population loss satisfies a moment condition and holds without trimming or assumptions of boundedness or Lipschitz continuity on the spectrum. Other recent works employ concentration of the empirical measure in Wasserstein distance to give concentration inequalities for spectral risk measures under sub-Gaussian conditions and moment conditions similar to ours [Prashanth and Bhat, 2022, Bhat and Prashanth, 2019, Pandey et al., 2019].

When we are willing to assume boundedness of the random variable in question, we may place smoothness conditions on the inverse CDF to achieve a decay of n^{-1}.

Proposition 2.3.3. *Assume that F^{-1} is twice differentiable on $(0,1)$, and the derivatives satisfy the conditions $\sup_{t\in(0,1)} |[F^{-1}]'(t)| \leq M_1$ and $\sup_{t\in(0,1)} |[F^{-1}]''(t)| \leq M_2$. Then,*

$$|\mathbb{E}_F \left[\mathbb{L}_s[F_n]\right] - \mathbb{L}_s[F]| \leq \frac{M_1 s(1)}{2n} + \frac{M_2}{n+1} + \frac{M_2 s(1)}{3(n+2)}.$$

Proof. The proof strategy will be to relate the order statistics of the observed data to the order statistics of uniform random variables via the inverse CDF transform. Then, using the properties of Beta-distributed random variables, we can explicitly compute bias terms where necessary. Denote by $X_{(1)} \leq \ldots \leq X_{(n)}$ the order statistics of the sample. Each of these order statistics can be written equivalently as $X_{(i)} = F^{-1}(U_{(i)})$, where $U_{(i)}$ denotes the i-th order statistic of an independently drawn sample $U_1, \ldots, U_n \sim \text{Unif}(0,1)$. Write

$$\begin{aligned}
\mathbb{E}_F \left[\mathbb{L}_s[F_n]\right] - \mathbb{L}_s[F] &= \mathbb{E}_F \left[\int_0^1 s(t) \left(F_n^{-1}(t) - F^{-1}(t)\right) \, \mathrm{d}t\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_F \left[\int_{(i-1)/n}^{i/n} s(t) \left(F_n^{-1}(t) - F^{-1}(t)\right) \, \mathrm{d}t\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_F \left[\int_{(i-1)/n}^{i/n} s(t) \left(F^{-1}(U_{(i)}) - F^{-1}(t)\right) \, \mathrm{d}t\right].
\end{aligned}$$

Due to the boundedness of s and absolute integrability of the random variable X_1, we may

apply Fubini's theorem so that

$$\mathbb{E}_F\left[\mathbb{L}_s[F_n]\right] - \mathbb{L}_s[F] = \int_{(i-1)/n}^{i/n} s(t)\mathbb{E}_F\left[F^{-1}(U_{(i)}) - F^{-1}(t)\right]\,\mathrm{d}t.$$

Now, fix any $t \in ((i-1)/n, i/n)$, and apply a second-order Taylor expansion of F^{-1} so that

$$F^{-1}(U_{(i)}) - F^{-1}(t) = \underbrace{[F^{-1}]'(t)(U_{(i)} - t)}_{\text{first-order term}} + \underbrace{\frac{[F^{-1}]''(\tilde{U}_{(i)})}{2}(U_{(i)} - t)^2}_{\text{second-order term}},$$

where $\tilde{U}_{(i)}$ is a real-valued random variable which lies in between $U_{(i)}$ and t. Note that because $U_{(i)}$ follows the Beta$(i, n-i+1)$ distribution [David and Nagaraja, 2003], we have that $\mathbb{E}_{\text{Unif}(0,1)}\left[U_{(i)}\right] = i/(n+1)$, which we use in the upcoming calculations.

Controlling the first-order term. Applying the expectation and the sum, we then upper and lower bound the resulting quantity. Using that $[F^{-1}]'$ is bounded by M_1 and non-negative because F^{-1} is non-decreasing,

$$\begin{aligned}\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} s(t)[F^{-1}]'(t)\left(\frac{i}{n+1} - t\right) &\le \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} s(t)[F^{-1}]'(t)\left(\frac{i}{n+1} - \frac{i-1}{n}\right)\,\mathrm{d}t \\ &\le M_1\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} s(t)\left(\frac{i}{n+1} - \frac{i-1}{n}\right)\,\mathrm{d}t \\ &\le \frac{M_1 s(1)}{n}\sum_{i=1}^{n}\left(\frac{i}{n+1} - \frac{i-1}{n}\right). \end{aligned} \tag{2.9}$$

Similarly, we have that

$$\begin{aligned}\sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} s(t)[F^{-1}]'(t)\left(\frac{i}{n+1} - t\right) &\ge \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} s(t)[F^{-1}]'(t)\left(\frac{i}{n+1} - \frac{i}{n}\right)\,\mathrm{d}t \\ &\ge \frac{M_1 s(1)}{n}\sum_{i=1}^{n}\left(\frac{i}{n+1} - \frac{i}{n}\right), \end{aligned} \tag{2.10}$$

We compute the sums in (2.9) and (2.10) to give the upper bound. Observe that

$$\sum_{i=1}^{n}\left(\frac{i}{n+1} - \frac{i-1}{n}\right) = \frac{1}{2} \text{ and } \sum_{i=1}^{n}\left(\frac{i}{n+1} - \frac{i}{n}\right) = -\frac{1}{2},$$

so that

$$\left| \sum_{i=1}^{n} s(t)[F^{-1}]'(t)\mathbb{E}_{\text{Unif}(0,1)}\left[U_{(i)} - t\right] \right| \leq \frac{M_1 s(1)}{2n}.$$

Controlling the second-order term. Here, we use that $[F^{-1}]''$ is bounded by M_2, so that

$$\begin{aligned}
&\mathbb{E}_{\text{Unif}(0,1)}\left[[F^{-1}]''(\tilde{U}_{(i)})(U_{(i)} - t)^2\right] \\
&\leq M_2 \mathbb{E}_{\text{Unif}(0,1)}\left[(U_{(i)} - t)^2\right] \\
&= M_2 \left(\mathbb{V}\text{ar}\, U_{(i)} + \left(\frac{i}{n+1} - t \right)^2 \right) \\
&= M_2 \left(\frac{i(n-i+1)}{(n+1)^2(n+2)} + \left(\frac{i}{n+1} - t \right)^2 \right) \\
&= M_2 \left(\frac{i}{(n+1)(n+2)} - \frac{i^2}{(n+1)^2(n+2)} + \left(\frac{i}{n+1} - t \right)^2 \right) \\
&\leq M_2 \left(\frac{1}{n+1} - \frac{i^2}{(n+1)^2(n+2)} + \left(\frac{i}{n+1} - t \right)^2 \right),
\end{aligned}$$

where we used the bias-variance decomposition of the mean squared error and the moments of the Beta$(i, n-i+1)$ distribution. We group the second and third terms and take the integral, so that

$$\begin{aligned}
&\sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} s(t)\mathbb{E}_{\text{Unif}(0,1)}\left[[F^{-1}]''(\tilde{U}_{(i)})(U_{(i)} - t)^2\right] \mathrm{d}t \\
&\leq M_2 \left(\int_0^1 s(t)\,\mathrm{d}t \right) \left(\frac{1}{n+1} \right) + M_2 \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} s(t) \left[\left(\frac{i}{n+1} - t \right)^2 - \frac{i^2}{(n+1)^2(n+2)} \right] \mathrm{d}t \\
&\leq \frac{M_2}{n+1} + M_2 s(1) \sum_{i=1}^{n} \left[\int_{(i-1)/n}^{i/n} \left(\frac{i}{n+1} - t \right)^2 \mathrm{d}t - \frac{i^2}{n(n+1)^2(n+2)} \right]. \qquad (2.11)
\end{aligned}$$

It remains to control the second term of (2.11). Expanding the square and integrating, we

generate telescoping terms and sum them to achieve

$$\begin{aligned}\sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} \left(\frac{i}{n+1} - t\right)^2 \mathrm{d}t &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{i}{n+1}\right)^2 - 1 + \frac{1}{3} \\ &\leq \frac{1}{n(n+1)^2} \sum_{i=1}^{n} i^2 \\ &= \frac{(2n+1)}{6(n+1)}.\end{aligned}$$

Then, including the negative term from (2.11), we have that

$$\begin{aligned}&\sum_{i=1}^{n} \left[\int_{(i-1)/n}^{i/n} \left(\frac{i}{n+1} - t\right)^2 \mathrm{d}t - \frac{i^2}{n(n+1)^2(n+2)}\right] \\ &= \frac{2n+1}{6}\left(\frac{1}{n+1} - \frac{1}{n+2}\right) = \frac{2n+1}{6(n+1)(n+2)} \leq \frac{1}{3(n+2)}.\end{aligned}$$

Putting these steps together, the upper bound on the second-order term reads as

$$\mathbb{E}_{\mathrm{Unif}(0,1)}\left[[F^{-1}]''(\tilde{U}_{(i)})(U_{(i)} - t)^2\right] \leq \frac{M_2}{n+1} + \frac{M_2 s(1)}{3(n+2)},$$

which completes the proof. □

Note that the Lipschitzness of the inverse CDF was also employed in Levy et al. [2020] to establish guarantees on DRO using the χ^2-divergence and CVaR uncertainty sets. So far, the results we have shown have relied on the continuous viewpoint of SRMs as functionals of a real-valued random variable. We can derive useful closed-form expressions for the empirical variants of these quantities.

2.3.3 Empirical Spectral Risk Measures as L-Statistics

One key connection that we will exploit is that spectral risk measures can be expressed as a function of the order statistics of the sample in the empirical setting. This is because the empirical quantile function can be written in terms of the order statistics as $F_n^{-1}(t) = X_{(\lceil nt \rceil)}$, as seen in Figure 2.2 (top). Notice in particular that when $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right)$, we have that

$F_n^{-1}(t) = X_{(i)}$, where end-points are chosen to make F_n^{-1} left continuous. Then, write

$$\begin{aligned}\mathbb{L}_s\left[F_n\right] &:= \int_0^1 s(t) \cdot F_n^{-1}(t)\,\mathrm{d}t = \sum_{i=1}^{n}\left(\int_{(i-1)/n}^{i/n} s(t) \cdot X_{(\lceil nt \rceil)}\,\mathrm{d}t\right) \\ &= \sum_{i=1}^{n}\left(\int_{(i-1)/n}^{i/n} s(t)\,\mathrm{d}t\right) X_{(i)} = \sum_{i=1}^{n} \sigma_i X_{(i)} \end{aligned} \tag{2.12}$$

for $\sigma_i := \left(\int_{(i-1)/n}^{i/n} s(t)\,\mathrm{d}t\right)$. The discretized weights $\sigma_1, \ldots, \sigma_n$ are shown in Figure 2.2 (bottom). The expression (2.12) is called an *L-estimator* [Shorack, 2017, Maurer et al., 2021] for a generic linear combination of order statistics and an *L-risk* when the ordered elements are losses incurred on a training set [Maurer et al., 2021, Khim et al., 2020]. The σ_i's allow the practitioner to interpolate between the average-case ($\sigma_i = 1/n\ \forall i$) and worst-case ($\sigma_n = 1$) performance on the training set. Such objectives have garnered a flurry of recent interest in machine learning [Fan et al., 2017, Williamson and Menon, 2019, Khim et al., 2020, Maurer et al., 2021, Holland and Mehdi Haress, 2022, Leqi et al., 2019, Lee et al., 2020, Kawaguchi and Lu, 2020].

As for particular examples: for $\tau \in (0, 1)$, the *τ-CVaR* (a.k.a. superquantile) [Rockafellar and Royset, 2013, Kawaguchi and Lu, 2020, Laguel et al., 2021] requires that $k = n(1-\tau)$ elements of σ be non-zero with equal probability and that the remaining $n-k$ are zero. The *r-extremile* [Daouia et al., 2019] and *γ-exponential spectral risk measure* [Cotter and Dowd, 2006] define their spectra the equations below.

$$\sigma_i = \begin{cases} \frac{1}{n(1-\tau)} & \text{if } i \in \{\lceil n\tau \rceil, \ldots, n\} \\ 1 - \frac{\lfloor n(1-\tau) \rfloor}{n(1-\tau)} & \text{if } \lfloor n\tau \rfloor < i < \lceil n\tau \rceil \end{cases} \qquad \tau\text{-CVaR},\ \tau \in (0,1) \tag{2.13}$$

$$\sigma_i = \left(\tfrac{i}{n}\right)^r - \left(\tfrac{i-1}{n}\right)^r \qquad b\text{-extremile},\ r \geq 1 \tag{2.14}$$

$$\sigma_i = \frac{e^{\gamma}\left(e^{\gamma i/n} - e^{\gamma(i-1)/n}\right)}{1 - e^{-\gamma}} \qquad \gamma\text{-ESRM},\ \gamma > 0 \tag{2.15}$$

The multiple cases in the CVaR definition account for the instance in which $n(1-\tau)$ is not an integer.

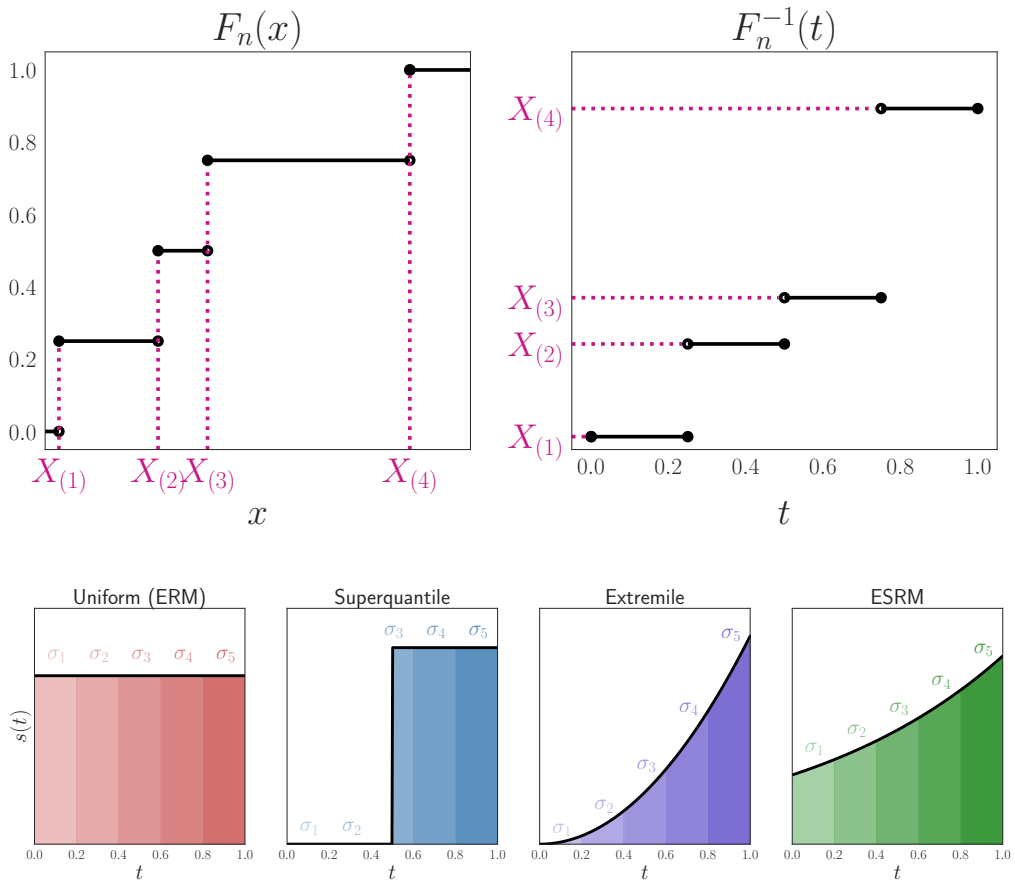

Figure 2.2: **Illustration of Spectral Risk and Quantile Functions.** The relationship between the order statistics, quantile functions, and the discretized spectrum is shown. **Top:** Empirical CDF F_n and quantile function F_n^{-1} of $X_1, \cdots, X_4$. **Bottom:** Continuous spectra $s(t)$ and their discretization $(\sigma_1, \ldots, \sigma_5)$ for various risk measures.

While the bias bounds established above are general-purpose, we return to the optimization problem introduced in Section 2.1 in the next section. We provide a variational form of (2.12) in the empirical setting, which will provide the relationship with distributionally robust optimization by defining a corresponding uncertainty set for SRMs. Observe that because s is non-decreasing, it holds that $\sigma_1 \leq \ldots \leq \sigma_n$. Thus, for any vector

$\boldsymbol{l} = (l_1, \ldots, l_n) \in \mathbb{R}^n$ with ordered entries $l_{(1)} \leq \ldots \leq l_{(n)}$, we have that

$$\sum_{i=1}^{n} \sigma_i l_{(i)} = \max_{\text{permutations } \pi} \sum_{i=1}^{n} \sigma_{\pi(i)} l_i = \max_{\boldsymbol{q} \in \mathcal{Q}(\sigma)} \sum_{i=1}^{n} q_i l_i,$$

where

$$\mathcal{Q}(\sigma) := \operatorname{conv} \{\text{permutations of } \sigma\}, \tag{2.16}$$

also known as the *permutahedron* of σ in $\mathbb{R}^n$. Using this uncertainty set, we address the optimization problem from Section 2.1.

2.4 Smoothing Maximum-Type Objectives

We now write our objective in the form of (2.1), which is amenable to being studied as an optimization problem. First, for notational ease, define $\ell_i(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}, \xi_i)$ and the vector $\ell(\boldsymbol{\theta}) = (\ell_1(\boldsymbol{\theta}), \ldots, \ell_n(\boldsymbol{\theta}))$. Then,

$$\max_{\beta \in \mathcal{B}(P_n)} \frac{1}{n} \sum_{i=1}^{n} \beta(\xi_i) \ell(\boldsymbol{\theta}, \xi_i) = \max_{\boldsymbol{q} \in \mathcal{Q}} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta}) \rangle, \tag{2.17}$$

where $\mathcal{Q} := \{\beta(\xi_i)/n : \beta \in \mathcal{B}(P_n)\}$. As shown in Section 2.1, the elements of any $\boldsymbol{q} = (q_1, \ldots, q_n) \in \mathcal{Q}$ are weights of a probability mass function, hence $\sum_{i=1}^{n} q_i = 1$. As an abuse of terminology, we will also refer to $\mathcal{Q}$ as the *uncertainty set* in this context.

As is common in statistical and machine learning, we will regularize the quantity $\langle \boldsymbol{q}, \ell(\boldsymbol{\theta}) \rangle$ from (2.17) both on the primal and dual side, with respective hyperparameters $\mu > 0$ and $\nu > 0$, and consider the parameter space $\Theta = \mathbb{R}^d$. To emphasize the statistical context, rather than a primal-dual viewpoint, we consider the objective to be a composition of two functions. The first function $\mathcal{A}_\nu$ aggregates the vector (or equivalently, the histogram) of losses $\ell(\boldsymbol{\theta})$ for a particular $\boldsymbol{\theta} \in \Theta$ and produces a univariate summary, given by

$$\mathcal{A}_\nu(\boldsymbol{l}) := \max_{\boldsymbol{q} \in \mathcal{Q}} \Big\{ \langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu D_f(\boldsymbol{q} \| \mathbf{1}/n) \Big\},$$

where $D_f(\boldsymbol{q} \| \mathbf{1}/n) = \frac{1}{n} \sum_{i=1}^{n} f(n q_i)$ is an f-divergence with generator f over empirical mea-

sures. Because ν controls how much we penalize $\boldsymbol{q}$ from shifting to far from the original uniform weights, we will refer to this parameter as the *shift cost*. The map $\mathcal{A}_\nu$ is in fact differentiable (as opposed to $\boldsymbol{l} \mapsto \max_{\boldsymbol{q}\in\mathcal{Q}} \langle \boldsymbol{q}, \boldsymbol{l}\rangle$), which will be an essential property when deriving the optimization algorithm and establishing the convergence guarantees. Thus, our problem of interest can be written as

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \left[\mathcal{L}(\boldsymbol{\theta}) := \mathcal{A}_\nu(\ell(\boldsymbol{\theta})) + \frac{\mu}{2}\left\|\boldsymbol{\theta}\right\|_2^2\right], \tag{2.18}$$

As introduced in Section 2.2 maintain two prototypical classes of examples for the uncertainty set $\mathcal{Q}$: f-DRO and SRMs. Our theoretical analyses simply rely on $\mathcal{Q}$ being a closed, convex subset of Δ^{n-1}, whereas implementation details (such as how to compute and differentiate the map $\mathcal{A}_\nu$) will depend heavily on the specific uncertainty set used and are discussed in Section 2.7. Our goal is to design and analyze an iterative algorithm that produces a sequence $(\boldsymbol{\theta}^{(k)})_{t\geq 0}$ converging to the solution of (2.18) with a favorable dependence on the sample size n and constants associated to the loss ℓ. Consider the following standard assumptions.

Assumption 2.4.1. Each ℓ_i is convex, G-Lipschitz continuous w.r.t. $\left\|\cdot\right\|_2$, and L-smooth w.r.t. $\left\|\cdot\right\|_2$ (i.e., $\nabla\ell_i$ exists and is L-Lipschitz continuous). The generator f of D_f is α_n-strongly convex on the interval $[0, n]$. The constants μ, ν, and α_n are positive.

Common examples of α_n are $2n$ for the χ^2-divergence and 1 for the KL-divergence. Before proposing an algorithm, we use gradient descent as a baseline. This involves showing that the composition $\boldsymbol{\theta} \mapsto \mathcal{A}_\nu(\ell(\boldsymbol{\theta}))$ is indeed differentiable (in fact, smooth) and characterizing the properties of the objective (2.18).

By Proposition A.1.1, we see that if f is α_n-strongly convex function over $[0, n]$, the f-divergence $D_f(\cdot\|\mathbf{1}/n)$ will be $(\alpha_n n)$-strongly convex. Given Assumption 2.4.1, we assume that D_f is rescaled to be 1-strongly convex for ease of presentation, define $\mathrm{Reg}(\boldsymbol{q}) := (\alpha_n n)^{-1} D_f(q\|\mathbf{1}/n)$, and rewrite our risk functional as

$$\mathcal{A}_\nu(\boldsymbol{l}) := \max_{\boldsymbol{q}\in\mathcal{Q}} \Big\{ \langle \boldsymbol{q}, \boldsymbol{l}\rangle - \nu\,\mathrm{Reg}(\boldsymbol{q})\Big\}, \tag{2.19}$$

In the following, let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ denote its associated dual norm.

Lemma 2.4.1. *When $\mathcal{Q}$ is closed and convex, and the map* Reg *is* 1*-strongly convex over $\mathcal{Q}$ with respect to $\|\cdot\|$, then $\mathcal{A}_\nu$ is continuously differentiable with gradient given by*

$$q^{\mathrm{opt}}(\boldsymbol{l}) := \nabla \mathcal{A}_\nu(\boldsymbol{l}) = \operatorname*{arg\,max}_{\boldsymbol{q} \in \mathcal{Q}} \{\langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{q})\} \in \mathbb{R}^n.$$

Moreover, q^{opt} is $(1/\nu)$-Lipschitz continuous with respect to $\|\cdot\|_*$.

Proof. Because $\mathcal{Q}$ is a closed subset of the compact set Δ^{n-1}, it is also compact. Because $\mathcal{Q}$ is compact and convex, the strongly concave function $\boldsymbol{q} \mapsto \langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{q})$ has a unique maximizer. By Danskin's theorem [Bertsekas, 1997, Proposition B.25], we have that $\mathcal{A}_\nu$ is continuously differentiable with the given gradient formula. By Nesterov [2018, Lemma 6.1.2], it is also $(1/\nu)$-Lipschitz continuous with respect to $\|\cdot\|_*$. □

Finally, we can pass the smoothness result of Lemma 2.4.1 onto $\mathcal{L}$ through a simple application of the chain rule. Letting $\nabla \ell(\boldsymbol{\theta}) \in \mathbb{R}^{n \times d}$ be the Jacobian of ℓ, we have that the gradient of $\boldsymbol{\theta} \mapsto \mathcal{A}_\nu(\ell(\boldsymbol{\theta}))$ is given by $\nabla \ell(\boldsymbol{\theta})^\top q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}))$. Then, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, we have that

$$\begin{aligned}
&\left\|\nabla \ell(\boldsymbol{\theta})^\top q^{\mathrm{opt}}(\ell(\boldsymbol{\theta})) - \nabla \ell(\boldsymbol{\theta}')^\top q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}'))\right\|_2 \\
&\leq \left\|\nabla \ell(\boldsymbol{\theta})^\top (q^{\mathrm{opt}}(\ell(\boldsymbol{\theta})) - q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}')))\right\|_2 + \left\|(\nabla \ell(\boldsymbol{\theta}) - \nabla \ell(\boldsymbol{\theta}'))^\top q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}'))\right\|_2] \\
&\leq \frac{nG}{\nu} \left\|\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}')\right\|_2 + L \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2^2 \qquad (2.20) \\
&\leq \left(\frac{nG^2}{\nu} + L\right) \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2^2. \qquad (2.21)
\end{aligned}$$

As a result $\mathcal{L}$ is $(L + nG^2/\nu + \mu)$-smooth and μ-strongly convex with respect to $\|\cdot\|_2$. By Nesterov [2018, Theorem 2.1.15], using the gradient descent update

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})$$

Algorithm 1 Prospect

Inputs: Initial points $\boldsymbol{\theta}^{(0)}$, stepsize $\eta > 0$, number of iterations t.

Set $\tilde{\boldsymbol{\theta}}_i^{(0)} = \hat{\boldsymbol{\theta}}_i^{(0)} = \boldsymbol{\theta}^{(0)}$ for all $i \in [n]$, with $\tilde{\boldsymbol{\theta}}^{(0)} = (\tilde{\boldsymbol{\theta}}^{(0)})_{i=1}^n$ and $\hat{\boldsymbol{\theta}}^{(0)} = (\hat{\boldsymbol{\theta}}^{(0)})_{i=1}^n$.

$\boldsymbol{q}^{(0)} = q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}^{(0)}))$, $\boldsymbol{\rho}^{(0)} = \boldsymbol{q}^{(0)}$.

Set $\boldsymbol{l}^{(0)} = (\ell_i(\hat{\boldsymbol{\theta}}_i^{(0)}))_{i=1}^n \in \mathbb{R}^n$, $\boldsymbol{g}^{(0)} = (\nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(0)}))_{i=1}^n \in \mathbb{R}^{d \times n}$.

for $k = 0, \ldots, t-1$ **do**

  $i_k \sim \mathrm{Unif}([n])$, $j_k \sim \mathrm{Unif}([n])$.

  $\boldsymbol{v}^{(k)} = n q_{i_k}^{(k)} \nabla r_{i_k}(\boldsymbol{\theta}^{(k)}) - (n \rho_{i_k}^{(k)} \nabla r_{i_k}(\tilde{\boldsymbol{\theta}}_{i_k}^{(k)}) - \sum_{i=1}^n \rho_i^{(k)} \boldsymbol{g}_i^{(k)}$.

  $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta \boldsymbol{v}^{(k)}$.

  $\hat{\boldsymbol{\theta}}^{(k+1)} = \mathrm{UpdateBiasReductionTable}(\boldsymbol{\theta}^{(k)}, \hat{\boldsymbol{\theta}}^{(k)}, i_k)$.

  $\boldsymbol{l}^{(k+1)} = \ell(\hat{\boldsymbol{\theta}}^{(k+1)})$.

  $\boldsymbol{q}^{(k+1)} = q^{\mathrm{opt}}(\boldsymbol{l}^{(k+1)})$.

  $\tilde{\boldsymbol{\theta}}^{(k+1)}, \boldsymbol{\rho}^{(k+1)} = \mathrm{UpdateVarianceReductionTable}(\boldsymbol{\theta}^{(k)}, \tilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}^{(k)}, i_k)$.

  $\boldsymbol{g}^{(k+1)} = (\nabla r_i(\tilde{\boldsymbol{\theta}}^{(k+1)}))_{i=1}^n$.

Output: Final point $\boldsymbol{\theta}^{(t)}$

would achieve ε-suboptimality in

$$t = O\left(n\left(\frac{L}{\mu} + \frac{nG^2}{\mu\nu} + 1\right)\ln\frac{1}{\varepsilon}\right)$$

calls to the first-order oracles $\{(\ell_i, \nabla\ell_i)\}_{i=1}^n$. Given this baseline, the primary goal in designing the upcoming algorithm is to decouple the term n (representing a full-batch of oracle calls on a single iteration) and the condition number term $\frac{L}{\mu} + \frac{nG^2}{\mu\nu} + 1$ from being multiplied to being summed. This type of improvement has been achieved in the setting of empirical risk minimization via variance reduction [Johnson and Zhang, 2013, Defazio et al., 2014]. We devise a similar class of algorithms in the next section.

2.5 Stochastic Optimization with Bias and Variance Reduction

2.5.1 Algorithm Description

For ease of presentation, we use r_i to denote the regularized losses

$$r_i(\boldsymbol{\theta}) := \ell_i(\boldsymbol{\theta}) + \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2 .$$

We present the Prospect method in Algorithm 1. At a high level, the algorithm stores a table of losses $\boldsymbol{l}^{(k)} \in \mathbb{R}^n$ and gradients $\boldsymbol{g}^{(k)} \in \mathbb{R}^{n \times d}$ which are used to approximate $\ell(\boldsymbol{\theta}^{(k)})$ and $\nabla r(\boldsymbol{\theta}^{(k)})$), respectively. In order to accompany the analysis, we use the notation $\hat{\boldsymbol{\theta}}_i^{(k)}$ to denote an element of $\mathbb{R}^d$ such that $l_i^{(k)} = \ell_i(\hat{\boldsymbol{\theta}}_i^{(k)})$ and $\tilde{\boldsymbol{\theta}}_i^{(k)}$ to denote an element of $\mathbb{R}^d$ such that $\boldsymbol{g}_i^{(k)} = \nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(k)})$. In implementation, however, it may not be necessary to store the tables $\hat{\boldsymbol{\theta}}^{(k)} \in \mathbb{R}^{n \times d}$ or $\tilde{\boldsymbol{\theta}}^{(k)} \in \mathbb{R}^{n \times d}$. In fact, $\boldsymbol{g}^{(k)}$ itself may not need to be stored either. The algorithm is written in this way because lines 10 and 14 involve updates to these quantities that are left unspecified, which may be random (even conditionally on i_k).

2.5.2 Bias- and Variance-Reduced Updates

These updates may come in different forms, as long as Condition 2.5.1 and Condition 2.5.2 are satisfied. To understand them, we introduce the notation $\mathbb{E}_k[\cdot]$ to denote the conditional expectation given $\boldsymbol{\theta}^{(k)}$. This integrates the randomness of the randomly drawn index i_k as well as any additional randomness resulting from lines 10 and 14.

Condition 2.5.1 (Bias)**.** The update in line 10 of Algorithm 1 satisfies the following. For functions $h_j : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for $j \in [n]$, it holds that

$$\sum_{j=1}^{n} \mathbb{E}_k \left[h_j(\boldsymbol{\theta}^{(k+1)}, \hat{\boldsymbol{\theta}}_j^{(k+1)}) \right] = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_k \left[h_j(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \right] + \left(1 - \frac{1}{n}\right) \sum_{j=1}^{n} \mathbb{E}_k \left[h_j(\boldsymbol{\theta}^{(k+1)}, \hat{\boldsymbol{\theta}}_j^{(k)}) \right] .$$

Condition 2.5.2 (Variance)**.** The update in line 14 of Algorithm 1 satisfies the following.

For functions $h_i : \mathbb{R}^d \to \mathbb{R}$ for $i \in [n]$, it holds that

$$\sum_{i=1}^{n} \mathbb{E}_k \left[h_i \left(\rho_i^{(k+1)} \boldsymbol{g}_i^{(k+1)} \right) \right] = \frac{1}{n} \sum_{i=1}^{n} h_i \left(q_i^{(k)} \nabla \ell_i(\boldsymbol{\theta}^{(k)}) \right) + \left(1 - \frac{1}{n} \right) \sum_{i=1}^{n} h_i \left(\rho_i^{(k)} \boldsymbol{g}_i^{(k)} \right).$$

After instantiating updates that satisfy these conditions, the implementation of the algorithm may be optimized to ensure an $O(n+d)$ per-iteration time complexity and an $O(nd)$ (or even $O(n+d)$) space complexity. We comment on two particular cases and proceed with the analysis in Section 2.6. Instantiating an algorithm of the form Algorithm 1 involves describing the bias and variance reduction table updates, and then "streamlining" the implementation to ensure that the time complexity of each update is $O(n+d)$. This also depends on which method is used in the computation and recomputation of the weights in line 12.

We present two examples of instantiations of Algorithm 1, and prove that they satisfy Condition 2.5.1 and Condition 2.5.2. In doing so, we gain intuition as to what the role of the tables $\left\{ \hat{\boldsymbol{\theta}}_i^{(k)} \right\}_{i=1}^{n}$ and $\left\{ \tilde{\boldsymbol{\theta}}_i^{(k)} \right\}_{i=1}^{n}$, that is to contain examples that approximately track $\boldsymbol{\theta}^{(k)}$ in expectation along the course of the optimization.

Example 2.5.1 (Prospect-Style Bias Reduction)**.** In Mehta et al. [2024b], the table $\hat{\boldsymbol{\theta}}^{(k)}$ is updated by drawing an uniformly random index j_k (independent of i_k) and applying

$$\hat{\boldsymbol{\theta}}_j^{(k+1)} = \begin{cases} \boldsymbol{\theta}^{(k)} & \text{if } j = j_k \\ \hat{\boldsymbol{\theta}}_j^{(k)} & \text{if } j \neq j_k \end{cases}.$$

We confirm that Condition 2.5.1 holds by first take the expected value with respect to j_k:

$$\begin{aligned}
&\sum_{j=1}^{n} \mathbb{E}_k \left[h_j(\boldsymbol{\theta}^{(k+1)}, \hat{\boldsymbol{\theta}}_j^{(k+1)}) \right] \\
&= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_k \left[h_i(\boldsymbol{\theta}^{(k+1)}, \hat{\boldsymbol{\theta}}_i^{(k+1)}) \mid j_k = j \right] + \left(1 - \frac{1}{n} \right) \sum_{j=1}^{n} \mathbb{E}_k \left[h_j(\boldsymbol{\theta}^{(k+1)}, \hat{\boldsymbol{\theta}}_i^{(k+1)}) \mid j_k \neq j \right] \\
&= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_k \left[h_i(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}) \right] + \left(1 - \frac{1}{n} \right) \sum_{j=1}^{n} \mathbb{E}_k \left[h_j(\boldsymbol{\theta}^{(k+1)}, \hat{\boldsymbol{\theta}}_i^{(k)}) \right],
\end{aligned}$$

as desired. To perform this update in practice, we need not store $\hat{\boldsymbol{\theta}}^{(k)}$ at all; instead, we store

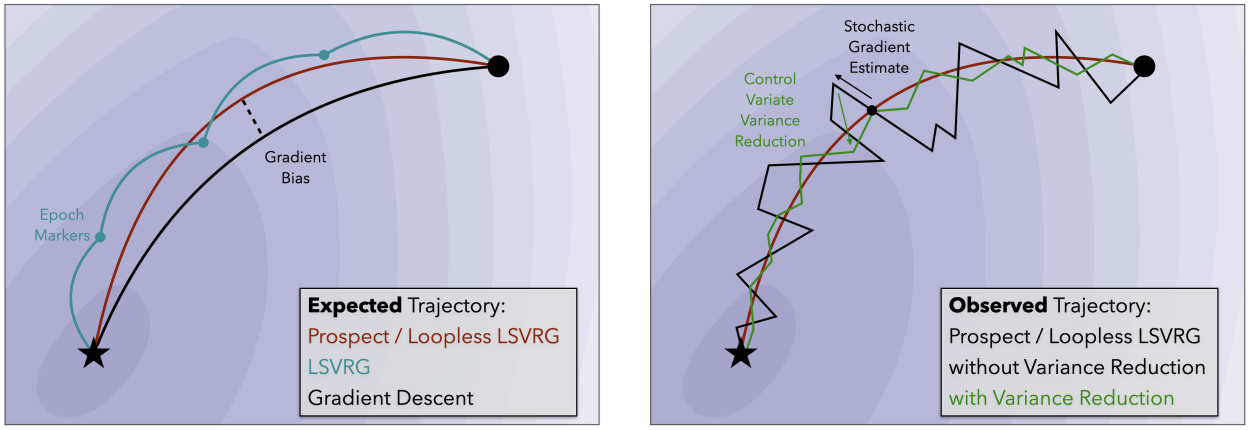

Figure 2.3: **Idealized Visualization of Bias- and Variance-Reduced Algorithms.** The Prospect [Mehta et al., 2024b], LSVRG [Mehta et al., 2023], and Loopless LSVRG (Example 2.5.2) algorithms are shown. **Left:** Expected trajectory over algorithmic randomness. This displays the gradient bias (as compared to full batch gradient descent of each method). LSVRG, unlike the updates in lines 14 and 14 for Algorithm 1, operators in epochs. **Right:** Observed trajectory of either Prospect/Loopless LSVRG. The variance reduction is interpreted as a control variate correction applied to the initial stochastic gradient estimate.

the $O(n)$ sized vector $\boldsymbol{l}^{(k)}$ and update element j_k on iteration t. The memory requirement for this method is then $O(n)$.

Example 2.5.2 (LSVRG-Style Variance Reduction)**.** A loopless variance of LSVRG [Mehta et al., 2023] is described by the following update:

$$(\tilde{\boldsymbol{\theta}}_i^{(k+1)}, \rho_i^{(k+1)}) = \begin{cases} (\boldsymbol{\theta}^{(k)}, q_i^{(k+1)}) \ \forall i & \text{w.p. } \frac{1}{n} \\ (\tilde{\boldsymbol{\theta}}_i^{(k)}, \rho_i^{(k)}) \ \forall i & \text{w.p. } 1 - \frac{1}{n} \end{cases}.$$

It can be seen immediately that Condition 2.5.2 holds. Computationally, we see that $\tilde{\boldsymbol{\theta}}_1^{(k)} = \ldots \tilde{\boldsymbol{\theta}}_n^{(k)}$ so we need only store a d-length "checkpoint" vector. Rather than storing $(g_1^{(k)}, \ldots, g_n^{(k)})$, we may simply store $\boldsymbol{\rho}^{(k)}$ (the weights) and $\bar{\boldsymbol{g}}^{(k)} := \sum_{i=1}^n \rho_i^{(k)} \boldsymbol{g}_i^{(k)}$ (aggregation). When computing line 7, we may simply recompute $\nabla r_{i_k}(\tilde{\boldsymbol{\theta}}_{i_k}^{(k)})$ using the checkpoint at a constant additional cost. Thus, the memory requirement for this step is $O(n + d)$.

2.5.3 Main Results

The convergence analysis given in the next section provides two different convergence rates depending on the choice of the shift cost ν. As shown in (2.21) (Section 2.4), this quantity is inversely proportional to the smoothness constant of the objective, thus directly affecting the conditioning of the objective. When $\nu = \Omega(nG^2/\mu)$ (which we refer to as the *large cost* setting), we see that the constant $(nG^2/\nu + L) = O(\mu + L)$, which is indeed the smoothness constant of the empirical risk minimization (ERM) objective $\frac{1}{n}\sum_{i=1}^{n} \ell_i(\cdot) + \frac{\mu}{2}\|\cdot\|$. Thus, for large values of ν, the rate should be similar to ERM. However, the bias in the gradient estimate remains even when ν is large, which contributes an additional condition number factor

$$\kappa_{\mathcal{P}} = n \max_{i\in[n], q\in\mathcal{P}} q_i$$

in addition to the usual $\kappa = (L+\mu)/\mu$ condition number, and preventing achieving the $(n+\kappa)\ln(1/\varepsilon)$ convergence rate. Note that $\kappa_{\mathcal{P}} = 1$ when $\mathcal{P} = \{\mathbf{1}/n\}$, thus making the large shift cost result a proper generalization of the analysis of ERM. Note that due to strong convexity, there exists a unique minimizer

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \tag{2.22}$$

which is referenced in the upcoming results. We first present the large cost convergence rate.

Theorem 2.5.1. *Suppose the shift cost satisfies*

$$\nu \geq 8nG^2/\mu.$$

Then, the sequence of iterates produced by Algorithm 1 with $\eta = 1/(12(\mu + M)\kappa_{\mathcal{P}})$ *achieves*

$$\mathbb{E}_0\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^\star\|_2^2 \leq (1 + \sigma_n^{-1} + \sigma_n^{-2})\exp(-t/\tau)\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2\,.$$

with

$$\tau = 2\max\{n, 24\kappa_{\mathcal{P}}(\kappa+1)\}.$$

To describe the rate for small costs, define $\delta := nG^2/(\mu\nu)$. The quantity δ captures the effect of the primal regularizer μ and dual regularizer ν as compared to the inherent continuity properties of the underlying losses.

Theorem 2.5.2. *Assume that $n \geq 2$ and that the shift cost $\nu \leq 8nG^2/\mu$. The sequence of iterates produced by Algorithm 1 with*

$$\eta = \frac{1}{16n\mu}\min\left\{\frac{1}{6[8\delta + (\kappa+1)\kappa_{\mathcal{P}}]}, \frac{1}{4\delta^2\max\{2n\kappa^2, \delta\}}\right\}$$

achieves

$$\mathbb{E}_0\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^\star\right\|_2^2 \leq \left(5 + 16\delta + \frac{6\kappa^2}{\sigma_n}\right)\exp\left(-t/\tau\right)\left\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\right\|_2^2$$

for

$$\tau = 32n\max\left\{6[8\delta + (\kappa+1)\kappa_{\mathcal{P}}], 4\delta^2\max\left\{2n\kappa^2, \delta\right\}, 1/16\right\}.$$

Clearly, the small cost rate is a high-degree polynomial in both n and κ, and does not reduce to the ERM rate. One intuition behind this is that the updates for the sequence $(\boldsymbol{q}^{(k)})_{t\geq 1}$ cannot be obviously cast as gradient descent steps or even non-Euclidean proximal steps; they are full maximization steps without any regularization to ensure that $\boldsymbol{q}^{(k+1)}$ is close to $\boldsymbol{q}^{(k)}$. One benefit of full maximization, however, is that there is no additional dual learning rate to tune.

Note that the guarantees in Theorem 2.5.1 and Theorem 2.5.2 also provide guarantees on the primal-dual gap, which is a common convergence criterion for min-max problems. To describe this quantity, consider the saddle point objective

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q}) = \langle \boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle - \nu D_f(\boldsymbol{q}\|\mathbf{1}/n) + \frac{\mu}{2}\left\|\boldsymbol{\theta}\right\|_2^2$$

so that $\mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{q}\in\mathcal{Q}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q})$. Due to strong convexity and strong concavity, there exists a unique saddle point $(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star)$ satisfying

$$\mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}) \le \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star) \le \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q}^\star)$$

for all $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{q} \in \mathcal{Q}$, where $\boldsymbol{\theta}^\star$ coincides with (2.22). Then, we may the *primal-dual gap* at any point $(\boldsymbol{\theta}, \boldsymbol{q})$ as

$$\mathrm{Gap}(\boldsymbol{\theta}, \boldsymbol{q}) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}) \ge 0, \tag{2.23}$$

which is uniquely minimized at $(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star)$. The following corollary can convert the distance-to-opt bounds into a bound on the primal-dual gap in the case of a smooth dual.

Corollary 2.5.1. *Assume that the rescaled χ^2-divergence $D_f(\boldsymbol{q}\|\mathbf{1}/n) = \frac{1}{2}\|\boldsymbol{q} - \mathbf{1}/n\|_2^2$ is used for the shift penalty. Then,*

$$\frac{\mu}{2}\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \le \mathrm{Gap}(\boldsymbol{\theta}^{(k)}, q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}^{(k)}))) \le \frac{1}{2}\left(L + \mu + \frac{nG^2}{\nu}\right)\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 .$$

Thus, the expected primal-dual gap $\mathbb{E}_k[\mathrm{Gap}(\boldsymbol{\theta}^{(k)}, q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}^{(k)})))]$ achieves the same convergence rate as $\mathbb{E}_k\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2$ under the conditions of Theorem 2.5.1 and Theorem 2.5.2, respectively.

Proof. The gap criterion decomposes as

$$\mathrm{Gap}(\boldsymbol{\theta}, \boldsymbol{q}) = \underbrace{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star)}_{\ge \frac{\mu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|_2^2} + \underbrace{\mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q})}_{\ge \frac{\nu}{2}\|\boldsymbol{q}-\boldsymbol{q}^\star\|_2^2},$$

where the inequalities follow by strong convexity in $\boldsymbol{\theta}$ and $\boldsymbol{q}$, respectively. This gives the lower bound by using that $\|q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}^{(k)})) - \boldsymbol{q}^\star\|_2^2 \ge 0$. The function $\boldsymbol{\theta} \mapsto \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star)$ is $(L+\mu)$-smooth and is minimized at $\boldsymbol{\theta}^\star$, indicating that for any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star) \le \frac{L+\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 .$$

Similarly, the function $\boldsymbol{q} \mapsto \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q})$ is ν-smooth and minimized at $\boldsymbol{q}^\star$, which

gives that for any $\boldsymbol{q} \in \mathcal{Q}$

$$\mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star) - \mathcal{L}(\boldsymbol{\theta}^\star, \boldsymbol{q}) \leq \frac{\nu}{2} \|\boldsymbol{q} - \boldsymbol{q}^\star\|_2^2 .$$

Then, we have by $(1/\nu)$-Lipschitzness (w.r.t. $\|\cdot\|_2$) of q^{opt} and G-Lipschitzness of each ℓ_i that

$$\begin{aligned} \left\|q^{\text{opt}}(\ell(\boldsymbol{\theta}^{(k)})) - \boldsymbol{q}^\star\right\|_2^2 &= \left\|q^{\text{opt}}(\ell(\boldsymbol{\theta}^{(k)})) - q^{\text{opt}}(\ell(\boldsymbol{\theta}^\star))\right\|_2^2 \\ &\leq \frac{1}{\nu^2} \|\ell(\boldsymbol{\theta}^{(k)}) - \ell(\boldsymbol{\theta}^\star)\|_2^2 \\ &\leq \frac{nG^2}{\nu^2} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 . \end{aligned}$$

Adding together both bounds gives the desired result. □

Note that the primal-dual gap, while a convenient theoretical criterion, is still not computable by the algorithm in practice and thus cannot be used as a stopping criterion. However, it can be upper bounded by computable quantities. In Section 3.7.1, we derive an upper bound for a more general convergence criterion, called the *smoothed duality gap*, which contains (2.23) as a special case. In software implementations, we use the smoothed duality gap as an online certificate, or a measure of suboptimality that can be computed by the algorithm as it runs to determine a finite termination point.

The next subsection contains the convergence analysis that yields the two theorems above. The analysis of both methods will follow similarly, but differ in one key step: their usage of a generic bias bound (namely, the upcoming Proposition 2.6.1).

2.6 Convergence Analysis

We first give the broad outline of the analysis. Some key steps are proved in the main text, whereas the remaining proofs can be found in Appendix A.2. The optimum of (2.18) is denoted $\boldsymbol{\theta}^\star$ and satisfies

$$\nabla(\boldsymbol{q}^{\star\top} r(\boldsymbol{\theta}^\star)) = 0, \text{ for } \boldsymbol{q}^\star = \arg\max_{\boldsymbol{q} \in \mathcal{Q}} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta}^\star) \rangle - \nu \operatorname{Reg}(\boldsymbol{q}). \tag{2.24}$$

The optimum $\boldsymbol{\theta}^\star$ (or equivalently the saddle point $(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star)$) exists due to the respective strong convexity and strong concavity of the objective and because $\mu, \nu > 0$.

When analyzing stochastic gradient methods in the smooth and strongly convex setting, we typically expand

$$\mathbb{E}_k\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2 = \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \underbrace{-2\eta \left\langle \mathbb{E}_k[\boldsymbol{v}^{(k)}], \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star \right\rangle}_{\text{descent term}} + \underbrace{\eta^2 \mathbb{E}_k\|\boldsymbol{v}^{(k)}\|_2^2}_{\text{noise term}}. \tag{2.25}$$

First, note that the expectation of the primal gradient estimate $\boldsymbol{v}^{(k)}$ is $\nabla r(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}$, where $\boldsymbol{q}^{(k)} = q^{\text{opt}}(\boldsymbol{l}^{(k)})$ and $\boldsymbol{l}^{(k)} \in \mathbb{R}^n$ denotes the estimate of the full loss vector. Applying standard convex inequalities to the descent term (in particular, Theorem A.1.3) yields

$$\begin{aligned} -\left\langle \nabla r(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star \right\rangle \leq & -\frac{\mu M}{\mu + M} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \\ & - \frac{1}{\mu + L} \sum_{i=1}^n q_i^{(k)} \|\nabla r_i(\boldsymbol{\theta}^{(k)}) - \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 \\ & - \left\langle \nabla r(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star \right\rangle. \end{aligned}$$

The first two terms on the right-hand side are negative, which provides a decrease in the expected distance-to-optimum value on every iterate. In the empirical risk minimization setting, the final term on the right-hand side would be zero due to the first-order optimality conditions on $\boldsymbol{\theta}^\star$, as $\boldsymbol{q}^{(k)} = \mathbf{1}/n$, implying the decrease of $\mathbb{E}_k\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2$ for small enough η. However, because $\boldsymbol{q}^{(k)}$ is a potentially non-uniform vector estimated using the table of losses $\boldsymbol{l}^{(k)}$ (as opposed to the loss vector $\ell(\boldsymbol{\theta}^\star)$ at optimum), the term $-\left\langle \nabla r(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star \right\rangle$ is non-zero. Additionally, this term is multiplied only by the learning rate η, instead of the noise terms, which are multiplied by η^2. Thus, this bias term must be bounded carefully in order to achieve the convergence guarantee under this regime. The convergence analysis will proceed in three parts.

1. First, we separately upper bound the descent and noise terms appearing in (2.25). In doing so, we will introduce four non-negative terms, denoted $S^{(k)}$, $T^{(k)}$, $U^{(k)}$, and $R^{(k)}$. These will all be incorporated into a *Lyapunov function* $V^{(k)}$, given by the "main term"

$\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2$ and a weighted combination of the other terms, namely,

$$V^{(k)} = \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 + c_1 S^{(k)} + c_2 T^{(k)} + c_3 U^{(k)} + c_4 R^{(k)}. \tag{2.26}$$

The colors are used to track the quantities easily across lemmas.

2. We then study the per-iterate evolution of the additional Lyapunov terms $S^{(k)}$, $T^{(k)}$, $U^{(k)}$, and $R^{(k)}$. We note that in the next step, for certain values of the parameter ν, we need only use $S^{(k)}$ and $T^{(k)}$ in the Lyapunov function.

3. We combine the upper bounds from the previous steps and set the constants (c_1, c_2, c_3, c_4) to achieve a single-iterate progress bound of the form

$$\mathbb{E}_k\left[V^{(k+1)}\right] \leq (1 - \tau^{-1})V^{(k)}$$

for $\tau > 1$, which is interpreted as a "half-life" parameter. The final rate will depend on an additional quantity: the condition number $\kappa_{\mathcal{Q}} := n \max_i \max_{q \in \mathcal{Q}} q_i$, which is equal to 1 in the empirical risk minimization setting.

When ν is large, the analysis is simplified significantly by setting $c_3 = c_4 = 0$. However, to achieve unconditional linear convergence of the algorithm under Assumption 2.4.1 using this technique, the terms $U^{(k)}$ and $R^{(k)}$ are needed. We emphasize that this proof technique applies to a class of algorithms, namely those that satisfy Condition 2.5.1 and Condition 2.5.2. The terms are defined as

$$S^{(k)} = \frac{1}{n}\sum_{i=1}^n \|n\rho_i^{(k)} \nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(k)}) - n q_i^* \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2, \quad T^{(k)} = \sum_{i=1}^n \|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\|_2^2,$$
$$U^{(k)} = \frac{1}{n}\sum_{j=1}^n \|\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_j^{(k)}\|_2^2, \quad R^{(k)} = 2\eta n (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top (\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star).$$

Though not included in the Lyapunov function, we will also introduce

$$Q^{(k)} = \frac{1}{n}\sum_{i=1}^n \|n q_i^{(k)} \nabla r_i(\boldsymbol{\theta}^{(k)}) - n q_i^\star \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2,$$

which appears as a term to be canceled. The three steps in the outline comprise the next three subsections. We start with the "main" Lyapunov term $\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2$, whose upper bound will expose the other terms.

2.6.1 Bounding the Distance-to-Optimum

Because the main innovation in this proof technique is the bounding of the descent term with bias, we highlight this step in a result of independent interest.

Proposition 2.6.1 (Bias Bound)**.** *Consider any $\boldsymbol{\theta} \in \mathbb{R}^d$, $\boldsymbol{l} \in \mathbb{R}^n$, and $\bar{\boldsymbol{q}} \in \mathcal{Q}$. Define*

$$\boldsymbol{q} := q^{\mathrm{opt}}(\boldsymbol{l}) = \arg\max_{\boldsymbol{p} \in \mathcal{Q}} \langle \boldsymbol{p}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{p}).$$

For any $\alpha_1 \in [0, 1]$,

$$\begin{aligned}
&- (\nabla r(\boldsymbol{\theta})^\top \boldsymbol{q} - \nabla r(\boldsymbol{\theta}^\star)^\top \bar{\boldsymbol{q}})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\
&\leq -(\boldsymbol{q} - \bar{\boldsymbol{q}})^\top (\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^\star)) - \frac{\mu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \\
&- \frac{\alpha_1}{4(M+\mu)\kappa_{\mathcal{Q}}} \frac{1}{n} \sum_{i=1}^n \|n q_i \nabla r_i(\boldsymbol{\theta}) - n q_i^\star \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{2\alpha_1 G^2}{\nu(M+\mu)\kappa_{\mathcal{Q}}} n (\boldsymbol{q} - \boldsymbol{q}^\star)^\top (\boldsymbol{l} - \boldsymbol{l}^\star).
\end{aligned}$$

Notice in Proposition 2.6.1 that when $\bar{\boldsymbol{q}} = \boldsymbol{q}^\star$, we have that $\nabla r(\boldsymbol{\theta}^\star)^\top \bar{\boldsymbol{q}}$ vanishes, as $\boldsymbol{q}^\star = q^{\mathrm{opt}}(\ell(\boldsymbol{\theta}^\star))$. Armed with Proposition 2.6.1 and a relatively simple bound on the noise, we can upper bound the distance-to-optimum.

Lemma 2.6.1 (Analysis of distance-to-optimum term)**.** *For any constants $\alpha_1 \in [0, 1]$ and*

$\alpha_2 > 0$, and any $\bar{\boldsymbol{q}} \in \mathcal{Q}$, we have that

$$\begin{aligned}\mathbb{E}_k\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2 \leq (1-\eta\mu)\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\ - 2\eta(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^\star)\bar{\boldsymbol{q}} \\ - \eta\left(\frac{\alpha_1}{2(M+\mu)\kappa_{\mathcal{Q}}} - \eta(1+\alpha_2)\right) Q^{(k)} + \eta^2(1+\alpha_2^{-1})S^{(k)} \\ + \frac{2\alpha_1 G^2}{\nu(M+\mu)\kappa_{\mathcal{Q}}} R^{(k)} - 2\eta(\boldsymbol{q}^{(k)} - \bar{\boldsymbol{q}})^\top(\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^\star)).\end{aligned}$$

Proof. Recall the expansion given in (2.25), which is:

$$\mathbb{E}_k\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2 = \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 - 2\eta\left\langle \mathbb{E}_k[\boldsymbol{v}^{(k)}], \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\rangle + \eta^2\mathbb{E}_k\|\boldsymbol{v}^{(k)}\|_2^2. \tag{2.27}$$

Observe that

$$\mathbb{E}_k[\boldsymbol{v}^{(k)}] = \sum_{i=1}^n \boldsymbol{q}_i^{(k)} \nabla r(\boldsymbol{\theta}^{(k)}) = \nabla r(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}$$

By Proposition 2.6.1 with $\boldsymbol{l} = \boldsymbol{l}^{(k)}$, $\boldsymbol{q} = \boldsymbol{q}^{(k)}$, $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, and multiplying by 2η, we have that

$$\begin{aligned}&- 2\eta(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)} \\ &\leq -2\eta(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^\star)\bar{\boldsymbol{q}} - 2\eta(\boldsymbol{q}^{(k)} - \bar{\boldsymbol{q}})^\top(\ell(\boldsymbol{\theta}^{(k)}) - \ell(\boldsymbol{\theta}^\star)) \\ &- \mu\eta\,\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 - \frac{\eta\alpha_1}{2(M+\mu)\kappa_{\mathcal{Q}}} Q^{(k)} + \frac{2\alpha_1 G^2}{\nu(M+\mu)\kappa_{\mathcal{Q}}} R^{(k)}.\end{aligned}$$

The noise term is bounded by applying Young's inequality with parameter $\alpha_2 > 0$ and the identity $\mathbb{E}\|X - \mathbb{E}[X]\|_2^2 = \mathbb{E}\|X\|_2^2 - \|\mathbb{E}[X]\|_2^2$ (i.e., the variance is upper bounded by the

second moment). Noting that $\sum_{i=1}^{n} q_i^\star \nabla r_i(\boldsymbol{\theta}^\star) = 0$, we get

$$
\begin{aligned}
&\mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] \\
&= \mathbb{E}_k\left[\|\boldsymbol{v}^{(k)} - \nabla(\boldsymbol{q}^{*\top} r)(\boldsymbol{\theta}^\star)\|_2^2\right] \\
&= \mathbb{E}_k\Big[\|nq_{i_k}^{(k)}\nabla r_{i_k}(\boldsymbol{\theta}^{(k)}) - nq_{i_k}^\star \nabla r_{i_k}(\boldsymbol{\theta}^\star) \\
&\quad + nq_{i_k}^\star \nabla r_{i_k}(\boldsymbol{\theta}^\star) - n\rho_{i_k}^{(k)}\nabla r_{i_k}(\tilde{\boldsymbol{\theta}}_{i_k}^{(k)}) - (\nabla(\boldsymbol{q}^{\star\top} r)(\boldsymbol{\theta}^\star) - \textstyle\sum_{i=1}^n \rho_i^{(k)}\nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(k)}))\|_2^2\Big] \\
&\leq (1+\alpha_2)\mathbb{E}_k\Big[\|nq_{i_k}^{(k)}\nabla r_{i_k}(\boldsymbol{\theta}^{(k)}) - nq_{i_k}^\star \nabla r_{i_k}(\boldsymbol{\theta}^\star)\|^2\Big] \\
&\quad + (1+\alpha_2^{-1})\mathbb{E}_k\Big[nq_{i_k}^\star \nabla r_{i_k}(\boldsymbol{\theta}^\star) - n\rho_{i_k}^{(k)}\nabla r_{i_k}(\tilde{\boldsymbol{\theta}}_{i_k}^{(k)}) - (\nabla(\boldsymbol{q}^{\star\top} r)(\boldsymbol{\theta}^\star) - \textstyle\sum_{i=1}^n \rho_i^{(k)}\nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(k)}))\|_2^2\Big] \\
&\leq (1+\alpha_2)\mathbb{E}_k\Big[\|nq_{i_k}^{(k)}\nabla r_{i_k}(\boldsymbol{\theta}^{(k)}) - nq_{i_k}^\star \nabla r_{i_k}(\boldsymbol{\theta}^\star)\|^2\Big] \\
&\quad + (1+\alpha_2^{-1})\mathbb{E}_k\Big[\|nq_{i_k}^\star \nabla r_{i_k}(\boldsymbol{\theta}^\star) - n\rho_{i_k}^{(k)}\nabla r_{i_k}(\tilde{\boldsymbol{\theta}}_{i_k}^{(k)})\|_2^2\Big].
\end{aligned}
$$

Substituting the definitions of $Q^{(k)}$ and $S^{(k)}$ we have finally that

$$\eta^2\mathbb{E}_k\|\boldsymbol{v}^{(k)}\|_2^2 \leq \eta^2(1+\alpha_2)Q^{(k)} + \eta^2(1+\alpha_2^{-1})S^{(k)}.$$

Combine the two displays above to get the desired result. □

2.6.2 Bounding the Evolution of the Lyapunov Function Terms

Notice that the steps above did not depend on the updates to the bias and variance reduction tables. When bounding the evolution of the Lyapunov function terms, we see the conditions being used. We display the proof of Lemma 2.6.2 as an example of the conditions being used, and defer the proofs of the more technical Lemma 2.6.3 and Lemma 2.6.4 to Appendix A.2. Recall that

$$Q^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\|nq_i^{(k)}\nabla r_i(\boldsymbol{\theta}^{(k)}) - nq_i^\star\nabla r_i(\boldsymbol{\theta}^\star)\|_2^2.$$

Lemma 2.6.2. *Given Condition 2.5.1 and Condition 2.5.2, we have that*

$$\mathbb{E}_k\left[S^{(k+1)}\right] = \frac{1}{n}Q^{(k)} + \left(1-\frac{1}{n}\right)S^{(k)},$$
$$\mathbb{E}_k\left[T^{(k+1)}\right] = \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 + \left(1-\frac{1}{n}\right)T^{(k)}.$$

Proof. First, apply Condition 2.5.2 using the collection of functions

$$h_i(\boldsymbol{x}) := \frac{1}{n}\|n\boldsymbol{x} - nq_i^\star \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2$$

in the $(*)$ line below to achieve

$$\begin{aligned}
&\mathbb{E}_k\left[S^{(k+1)}\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_k\left[\|n\rho_i^{(k+1)}\nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(k+1)}) - nq_i^*\nabla r_i(\boldsymbol{\theta}^\star)\|_2^2\right] \\
&\overset{(*)}{=} \frac{1}{n}\sum_{i=1}^n \left[\frac{1}{n}\|nq_i^{(k)}\nabla r_i(\boldsymbol{\theta}^{(k)}) - nq_i^\star\nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \left(1-\frac{1}{n}\right)\|n\rho_i^{(k)}\nabla r_{i_k}(\tilde{\boldsymbol{\theta}}_i^{(k)}) - nq_i^*\nabla r_i(\boldsymbol{\theta}^\star)\|_2^2\right] \\
&= \frac{1}{n}Q^{(k)} + \left(1-\frac{1}{n}\right)S^{(k)}.
\end{aligned}$$

Similarly, using Condition 2.5.1 with the functions

$$h_i(\boldsymbol{u},\boldsymbol{x}) := \|\boldsymbol{x} - \boldsymbol{\theta}^\star\|_2^2$$

in the $(\circ)$ line below to achieve

$$\begin{aligned}
\mathbb{E}_k\left[T^{(k+1)}\right] &= \sum_{i=1}^n \mathbb{E}_k\left[\|\hat{\boldsymbol{\theta}}_i^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2\right] \\
&\overset{(\circ)}{=} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 + \left(1-\frac{1}{n}\right)\sum_{i=1}^n \|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\
&= \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 + \left(1-\frac{1}{n}\right)T^{(k)},
\end{aligned}$$

completing the proof. □

The remaining two lemmas give the upper bounds for $U^{(k)}$ and $R^{(k)}$.

Lemma 2.6.3. *For any value of $\alpha_2 > 0$, we have that*

$$\mathbb{E}_k\left[U^{(k+1)}\right] \leq \eta^2(1+\alpha_2)Q^{(k)} + \eta^2(1+\alpha_2^{-1})S^{(k)} + \frac{\eta M^2}{\mu n}\left(1-\frac{1}{n}\right)T^{(k)} + \left(1-\frac{1}{n}\right)\frac{G^2}{2\nu\mu n}R^{(k)} + \left(1-\frac{1}{n}\right)U^{(k)}.$$

Lemma 2.6.4. *For any $\alpha_3 > 0$, it holds that*

$$\mathbb{E}_k\left[R^{(k+1)}\right] \leq 2\eta(\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top(\ell(\boldsymbol{\theta}^{(k)}) - \boldsymbol{l}^\star) + \left(1-\frac{1}{n}\right)R^{(k)} + \frac{\eta G^2 n}{2\nu}\alpha_3^{-1}T^{(k)} + \frac{2\eta G^2 n}{\nu}(1+\alpha_3)U^{(k)}.$$

2.6.3 Tuning Constants and Final Rate

Finally, we combine all previous upper bounds into an upper bound on $\mathbb{E}_k\left[V^{(k+1)}\right]$ in terms of $V^{(k)}$. For Theorem 2.5.1, we may in fact set $c_3 = c_4 = 0$, simplifying the analysis significantly. Of the remaining two terms $S^{(k)}$ and $T^{(k)}$, $S^{(k)}$ is particularly similar to the (only) Lyapunov term in the simplest analysis of the SAGA algorithm (see Bach [2024, Proposition 5.9]). However, in the case of SAGA for ERM, the gradient estimate remains unbiased. In our case, the additional bias requires us to incorporate $T^{(k)}$, which compares the elements in the table used to estimate the dual variables with those at the optimum. For Theorem 2.5.2, the additional terms are necessary to achieve the suboptimal convergence rate, but without any conditions on ν. One weakness of the analysis is the attempt at using a primal-only strategy (analyzing $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2$) in a primal-dual setting. In Chapter 3, we employ a strategy based on a primal-dual gap criterion which yields optimal rates in various problem classes. For the remainder of the chapter, we focus on practical questions regarding DRO, such as selecting the uncertainty set and shift cost, solving the maximization problem, and extensions to modern machine learning models such as neural networks.

2.6.4 Adaptive Sampling and Non-Uniformity

In Chapter 3, we expand on the optimization aspects of the problem and target a problem class that is more general than DRO. One of the essential ideas will be *adaptive sampling*, which we briefly motivate using the analysis above. For simplicity, consider the non-smooth variant of the objective (2.17), which can be written in the form of a maximum expectation

$$\max_{\boldsymbol{q}\in\mathcal{Q}} \sum_{i=1}^{n} q_i \ell(\boldsymbol{\theta}, \xi_i) = \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}_{i\sim\text{Unif}[n]} \left[n q_i \ell(\boldsymbol{\theta}, \xi_i) \right].$$

We view $nq_i = \beta(\xi_i)$ as the Radon-Nikodym derivative as alluded to before. Technically, the same objective can be considered by considering a weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)$ satisfying $\boldsymbol{w} > 0$ and $\sum_{i=1}^{n} w_i = 1$, and writing

$$\max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}_{i\sim\text{Unif}[n]} \left[n q_i \ell(\boldsymbol{\theta}, \xi_i) \right] = \max_{\boldsymbol{q}\in\mathcal{Q}} \mathbb{E}_{i\sim\boldsymbol{w}} \left[q_i \frac{\ell(\boldsymbol{\theta}, \xi_i)}{w_i} \right].$$

Minimizing either stochastic formulations in $\boldsymbol{\theta}$ yields the same solution. Furthermore, the stochastic gradient estimate $\boldsymbol{v}^{(k)}$ from line 7 of Algorithm 1 will retain the same bias properties if $i_k \sim \boldsymbol{w}$ and if we replace r_i with $r_i/(nw_i)$. While the analysis of the Lyapunov function terms will change due to the change in the sampling scheme, to give intuition as to why a non-uniform sampling scheme can yield complexity improvements, notice that the results in Theorem 2.5.1 and Theorem 2.5.2 depend on the maximum Lipschitz constant G and the maximum smoothness constant L. That is, if each ℓ_i is G_i-Lipschitz continuous and L_i-smooth, then necessarily, we must have $G = \max_i G_i$ and $L = \max_i L_i$. On the other hand, the functions $\{\ell_i/(nw_i)\}_{i=1}^{n}$ would be governed by Lipschitz and smoothness constants

$$G(\boldsymbol{w}) = \max_{i=1,\ldots,n} \frac{G_i}{nw_i} \text{ and } L(\boldsymbol{w}) = \max_{i=1,\ldots,n} \frac{L_i}{nw_i}.$$

For example, by setting $w_i \propto G_i + L_i$, we get that $\max\{G(\boldsymbol{w}), L(\boldsymbol{w})\} \leq \frac{1}{n}\sum_{i=1}^{n}(G_i + L_i)$, which is the sum of *average* Lipschitz constants, which can be up to a factor n smaller than the maximum Lipschitz constants in the most non-uniform setting. One challenge that exists here, as opposed to a standard primal-only SGD analysis (e.g., as in Needell et al. [2014]),

is that the weights $\boldsymbol{w}$ seemingly need to adapt to both Lipschitz constants and smoothness constants simultaneously. This will be handled using a primal-dual viewpoint in Chapter 3, in which we view $G_1, \ldots, G_n$ not as primal Lipschitz constants, but as *dual smoothness* constants.

2.7 Uncertainty Sets and Shift Costs

Regarding the uncertainty set $\mathcal{Q}$ and shift cost ν, two concrete questions exist: 1) how should we select them, and 2) once selected, how do we compute (2.19) (i.e., solve the dual problem)? We consider the second question and then the first, for the two canonical examples of spectral risk measures, and then for divergence balls.

In both cases, we provide a duality result that converts the maximization into a minimization problem which can be solved by a near-exact iterative procedure. These duality relations also provide guidance on selecting the shift cost ν. As before, we will use the notation $\mathrm{Reg}(\boldsymbol{q}) = D_f(q\|\mathbf{1}/n)$ to denote an f-divergence as described in Section 2.4. Additionally, while the duality results form the basis of the computational routines, we focus on mathematical details in this section and provide detailed implementation instructions in Appendix A.3.

2.7.1 Duality of Spectral Risk Measures

As in (2.16) we parametrize the uncertainty set as $\mathcal{Q} = \mathcal{Q}(\sigma)$ for spectrum σ.

Proposition 2.7.1. *Let $\boldsymbol{l} \in \mathbb{R}^n$ be a vector such that $l_1 \leq \ldots \leq l_n$. We have the dual relation*

$$\max_{\boldsymbol{q}\in\mathcal{Q}(\sigma)} \{\langle \boldsymbol{q}, \boldsymbol{l}\rangle - \nu\,\mathrm{Reg}(\boldsymbol{q})\} = \min_{\substack{\boldsymbol{z}\in\mathbb{R}^n \\ z_1\leq\ldots\leq z_n}} \sum_{i=1}^{n} \sigma_i z_i + \tfrac{1}{n}\, f^*\left(\tfrac{l_i - z_i}{\nu}\right). \tag{2.28}$$

The optima of both problems, denoted

$$z^{\mathrm{opt}}(\boldsymbol{l}) = \operatorname*{arg\,min}_{\substack{\boldsymbol{z}\in\mathbb{R}^n \\ z_1\leq\ldots\leq z_n}} \sum_{i=1}^{n} \sigma_i z_i + \tfrac{1}{n}\, f^*\left(\tfrac{l_i - z_i}{\nu}\right),\ q^{\mathrm{opt}}(\boldsymbol{l}) = \operatorname*{arg\,max}_{\boldsymbol{q}\in\mathcal{Q}(\sigma)} \langle \boldsymbol{q}, \boldsymbol{l}\rangle - \nu\,\mathrm{Reg}(\boldsymbol{q}),$$

are related as

$$q_i^{\mathrm{opt}}(\boldsymbol{l}) = \frac{1}{n}[f^*]'\left(\tfrac{l_i - z_i^{\mathrm{opt}}(\boldsymbol{l})}{\nu})\right). \tag{2.29}$$

Proof. Let $\iota_{\mathcal{Q}(\sigma)}$ denote the convex indicator function of the permutahedron $\mathcal{Q}(\sigma)$, which is 0 inside $\mathcal{Q}(\sigma)$ and $+\infty$ outside of $\mathcal{Q}(\sigma)$. Its convex conjugate is the support function of the permutahedron, i.e.,

$$\iota^*_{\mathcal{Q}(\sigma)}(\boldsymbol{l}) = \max_{\boldsymbol{q}\in\mathcal{Q}(\sigma)} \langle \boldsymbol{q}, \boldsymbol{l}\rangle \,.$$

For two closed convex functions h_1 and h_2 that are bounded from below, the convex conjugate of their sum is the infimal convolution of their conjugate [Hiriart-Urruty and Lemaréchal, 2004, Proposition 6.3.1]:

$$(h_1 + h_2)^*(\boldsymbol{x}) = \inf_{\boldsymbol{y}\in\mathbb{R}^d} \{h_1^*(\boldsymbol{y}) + h_2^*(\boldsymbol{x}-\boldsymbol{y})\} \,.$$

Provided that $h_1 + h_2$ is strictly convex, we have that the maximizer defining the conjugate is unique and equal to the gradient, that is,

$$\nabla(h_1 + h_2)^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{z}\in\mathbb{R}^d} \{\langle \boldsymbol{z}, \boldsymbol{x}\rangle - (h_1 + h_2)(\boldsymbol{z})\} \,.$$

If, in addition, $h_1^* + h_2^*$ is strictly convex and h_2^* is differentiable, we have, by Danskin's theorem [Bertsekas, 1997],

$$\nabla(h_1 + h_2)^*(\boldsymbol{x}) = \nabla h_2^*(\boldsymbol{x} - \boldsymbol{y}^\star(\boldsymbol{x})) \text{ for } \boldsymbol{y}^\star(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}\in\mathbb{R}^d} \{h_1^*(\boldsymbol{y}) + h_2^*(\boldsymbol{x}-\boldsymbol{y})\} \,.$$

Consider then $h_1(\boldsymbol{q}) = \iota_{\mathcal{Q}(\sigma)}(\boldsymbol{q})$ and $h_2(\boldsymbol{q}) = \mathrm{Reg}(\boldsymbol{q}) := \nu D_f(q\|\mathbf{1}_n/n)$. Provided that f is strictly convex with f^* strictly convex, D_f is also strictly convex with D_f^* strictly convex

since D_f just decomposes as a sum of f on independent variables. We have then

$$\begin{aligned}
\sup_{\boldsymbol{q}\in\mathcal{Q}(\sigma)} \{\langle \boldsymbol{q}, \boldsymbol{l}\rangle - \mathrm{Reg}(\boldsymbol{q})\} &= \sup_{\boldsymbol{q}\in\mathbb{R}^n} \left\{\langle \boldsymbol{q}, \boldsymbol{l}\rangle - (\iota_{\mathcal{Q}(\sigma)}(\boldsymbol{q}) + \mathrm{Reg}(\boldsymbol{q}))\right\} \\
&= (\iota_{\mathcal{Q}(\sigma)} + \mathrm{Reg})^*(\boldsymbol{l}) \\
&= \inf_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\iota^*_{\mathcal{Q}(\sigma)}(\boldsymbol{y}) + \mathrm{Reg}^*(\boldsymbol{l}-\boldsymbol{y})\right\} \\
&= \inf_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\max_{\boldsymbol{q}\in\mathcal{Q}(\sigma)} \langle \boldsymbol{q}, \boldsymbol{y}\rangle + \mathrm{Reg}^*(\boldsymbol{l}-\boldsymbol{y})\right\} \\
&= \inf_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\sum_{i=1}^n \sigma_i y_{(i)} + \mathrm{Reg}^*(\boldsymbol{l}-\boldsymbol{y})\right\}, \qquad (2.30)
\end{aligned}$$

where $y_{(1)} \le \ldots \le y_{(n)}$ are the ordered values of $\boldsymbol{y} \in \mathbb{R}^n$. Moreover, we have that

$$\arg\max_{\boldsymbol{q}\in\mathcal{Q}(\sigma)} \{\langle \boldsymbol{q}, \boldsymbol{l}\rangle - \mathrm{Reg}(\boldsymbol{q})\} = \nabla\,\mathrm{Reg}^*(\boldsymbol{l}-\boldsymbol{y}^\star(\boldsymbol{l})) \text{ for } \boldsymbol{y}^\star(\boldsymbol{l}) = \arg\min_{\boldsymbol{y}\in\mathbb{R}^n} \left\{\sum_{i=1}^n \sigma_i y_{(i)} + \mathrm{Reg}^*(\boldsymbol{l}-\boldsymbol{y})\right\}.$$

Since for any $\boldsymbol{x} \in \mathbb{R}^n$, Reg is decomposable into a sum of identical functions evaluated at the coordinates $(x_1, \ldots, x_n)$, that is, $\mathrm{Reg}(\boldsymbol{x}) = \sum_{i=1}^n \mathrm{Reg}_i(x_i)$, its convex conjugate is $\mathrm{Reg}^*(\boldsymbol{y}) = \sum_{i=1}^n \mathrm{Reg}_i^*(y_i)$. In our case, $\mathrm{Reg}_i(x_i) = \frac{\nu}{n} f(nx_i)$, so $\mathrm{Reg}_i^*(y_i) = (\nu/n) f^*(y_i/\nu)$.

Next, by convexity of each Reg_i^*, we have that if for scalars l_i, l_j, y_i, y_j such that $l_i \le l_j$ and $y_i \ge y_j$, then using Lemma A.1.2, we have that

$$\mathrm{Reg}_i^*(l_i - y_i) + \mathrm{Reg}_i^*(l_j - y_j) \ge \mathrm{Reg}_i^*(l_i - y_j) + \mathrm{Reg}_i^*(l_j - y_i).$$

Hence for $\boldsymbol{y}$ to minimize $\mathrm{Reg}^*(\boldsymbol{l}-\boldsymbol{y}) = \sum_{i=1}^n \mathrm{Reg}_i^*(l_i - y_i)$, the coordinates of $\boldsymbol{y}$ must be ordered as $\boldsymbol{l}$. That is, if π is an argsort for $\boldsymbol{l}$, s.t. $l_{\pi(1)} \le \ldots \le l_{\pi(n)}$, then $y_{\pi(1)} \le \ldots \le y_{\pi(n)}$. Since $\iota^*_{\mathcal{Q}(\sigma)}(\boldsymbol{y}) = \sum_{i=1}^n \sigma_i y_{(i)}$ does not depend on the ordering of y, the solution of (2.30) must also be ordered as $\boldsymbol{l}$ such that the dual problem (2.30) can be written as

$$\inf_{\substack{\boldsymbol{y}\in\mathbb{R}^n \\ y_1\le\ldots\le y_n}} \sum_{i=1}^n \sigma_i y_i + \frac{1}{n}\, f^*\left(\frac{l_i - y_i}{\nu}\right) = \min_{\substack{\boldsymbol{z}\in\mathbb{R}^n \\ z_1\le\ldots\le z_n}} \sum_{i=1}^n \sigma_i z_i + \frac{1}{n}\, f^*\left(\frac{l_i - z_i}{\nu}\right).$$

By differentiating both sides of (2.28) with respect to l_i and applying Dankin's theorem [Bertsekas, 1997], we have the relation (2.29). □

Computationally, we see that (2.28) is an instance of an isotonic regression problem, which can be solved by a version of the Pool Adjacent Violators (PAV) algorithm that is adapted to the generator function f. Thus, the dual problem can be solved exactly using an iterative algorithm that terminates in $O(n \log n)$ time.

2.7.2 Duality of f-Divergence Ball

We parametrize the uncertainty as $\mathcal{Q} = \mathcal{Q}(\rho)$ for radius ρ. Recall that Δ^{n-1} denotes the n-dimensional probability simplex. The feasible set can equivalently be written as

$$\mathcal{Q}(\rho) = \left\{ \boldsymbol{q} \in \Delta^{n-1} : \operatorname{Reg}(\boldsymbol{q}) \leq \rho \right\}. \tag{2.31}$$

Observe the following.

Proposition 2.7.2. *Let $\boldsymbol{l} \in \mathbb{R}^n$ be a vector. We have the dual relation*

$$\max_{\boldsymbol{q} \in \mathcal{Q}(\rho)} \left\{ \langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{q}) \right\} = \min_{t \geq \nu} \left\{ t \left((\operatorname{Reg} + \iota_{\Delta^{n-1}})^* \left(\tfrac{\boldsymbol{l}}{t} \right) + \rho \right) - \rho\nu \right\}. \tag{2.32}$$

The optima of both problems, denoted

$$t^{\mathrm{opt}}(\boldsymbol{l}) = \operatorname*{arg\,min}_{t \geq \nu} t \left((\operatorname{Reg} + \iota_{\Delta^{n-1}})^* \left(\tfrac{\boldsymbol{l}}{t} \right) + \rho \right), \; q^{\mathrm{opt}} = \operatorname*{arg\,max}_{\boldsymbol{q} \in \mathcal{Q}(\rho)} \langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{q}),$$

are related as

$$q_i^{\mathrm{opt}}(\boldsymbol{l}) = \nabla \left(\operatorname{Reg} + \iota_{\Delta^{n-1}} \right)^* \left(\boldsymbol{l} / t^{\mathrm{opt}}(\boldsymbol{l}) \right) = \operatorname*{arg\,max}_{\boldsymbol{q} \in \Delta^{n-1}} \left\{ \left\langle \boldsymbol{q}, l / t^{\mathrm{opt}}(\boldsymbol{l}) \right\rangle - \operatorname{Reg}(\boldsymbol{q}) \right\}. \tag{2.33}$$

Proof. First, using (2.31), we represent the left-hand side of (2.32) using a Lagrange multiplier λ applied to the extended real-valued objective

$$\begin{aligned}
\max_{\boldsymbol{q} \in \mathcal{Q}(\rho)} \langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{q}) &= \max_{\boldsymbol{q} \in \Delta^{n-1}} \min_{\lambda \geq 0} \langle \boldsymbol{q}, \boldsymbol{l} \rangle - (\nu + \lambda) \operatorname{Reg}(\boldsymbol{q}) + \lambda\rho \\
&= \min_{\lambda \geq 0} \max_{\boldsymbol{q} \in \Delta^{n-1}} \langle \boldsymbol{q}, \boldsymbol{l} \rangle - (\nu + \lambda) \operatorname{Reg}(\boldsymbol{q}) + \lambda\rho \\
&= \min_{\lambda \geq 0} (\nu + \lambda) \left(\max_{\boldsymbol{q} \in \Delta^{n-1}} \langle \boldsymbol{q}, \boldsymbol{l} / (\nu + \lambda) \rangle - \operatorname{Reg}(\boldsymbol{q}) \right) + \lambda\rho.
\end{aligned}$$

The second equality follows because $\mathbf{1}/n \in \mathcal{Q}(\rho)$, so Slater's condition is satisfied and strong duality holds. We then reparametrize the problem with $t = \nu + \lambda$, so that

$$\begin{aligned}\max_{\boldsymbol{q}\in\mathcal{Q}(\rho)} \langle \boldsymbol{q}, \boldsymbol{l}\rangle - \nu \operatorname{Reg}(\boldsymbol{q}) &= \min_{t\geq\nu} t\left(\max_{\boldsymbol{q}\in\Delta^{n-1}} \langle \boldsymbol{q}, \boldsymbol{l}/t\rangle - \operatorname{Reg}(\boldsymbol{q})\right) + \rho(t-\nu) \\ &= \min_{t\geq\nu} \left\{t\left((\operatorname{Reg} + \iota_{\Delta^{n-1}})^*\left(\tfrac{\boldsymbol{l}}{t}\right) + \rho\right) - \rho\nu\right\},\end{aligned}$$

which gives (2.32). Next, we differentiate both sides of (2.32) with respect to l and apply Dankin's theorem [Bertsekas, 1997] to achieve (2.33). □

To understand the computational implications of Proposition 2.7.2, we define the function

$$h_{\boldsymbol{l}}(t) = t\,(\operatorname{Reg} + \iota_{\Delta^{n-1}})^*\left(\tfrac{\boldsymbol{l}}{t}\right) + \rho(t-\nu),$$

and see that it is the sum of a perspective (in t) of a convex function on $\mathbb{R}^n$ and a linear function in t, hence convex. Its derivative is then given by

$$h'_{\boldsymbol{l}}(t) = -\tfrac{1}{t}\underbrace{\left\langle \nabla\,(\operatorname{Reg} + \iota_{\Delta^{n-1}})^*\left(\tfrac{\boldsymbol{l}}{t}\right), \tfrac{\boldsymbol{l}}{t}\right\rangle}_{\left\langle \boldsymbol{q}^\star_{\boldsymbol{l}}(t), \boldsymbol{l}/t\right\rangle} + \rho \text{ for } q^\star_{\boldsymbol{l}}(t) = \operatorname*{arg\,max}_{\boldsymbol{q}\in\Delta^{n-1}} \langle \boldsymbol{q}, \boldsymbol{l}/t\rangle - \operatorname{Reg}(\boldsymbol{q}). \tag{2.34}$$

Thus, if $h'_l(t)$ is computable, then $t^{\text{opt}}(\boldsymbol{l})$ can be computed by binary search. Thus, the strength of the duality result lies on the ability to compute $h'_{\boldsymbol{l}}(t)$ (or equivalently, $q^\star_{\boldsymbol{l}}(t)$).

The quintessential examples of this type of objective are when $\operatorname{Reg}_{\text{KL}}(\boldsymbol{q}) = \sum_{i=1}^n q_i \log(nq_i)$ (for the KL divergence) and $\operatorname{Reg}_{\chi^2}(\boldsymbol{q}) = n\,\|\boldsymbol{q} - \mathbf{1}/n\|_2^2$ (for the χ^2-divergence). In this case, we have that

$$q^\star_{\boldsymbol{l}}(t) = \operatorname*{arg\,max}_{\boldsymbol{q}\in\Delta^{n-1}} \langle \boldsymbol{q}, \boldsymbol{l}/t\rangle - \operatorname{Reg}_{\text{KL}}(\boldsymbol{q}) = \frac{e^{\boldsymbol{l}/t}}{\sum_{i=1}^n e^{l_i/t}}$$

and

$$q^\star_{\boldsymbol{l}}(t) = \operatorname*{arg\,max}_{\boldsymbol{q}\in\Delta^{n-1}} \langle \boldsymbol{q}, \boldsymbol{l}/t\rangle - \operatorname{Reg}_{\chi^2}(\boldsymbol{q}) = \operatorname*{arg\,min}_{\boldsymbol{q}\in\Delta^{n-1}} \|\boldsymbol{q} - \boldsymbol{l}/(2nt)\|_2^2 = \operatorname{proj}_{\Delta^{n-1}}(\boldsymbol{l}/(2nt)).$$

The first is a closed-form solution, whereas the second is a projection in ℓ_2-norm onto the probability simplex, which is studied for instance in Condat [2016]. As we saw in Section 2.3,

spectral risk measures have many more constraints on the likelihood ratio than divergence balls. This appears in the case of empirical measures as well, as the dual form of the maximization problem has n Lagrange multipliers for spectral risk measures but only 1 for f-divergence balls.

2.7.3 Setting Problem Parameters

Maximum Uncertainty Set Size The maximum size is governed by the spectrum σ for spectral risk measures and the radius ρ for divergence balls. In either case, the criterion used to determine the maximum size is whether the uncertainty set constraint is inactive (as compared to the standard probability simplex constraint). Recall that both uncertainty sets are permutation invariant, i.e., that $\boldsymbol{q} \in \mathcal{Q}$ implies that $\boldsymbol{q}_{\pi(\cdot)} = (q_{\pi(1)}, \ldots, q_{\pi(n)}) \in \mathcal{Q}$ for permutation π on $[n]$. We use the following equivalence

$$\mathcal{Q} = \Delta^{n-1} \text{ if and only if } \boldsymbol{e}_1 := (1, 0, \ldots, 0) \in \mathcal{Q}.$$

Observe the following special cases.

- For any spectral risk measure, $(1, 0, \ldots, 0) \in \mathcal{Q}$ if and only if $\sigma_n = 1$, which corresponds to the $(1 - 1/n)$-superquantile.

- For balls in χ^2-divergence, we measure the length of $n \left\| \boldsymbol{e}_1 - \mathbf{1}/n \right\|_2^2$:

$$n \left\| e_1 - \mathbf{1}/n \right\|_2^2 = n \left\| \boldsymbol{e}_1 \right\|_2^2 - 2n \left\langle \boldsymbol{e}_1, \mathbf{1}/n \right\rangle + n \left\| \mathbf{1}/n \right\|_2^2 = n - 1.$$

  Thus, we have that $\rho = n - 1$ is the upper bound for this uncertainty set.

- For balls in KL-divergence, we argue similarly by computing $\mathrm{KL}(\boldsymbol{e}_1 \| \mathbf{1}/n)$ (using the convention that $0 \log 0 = 0$):

$$\mathrm{KL}(\boldsymbol{e}_1 \| \mathbf{1}/n) = \log n,$$

  giving the upper bound $\rho = \log n$.

Minimum Uncertainty Set Size The user might wish to set the uncertainty set size to be at least large enough to achieve some form of robustness guarantee. One such guarantee is determined based on relating the solution of a distributionally robust optimization problem to the expected loss of the resulting procedure. In other words, we may use the minimum value of the distributionally robust objective to create a one-sided confidence interval for the population loss at a particular parameter value $\boldsymbol{\theta} \in \mathbb{R}^d$.

A central assumption that will be needed is that there exists some $B_{\boldsymbol{\theta}}$ such that for $\xi \sim P$,

$$\ell(\boldsymbol{\theta}, \xi) \leq B_{\boldsymbol{\theta}} \text{ almost surely under } P.$$

This is a reasonable assumption when the input space is bounded, as the loss may still be unbounded when w varies as well. Consider i.i.d. samples $\xi_1, \ldots, \xi_n \sim P$. We derive a method that will assign to any failure probability $\delta \in (0, 1]$ an uncertainty set $\mathcal{Q}_\delta \subseteq \mathbb{R}^{n+1}$ such that with probability at least $1 - \delta$,

$$\max_{\boldsymbol{q} \in \mathcal{Q}_\delta} \left\{ \sum_{i=1}^{n} q_i \ell(\boldsymbol{\theta}, \xi_i) + q_{n+1} B_{\boldsymbol{\theta}} \right\} \geq \mathbb{E}_{\xi \sim P} \left[\ell(\boldsymbol{\theta}, \xi) \right]. \tag{2.35}$$

While this is not a valid confidence bound for a random $\boldsymbol{\theta}$ that is dependent with $\{\xi_i\}_{i=1}^n$ (such as the minimizer of an empirical risk objective), it can still be computed on held-out data and be interpreted conditionally. This is indeed the result of Coppens and Patrinos [2023, Proposition III.1], although we provide an alternate proof below which uses only elementary tools.

To achieve this result, we first prove a technical lemma. Consider the partial order $\succeq$ on $\mathbb{R}^{n+1}$ given by $\boldsymbol{x} \succeq \boldsymbol{y}$ if and only if

$$\sum_{i=1}^{k} x_i \geq \sum_{i=1}^{k} y_i \ \forall k \in [n], \text{ and } \sum_{i=1}^{n+1} x_i = \sum_{i=1}^{n+1} y_i.$$

This is related to majorization but is not exactly the same, as we would be checking the inequality for the sorted versions of $\boldsymbol{x}$ and $\boldsymbol{y}$. Then, observe the following lemma.

Lemma 2.7.1. *Consider $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^{n+1}$ such that $\boldsymbol{x} \succeq \boldsymbol{y}$ and $z_1 \leq \ldots \leq z_{n+1}$, i.e., $\boldsymbol{z}$ is*

sorted. Then,

$$\boldsymbol{x}^\top \boldsymbol{z} \leq \boldsymbol{y}^\top \boldsymbol{z}.$$

Proof. We prove the result by induction. The case of $n = 0$ is satisfied trivially, as $x_1 = y_1 \implies x_1 z_1 = y_1 z_1$. Assume that the claim holds for vectors of size n; we show that it holds for vectors of size $n+1$. First, we define $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ such that $\bar{x}_i = x_i$ for $i \leq n-1$, and $\bar{x}_n = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n-1} x_i$. By construction, it holds that

$$\bar{\boldsymbol{x}} \succeq (y_1, \ldots, y_n).$$

Applying the inductive claim, we are assured that $\sum_{i=1}^{n} \bar{x}_i z_i \leq \sum_{i=1}^{n} y_i z_i$. Next, to achieve the desired result, write

$$\begin{aligned}
\sum_{i=1}^{n+1} x_i z_i &= \sum_{i=1}^{n} \bar{x}_i z_i + \left(\sum_{i=1}^{n-1} x_i - \sum_{i=1}^{n} y_i \right) z_n + x_n z_n + x_{n+1} z_{n+1} \\
&= \sum_{i=1}^{n} \bar{x}_i z_i + \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \right) z_n + x_{n+1} z_{n+1} \\
&\leq \sum_{i=1}^{n} y_i z_i + \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \right) z_n + x_{n+1} z_{n+1}.
\end{aligned}$$

Thus, the proof is complete if $\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i\right) z_n + x_{n+1} z_{n+1} \leq y_{n+1} z_{n+1}$. To see this, use $z_n \leq z_{n+1}$ and $\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \geq 0$ to achieve

$$\begin{aligned}
\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \right) z_n + x_{n+1} z_{n+1} &\leq \left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \right) z_{n+1} + x_{n+1} z_{n+1} \\
&= y_{n+1} z_{n+1}
\end{aligned}$$

where we use that $\sum_{i=1}^{n+1} x_i - \sum_{i=1}^{n+1} y_i = 0$ in the last step. □

We may now present our version of Coppens and Patrinos [2023, Proposition III.1].

Proposition 2.7.3. *Let U be uniformly distributed over Δ^n, the probability simplex in $n+1$ dimensions. Assume that $\ell(\boldsymbol{\theta}, \xi)$ is a continuous random variable on $\mathbb{R}$. Then, for uncertainty*

set $\mathcal{Q}_\delta \subseteq \Delta^n$, the event $\{U \succeq \boldsymbol{q}$ for some $\boldsymbol{q} \in \mathcal{Q}_\delta\}$ is contained within the event (2.35).

Proof. Define the random variables $L_i = \ell(\boldsymbol{\theta}, \xi_i)$ for $i \in [n]$, which are i.i.d. copies of $L := \ell(\boldsymbol{\theta}, \xi)$. Denote by F the (continuous) cumulative distribution function of L. Then, for particular realizations $l_{(1)} \leq \ldots \leq l_{(n)}$ of the order statistics of $(L_1, \ldots, L_n)$, define $l_{(0)} := 0$ and $l_{(n+1)} = B_{\boldsymbol{\theta}}$. Write

$$\begin{aligned}
\mathbb{E}_P[L] = \int_0^B x \, \mathrm{d}F(\boldsymbol{x}) &= \sum_{i=1}^{n+1} \int_{l_{(i-1)}}^{l_{(i)}} x \, \mathrm{d}F(\boldsymbol{x}) \\
&\leq \sum_{i=1}^{n+1} l_{(i)} \int_{l_{(i-1)}}^{l_{(i)}} \mathrm{d}F(\boldsymbol{x}) \\
&\leq \sum_{i=1}^{n+1} l_{(i)} F(l_{(i)}) - F(l_{(i-1)}).
\end{aligned}$$

Thus, for the random variables $(L_1, \ldots, L_n)$, we have that

$$\mathbb{E}_P[L] \leq \sum_{i=1}^{n+1} L_{(i)} U_i,$$

where $U_i = F(L_{(i)}) - F(L_{(i-1)})$. Because $L_1, \ldots, L_n$ are continuous, bounded random variables, it follows that U is uniformly distributed over Δ^{n+1}. We now show that

$$U \succeq \boldsymbol{q} \text{ for some } \boldsymbol{q} \in \mathcal{Q}(\Delta) \implies \max_{\boldsymbol{q} \in \mathcal{Q}(\Delta)} \sum_{i=1}^{n} q_i L_i + q_{n+1} L_{n+1} \geq \mathbb{E}_P[L].$$

Letting $\bar{\boldsymbol{q}}$ be such that $U \succeq \bar{\boldsymbol{q}}$ and because $L_{(1)} \leq \ldots \leq L_{(n+1)}$, we have by Lemma 2.7.1 that

$$\begin{aligned}
\sum_{i=1}^{n+1} L_{(i)} U_i &\leq \sum_{i=1}^{n+1} L_{(i)} \bar{\boldsymbol{q}}_i && \text{Lemma 2.7.1} \\
&\leq \max_{\boldsymbol{q} \in \mathcal{Q}(\Delta)} \sum_{i=1}^{n} q_i L_{(i)} + q_{n+1} L_{(n+1)} && \bar{\boldsymbol{q}} \in \mathcal{Q}(\Delta) \\
&= \max_{\boldsymbol{q} \in \mathcal{Q}(\Delta)} \sum_{i=1}^{n} q_i L_i + q_{n+1} L_{n+1},
\end{aligned}$$

where the final equality follows from permutation invariance of $\mathcal{Q}(\Delta)$. □

The significance of Proposition 2.7.3 is that given δ, one can parametrize the uncertainty set using a univariate parameter and compute the probability of $U \succeq \boldsymbol{q}$ via simulation. The size for which the probability matches $1-\delta$ can be made the minimum size. This is especially easy for spectral risk measures, given the following result.

Lemma 2.7.2. *Let $\mathcal{Q}(\sigma)$ denote a spectral risk measure uncertainty set with spectrum σ. Then,*

$$\mathbb{P}\left[U \succeq \boldsymbol{q} \text{ for some } \boldsymbol{q} \in \mathcal{Q}(\sigma)\right] = \mathbb{P}\left[U \succeq \sigma\right].$$

Proof. We prove the result by showing the equivalence of the corresponding events. First, because $\sigma \in \mathcal{Q}(\sigma)$, it holds trivially that

$$U \succeq \sigma \implies U \succeq \boldsymbol{q} \text{ for some } \boldsymbol{q} \in \mathcal{Q}(\sigma).$$

Conversely, consider $\boldsymbol{q} \in \mathcal{Q}(\sigma)$ such that $U \succeq \boldsymbol{q}$. We can write $\boldsymbol{q}$ as

$$\boldsymbol{q} = \sum_{\pi \in \Pi_{n+1}} \lambda_\pi \sigma_\pi,$$

where Π_{n+1} is the set of permutations on $[n+1]$, $\sum_{\pi \in \Pi_{n+1}} \lambda_\pi = 1$, each $\lambda_\pi \geq 0$, and

$$\sigma_\pi := \left(\sigma_{\pi(1)}, \ldots, \sigma_{\pi(n+1)}\right).$$

Then, for any $k \in [n+1]$, it holds that

$$\sum_{i=1}^{k} U_i \geq \sum_{\pi \in \Pi_{n+1}} \lambda_\pi \sum_{i=1}^{k} \sigma_{\pi(i)} \geq \min_{\pi \in \Pi_{n+1}} \sum_{i=1}^{k} \sigma_{\pi(i)} = \sum_{i=1}^{k} \sigma_i,$$

where we use in the last step that $\sigma_1 \leq \ldots \leq \sigma_{n+1}$. Thus, $U \succeq \boldsymbol{q}$, completing the proof. □

The probabilities can be computed exponentially quickly in the number of samples, rendering the simulation tractable.

Maximum Shift Cost This question is answered by considering uniform approximation bounds between the smoothed ($\nu > 0$) and unsmoothed ($\nu = 0$) objectives. Because the $\nu > 0$ case is typically considered a tool to improve optimization convergence rates in theory and in practice, we would like it to be as small as possible. Noting that

$$\max_{\boldsymbol{q}\in\mathcal{Q}} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle \geq \max_{\boldsymbol{q}\in\mathcal{Q}} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle - \nu \operatorname{Reg}(\boldsymbol{q}) \geq \max_{\boldsymbol{q}\in\mathcal{Q}} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle - \nu \max_{\boldsymbol{q}'\in\mathcal{Q}} \operatorname{Reg}(\boldsymbol{q}'),$$

we compute the $\max_{\boldsymbol{q}'\in\mathcal{Q}} \operatorname{Reg}(\boldsymbol{q}')$ for both uncertainty set classes. Because $\operatorname{Reg}(\boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^n f(nq_i)$ is a strongly convex, rotationally symmetric function centered at $\mathbf{1}/n$, over $\mathcal{Q}(\sigma)$ the maximum is attained at any vertex, yielding $\max_{\boldsymbol{q}'\in\mathcal{Q}} \operatorname{Reg}(\boldsymbol{q}') = \operatorname{Reg}(\sigma)$. For divergence balls, we have by definition that $\max_{\boldsymbol{q}'\in\mathcal{Q}} \operatorname{Reg}(\boldsymbol{q}') = \rho$. This, if we aim to achieve suboptimality ε for the $\nu = 0$ objective by optimizing the $\nu > 0$ objective, then setting

$$\nu_{\max} = \begin{cases} \frac{\varepsilon}{4\operatorname{Reg}(\sigma)} & \text{if } \mathcal{Q} = \mathcal{Q}(\sigma) \text{ (spectral risk measure)} \\ \frac{\varepsilon}{4\rho} & \text{if } \mathcal{Q} = \mathcal{Q}(\rho) \text{ (divergence ball)} \end{cases}$$

contributes $\varepsilon/2$ suboptimality by smoothing.

Minimum Shift Cost As before, we address this question separately for spectral risk measures and for f-DRO. To establish notation, let

$$\mathcal{L}^{(\nu)}(\boldsymbol{\theta}) := \max_{\boldsymbol{q}\in\mathcal{Q}} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle - \nu \operatorname{Reg}(\boldsymbol{q}) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2$$

denote the primal objective with shift cost $\nu \geq 0$ with respect to a particular uncertainty set $\mathcal{Q}$ and f-divergence penalty Reg (now indexed with ν). In both cases, we derive a computationally verifiable condition on w under which the following holds: for all $\nu, \bar{\nu} \geq 0$ small enough, we have that $\nabla\mathcal{L}^{(\nu)}(\boldsymbol{\theta}) = \nabla\mathcal{L}^{(\bar{\nu})}(\boldsymbol{\theta})$.

Proposition 2.7.4. *Let $\mathcal{Q} = \mathcal{Q}(\sigma)$ be a spectral risk measure uncertainty set with spectrum σ. Let $z^{(\nu)}$ denote the optimum map z^{opt} of (2.28) under a shift cost $\nu \geq 0$. If for some*

$\bar{\nu} \geq 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ we have that

$$z_i^{(\bar{\nu})}(\ell(\boldsymbol{\theta})) = \ell_{(i)}(\boldsymbol{\theta}) - \bar{\nu} f'(n\sigma_i), \tag{2.36}$$

then

$$\nabla \mathcal{L}^{(\bar{\nu})}(\boldsymbol{\theta}) = \nabla \mathcal{L}^{(\nu)}(\boldsymbol{\theta}) \text{ for all } \nu \in (0, \bar{\nu}] \text{ and } \nabla \mathcal{L}^{(\bar{\nu})}(\boldsymbol{\theta}) \in \partial \mathcal{L}^{(0)}(\boldsymbol{\theta}).$$

Furthermore, the condition (2.36) is equivalent to

$$\bar{\nu}\left(f'(n\sigma_{i+1}) - f'(n\sigma_i)\right) \leq \ell_{(i+1)}(\boldsymbol{\theta}) - \ell_{(i)}(\boldsymbol{\theta}) \text{ for } i = 1, \ldots, n. \tag{2.37}$$

Proof. The left-hand side of (2.36) results from taking the derivative of the right-hand side of (2.28). Thus, (2.36) is equivalent to the unconstrained solution of (2.28) being feasible in $\boldsymbol{z}$. Without loss of generality, assume that $\ell(\boldsymbol{\theta})$ is sorted in non-decreasing order. It follows that

$$z_i^{(\nu)}(\ell(\boldsymbol{\theta})) = \ell_i(\boldsymbol{\theta}) - \nu f'(n\sigma_i) \tag{2.38}$$

for all $\nu \leq \bar{\nu}$, as the right-hand side remains feasible for the problem (2.28) (in particular, due to the monotonicity of $z_i^{(\bar{\nu})}(\ell(\boldsymbol{\theta}))$). Defining similarly $q_i^{\text{opt}} = q_i^{(\nu)}$, it follows from Proposition 2.7.1 that

$$q_i^{(\nu)}(\ell(\boldsymbol{\theta})) = \frac{1}{n}[f^*]'\left(\frac{\ell_i(\boldsymbol{\theta}) - z_i^{(\nu)}(\ell(\boldsymbol{\theta}))}{\nu}\right) = \frac{1}{n}[f^*]'\left(f'(n\sigma_i)\right) = \sigma_i,$$

which is independent of ν and $\bar{\nu}$. First, consider $\nu > 0$, and notice that

$$\begin{aligned}
\nabla \mathcal{L}^{(\bar{\nu})}(\boldsymbol{\theta}) &= \nabla \ell(\boldsymbol{\theta})^\top q^{(\bar{\nu})}(\ell(\boldsymbol{\theta})) + \mu w \\
&= \nabla \ell(\boldsymbol{\theta})^\top q^{(\nu)}(\ell(\boldsymbol{\theta})) + \mu w \\
&= \nabla \mathcal{L}^{\nu}(\boldsymbol{\theta})
\end{aligned}$$

For the case of $\nu = 0$, note that $\ell(\boldsymbol{\theta})^\top \sigma + \mu \boldsymbol{\theta}^\star_{\bar{\nu}}$ is a subgradient of $\mathcal{L}^{(0)}$ (see Mehta et al. [2023, Proposition 2]). The second claim follows from the feasibility (in particular, the

monotonicity) of $z^{(\nu)}(\ell(\boldsymbol{\theta}))$. □

Proposition 2.7.5. *Let $\mathcal{Q} = \mathcal{Q}(\rho)$ be an f-divergence uncertainty set with radius ρ. Let $t^{(\nu)}$ denote the optimum of t^{opt} of (2.32) under a shift cost $\nu \geq 0$. If for some $\bar{\nu} \geq 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$, we have that*

$$h'_{\ell(\boldsymbol{\theta})}(t^{(\bar{\nu})}(\ell(\boldsymbol{\theta}))) = 0, \tag{2.39}$$

for h'_l defined in (2.34), then then

$$\nabla\mathcal{L}^{(\bar{\nu})}(\boldsymbol{\theta}) = \nabla\mathcal{L}^{(\nu)}(\boldsymbol{\theta}) \text{ for all } \nu \in (0, \bar{\nu}] \text{ and } \nabla\mathcal{L}^{(\bar{\nu})}(\boldsymbol{\theta}) \in \partial\mathcal{L}^{(0)}(\boldsymbol{\theta}).$$

Furthermore, the condition (2.39) is equivalent to the solution of $h'_{\ell(\boldsymbol{\theta})}(t) = 0$ in t being greater than or equal to $\bar{\nu}$.

Proof. As in the proof of Proposition 2.7.4, the left-hand side of (2.39) results from taking the derivative of the right-hand side of (2.32). Thus, (2.39) is equivalent to the unconstrained solution of (2.32) being feasible in t. The remainder of the proof follows identically to that of Proposition 2.7.4. □

Before showing numerical benchmarks, we also comment that there are numerous statistical ideas that appear in this chapter that have connections to broader literature, including the method of empirical likelihood. We clarify the similarities and differences between our work and these areas in the next section.

2.8 Comparison to Broader Literature

As reflected by the diversity of topics in this chapter, distributionally robust optimization can be studied from many statistical and computational angles. Accordingly, there are several opportunities to contextualize our results with respect to other subfields within statistics. These include quantitative finance, nonparametric statistics, classical viewpoints on distribution shift, and existing stochastic variance-reduced algorithms for empirical risk minimization.

Risk Measures In Section 2.1, we introduced closed balls f-divergence balls and spectral risk measures as the major classes of distributionally robust objectives. However, spectral risk measures originated from a parallel line of work on *risk measures*, with theoretical roots in convex analysis and applications in econometrics and finance [Rockafellar and Uryasev, 2013, Föllmer and Schied, 2002, He et al., 2022, Ben-Tal and Teboulle, 2007]. Risk measures are functionals of a real-valued random variable (such as the expectation) that quantify some notion of "tail error". Functionals that satisfied particular axiomatic properties were deemed *coherent* [Artzner et al., 1999], and among them were the class of spectral risk measures [Acerbi and Tasche, 2002]. In other words, spectral risk measures arose conceptually to quantify the tailedness of a random variable but can be viewed in the framework of DRO based on the variational arguments in the upcoming Section 2.3. On the other hand, the etymology of f-divergence uncertainty sets can be seen in reverse: starting as a natural framework for DRO, modern results demonstrated that these also can be represented as measures of tail error (albeit not in closed form). Indeed, by Shapiro [2017, Section 3.2], we have that

$$\sup_{Q \ll P : D_f(Q\|P) \leq \rho} \mathbb{E}_Q \left[\ell(\boldsymbol{\theta}, \xi)\right] = \inf_{\lambda \geq 0, \gamma \in \mathbb{R}} \mathbb{E}_P \left[\lambda f^* \left(\frac{\ell(\boldsymbol{\theta}, \xi) - \gamma}{\lambda}\right) + \lambda\rho + \gamma\right]. \tag{2.40}$$

We may actually relate the dual-minimization form to another notion of risk measure, which will help interpret (2.40). Let $\psi : \mathbb{R} \to \mathbb{R}_+$ be a convex function and define

$$\mathrm{oce}_P^{\psi}(X) = \inf_{\gamma \in \mathbb{R}} \mathbb{E}_P \left[\psi\left(X - \gamma\right)\right] + \gamma$$

as the *optimized certainty equivalent (OCE)* with *disutility function* ψ [Leqi et al., 2019, Lee et al., 2020]. This risk measure is often interpreted in quantitative finance as an optimal allocation between current and future losses; the investor is subject to a random loss X and may choose to incur γ of it while subjecting the remaining unknown portion to the function ψ, which represents the value of some loss/return in the future. Then, proceeding from

(2.40), we have that

$$\inf_{\lambda\geq 0,\gamma\in\mathbb{R}} \mathbb{E}_P\left[\lambda f^*\left(\frac{\ell(\boldsymbol{\theta},\xi)-\gamma}{\lambda}\right)+\lambda\rho+\gamma\right] = \inf_{\lambda\geq 0,\gamma\in\mathbb{R}} \mathbb{E}_P\left[\psi_\lambda\left(\ell(\boldsymbol{\theta},\xi)-\gamma\right)\right]+\gamma$$
$$= \inf_{\lambda\geq 0} \operatorname{oce}_P^{\psi_\lambda}(\ell(\boldsymbol{\theta},\xi)),$$

for $\psi_\lambda(x) = \lambda f^*(x/\lambda) + \lambda\rho$. Thus, the f-DRO objective is the smallest OCE over a class of disutility functions $\{\psi_\lambda : \lambda \geq 0\}$, which themselves may be interpreted as various investments.

A number of recent works study L-risks in the form of (2.12), with a focus on statistical properties. The works Khim et al. [2020] and Maurer et al. [2021] provide classical statistical learning bounds for L-risk objectives, and the latter focuses on unsupervised tasks like clustering. Holland and Mehdi Haress [2022] present a derivative-free learning procedure for general L-risk problems in the fully stochastic/streaming setting. As for optimizing these risk measures, Fan et al. [2017] and Kawaguchi and Lu [2020] study batch and stochastic algorithms, respectively, for the "average top-k" loss, which is exactly equivalent to the superquantile. We instead focus on developing incremental algorithms, akin to those for ERM [Mairal, 2014, Le Roux et al., 2012, Shalev-Shwartz and Zhang, 2013, Johnson and Zhang, 2013, Defazio et al., 2014], which apply to all L-risks.

Empirical Likelihood We review the method of empirical likelihood (EL) from nonparametric statistics [Owen, 1990] while comparing it to distributionally robust optimization (DRO) from both the statistical and computational perspectives. We show the following.

- The distributionally robust objective with the reverse-KL ball uncertainty set is equal to the upper limit of an empirical likelihood confidence region for mean estimation. The confidence level can be expressed in terms of the radius of the ball.

- The maximum over the weights q occurs under different constraints in DRO versus EL. In EL, the constraint is that the distribution induced by q has a particular mean (an affine constraint), as opposed to the small divergence constraint in DRO. In other

words, DRO maximizes a linear objective under a nonlinear constraint, whereas EL maximizes a nonlinear objective with a linear constraint.

EL is used to derive nonparametric confidence intervals for an arbitrary functional of a probability measure. For now, we consider the special case mean estimation from i.i.d. real-valued observations $X_1, \ldots, X_n$ with cumulative distribution function F. Rather than specifying a distributional class for F, the idea is to construct a likelihood ratio of a distribution with mean μ dominated by the empirical CDF F_n itself and reject the proposed mean μ when this likelihood ratio is small. Letting $\boldsymbol{q} = (q_1, \ldots, q_n)$ be the weights on the samples, we consider *all* valid likelihood ratios for such a distribution, and take their highest value, which we denote by $\mathcal{R}(\mu)$.

$$\mathcal{R}(\mu) := \max \left\{ \prod_{i=1}^{n} (nq_i) : \sum_{i=1}^{n} q_i X_i = \mu, \boldsymbol{q} \in \Delta^{n-1} \right\}. \tag{2.41}$$

The objective is log-strictly concave and the feasible set is compact in $\mathbb{R}^n$, so a unique maximizer assuredly exists. The confidence or acceptance region for μ is then given by

$$\{\mu : \mathcal{R}(\mu) \geq r\},$$

where r is the to-be-calibrated threshold on the likelihood ratio. To address the first bullet, observe the following.

Proposition 2.8.1. *Let $f(t) = -\log(t)$ for $t > 0$. Then,*

$$\max \left\{ \sum_{i=1}^{n} q_i X_i : D_f(\boldsymbol{q} \| \mathbf{1}/n) \leq \rho, \boldsymbol{q} \in \Delta^{n-1} \right\} = \max \left\{ \mu \in \mathbb{R} : \mathcal{R}(\mu) \geq e^{-n\rho} \right\}.$$

Proof. We start with the quantity on the left-hand side and derive the right-hand side as a result. For notational ease, set $T_n(\boldsymbol{q}) := \sum_{i=1}^{n} q_i X_i$. Introduce the auxiliary variable μ to

the DRO problem:

$$\begin{aligned}
&\max\left\{T_n(\boldsymbol{q}) : D_f(\boldsymbol{q}\|\mathbf{1}/n) \le \rho, \boldsymbol{q} \in \Delta^{n-1}\right\} \\
&= \max_{\mu\in\mathbb{R}} \max\left\{\mu : T_n(\boldsymbol{q}) = \mu, D_f(\boldsymbol{q}\|\mathbf{1}/n) \le \rho, \boldsymbol{q} \in \Delta^{n-1}\right\} \\
&= \max_{\mu\in\mathbb{R}} \max_{\substack{\boldsymbol{q}\in\Delta^{n-1} \\ T_n(\boldsymbol{q})=\mu}} \min_{\xi\ge 0} \mu - \xi(D_f(\boldsymbol{q}\|\mathbf{1}/n) - \rho),
\end{aligned}$$

where we introduced the Lagrange multiplier ξ. Because of the existence of a primal-dual strictly feasible point $(\boldsymbol{q}, \xi) = (\mathbf{1}/n, 1)$, we apply Slater's condition and strong duality to the inner max-min problem to achieve

$$\begin{aligned}
&\max_{\boldsymbol{q}\in\Delta^{n-1}} \{T_n(\boldsymbol{q}) : D_f(\boldsymbol{q}\|\mathbf{1}/n) \le \rho\} \\
&= \max_{\mu\in\mathbb{R}} \min_{\xi\ge 0} \max_{\substack{\boldsymbol{q}\in\Delta^{n-1} \\ T_n(\boldsymbol{q})=\mu}} \mu - \xi(D_f(\boldsymbol{q}\|\mathbf{1}/n) - \rho) \\
&= \max_{\mu\in\mathbb{R}} \min_{\xi\ge 0} \max_{\substack{\boldsymbol{q}\in\Delta^{n-1} \\ T_n(\boldsymbol{q})=\mu}} \mu - \frac{\xi}{n}\left(-\sum_{i=1}^{n}\log(nq_i) - n\rho\right) \\
&= \max_{\mu\in\mathbb{R}} \min_{\xi\ge 0} \mu + \frac{\xi}{n}\left(\max_{\substack{\boldsymbol{q}\in\Delta^{n-1} \\ T_n(\boldsymbol{q})=\mu}} \sum_{i=1}^{n}\log(nq_i) + n\rho\right) \\
&= \max_{\mu\in\mathbb{R}} \min_{\xi/n\ge 0} \mu - \frac{\xi}{n}\left(-\log(\mathcal{R}(\mu)) - n\rho\right).
\end{aligned}$$

Next, we interpret ξ/n as a Lagrange multiplier for the constraint $-\log(\mathcal{R}(\mu)) \le n\rho$, which is equivalent to $\mathcal{R}(\mu) \ge e^{-n\rho}$. Rewriting the above in constraint form completes the argument. □

Next, we consider the computational aspect. As stated before, DRO maximizes a linear objective under a nonlinear constraint, whereas EL maximizes a nonlinear objective with a

linear constraint:

$$\max_{\boldsymbol{q}\in\Delta^{n-1}} \{T_n(\boldsymbol{q}) : D_f(\boldsymbol{q}\|\mathbf{1}/n) \le \rho\}, \tag{DRO}$$

$$\max_{\boldsymbol{q}\in\Delta^{n-1}} \left\{\sum_{i=1}^{n} \log(nq_i) : T_n(\boldsymbol{q}) = \mu\right\}. \tag{EL}$$

Furthermore, EL does not compute a primal-optimal point in closed form; instead, it computes a saddle point of the Lagrangian when introducing a Lagrange multiplier for the constraint $T_n(\boldsymbol{q}) = \mu$. If λ denotes this multiplier, then we have that

$$q_i = \frac{1}{n}\frac{1}{1+\lambda(X_i-\mu)} \;\forall i \in \{1,\dots,n\}.$$

The proof of the univariate mean empirical likelihood theorem, or ELT [Owen, 2001, Theorem 2.2], i.e., the asymptotic χ^1_1 distribution of $-2\log\mathcal{R}(\mu)$ does actually compute λ (hence, does not have a closed form for q) but shows that $\lambda = O_p(n^{-1/2})$, from which asymptotic arguments follow. A potentially different computational scheme is required for every choice of T_n, and a subsequent analysis of the Lagrange multiplier. That is why there are different ELT variants for each function (quantiles, variance, regression parameters, etc.).

DRO for In-Distribution Generalization While typically motivated through considerations such as distribution shift from P to Q, there are applications of DRO methods even for traditional generalization guarantees from P_n to P. In particular, consider the two distributional parameters (which we assume to uniquely exist for the sake of discussion) as

$$\boldsymbol{\theta}^\star_{\text{in}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \underbrace{\mathbb{E}_{\xi\sim P}\left[\ell(\boldsymbol{\theta},\xi)\right]}_{\mathcal{R}_{\text{in}}(\boldsymbol{\theta},P)}, \text{ and } \boldsymbol{\theta}^\star_{\text{out}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \underbrace{\sup_{Q\in\mathcal{Q}(P)} \mathbb{E}_{\xi\sim Q}\left[\ell(\boldsymbol{\theta},\xi)\right]}_{\mathcal{R}_{\text{out}}(\boldsymbol{\theta},P)}.$$

The statistical and optimization techniques developed for DRO are often in service of estimating $\boldsymbol{\theta}^\star_{\text{out}}$. We now describe scenarios in which these same techniques can in fact be used to achieve desirable guarantees for the estimation of $\boldsymbol{\theta}^\star_{\text{in}}$ as well. In the statements below, all

probabilities are taken with respect to an independently drawn sample $\xi_1, \ldots, \xi_n \sim P$, with

$$\boldsymbol{\theta}_n \equiv \boldsymbol{\theta}_n(\xi_1, \ldots, \xi_n) := \arg\min_{\boldsymbol{\theta} \in \Theta} \underbrace{\sup_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_{\xi \sim Q} \left[\ell(\boldsymbol{\theta}, \xi) \right]}_{\mathcal{R}_{\text{in}}(\boldsymbol{\theta}, P_n)}.$$

The first type of guarantee concerns *certification*, or claiming that with high probability,

$$\max_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_{\xi \sim Q} \left[\ell(\boldsymbol{\theta}_n, \xi) | P_n \right] \geq \mathbb{E}_{\xi \sim P} \left[\ell(\boldsymbol{\theta}_n, \xi) | P_n \right]. \tag{2.42}$$

In other words, the optimal value of the distributionally robust empirical risk is, with high probability, an upper bound for the expected risk of the minimizer conditioned on the training data. Thus, while $\boldsymbol{\theta}_n$ is computed with a distributionally robust empirical objective, we still consider its performance on the data-generating distribution. Clearly, we may satisfy (2.42) if the sufficient condition

$$P \in \mathcal{Q}(P_n). \tag{2.43}$$

holds. However, this is only possible when $\mathcal{Q}(P_n)$ contains distributions that may not be absolutely continuous with respect to P_n. This precludes the likelihood ratio-based framework from Section 2.1. For these, we turn to methods that define $\mathcal{Q}(P_n)$ using a ball in the Wasserstein metric W. In this case, if ε is the radius of the Wasserstein ball, then (2.43) follows if and only if $\mathsf{W}(P, P_n) \leq \varepsilon$. Thus, the guarantees will follow by appealing to concentration results of $\mathsf{W}(P, P_n)$ (see Kuhn et al. [2019]). In fact, (2.43) implies the stronger condition that (2.42) holds uniformly, i.e.,

$$\max_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_{\xi \sim Q} \left[\ell(\boldsymbol{\theta}, \xi) \right] \geq \mathbb{E}_{\xi \sim P} \left[\ell(\boldsymbol{\theta}, \xi) \right] \ \forall \boldsymbol{\theta} \in \Theta. \tag{2.44}$$

Thus, if ε is fixed, we may say that if n is sufficiently large, then (2.42) holds.

For likelihood ratio-based DRO, as in the case of f-divergences, results typically operate in reverse: for a fixed sample size n, if the radius ρ is large enough, then we may achieve

(see Coppens and Patrinos [2023]) the following result:

$$\forall \boldsymbol{\theta} \in \Theta,\ \exists \rho : \max_{Q \in \mathcal{Q}_\rho(P_n)} \mathbb{E}_{\xi \sim Q}\left[\ell(\boldsymbol{\theta}, \xi)\right] \geq \mathbb{E}_{\xi \sim P}\left[\ell(\boldsymbol{\theta}, \xi)\right] \text{ w.h.p.} \tag{2.45}$$

Notice that the statement above is a random event with respect to the empirical measure P_n. Moreover, for continuous loss distributions, the selected ρ can be chosen independently of $\boldsymbol{\theta}$. We apply a similar result to select a minimum uncertainty set size in Section 2.7.

A second type of guarantee concerns *confidence*, or to say that with any prespecified probability δ, we may define a random upper bound U_n^δ based on observations $\xi_1, \ldots, \xi_n$ such that

$$\liminf_{n \to \infty} \mathbb{P}\left[U_n^\delta \geq \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\xi \sim P}\left[\ell(\boldsymbol{\theta}, \xi)\right]\right] \geq 1 - \delta. \tag{2.46}$$

This is simply a one-sided asymptotic $(1-\delta)$-confidence interval for the minimal value of $\boldsymbol{\theta} \mapsto \mathbb{E}_{\xi \sim P}\left[\ell(\boldsymbol{\theta}, \xi)\right]$. This setting can be interpreted as estimating a distributional parameter which is defined as the minimum of a particular functional of P defined for each $\boldsymbol{\theta}$, with associated uncertainty quantification. Notably, one possible choice of $U_n^\delta = \max_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_{\xi \sim Q}\left[\ell(\boldsymbol{\theta}_n, \xi) | P_n\right]$, because by (2.42), we have that

$$\max_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_{\xi \sim Q}\left[\ell(\boldsymbol{\theta}_n, \xi) | P_n\right] \geq \mathbb{E}_{\xi \sim P}\left[\ell(\boldsymbol{\theta}_n, \xi) | P_n\right] \geq \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\xi \sim P}\left[\ell(\boldsymbol{\theta}, \xi)\right] = \mathbb{E}_{\xi \sim P}\left[\ell(\boldsymbol{\theta}^\star_{\text{in}}, \xi)\right].$$

However, this may result in a loose interval, which can be seen explicitly in the upcoming expression for U_n^δ. As shown in Duchi et al. [2021], we have that

$$\begin{aligned} U_n^\delta &= \min_{\boldsymbol{\theta}, \Theta} \max_{Q \in \mathcal{Q}(P_n)} \mathbb{E}_{\xi \sim Q}\left[\ell(\boldsymbol{\theta}, \xi)\right], \\ \mathcal{Q}(P_n) &= \left\{ Q \ll P_n : D_f(Q \| P_n) \leq \frac{\chi^2_{1,1-2\delta}}{n} \right\} \end{aligned} \tag{2.47}$$

satisfies (2.46), where $\chi^2_{1,1-2\delta}$ denotes the $(1-2\delta)$-quantile of the χ^2-distribution with 1 degree of freedom. To comment on the looseness of the left-hand side of (2.42) for this purpose, we notice that the radius in (2.47) is shrinking at a rate of $O(n^{-1})$. Thus, as expected from a confidence interval, its size shrinks to zero as $n \to \infty$. In contrast, when

specifically pursuing distributional robustness, the radius is kept constant. Furthermore, the functional $\mathbb{E}_{\xi\sim P}[\ell(\boldsymbol{\theta},\xi)]$ can be replaced by a Hadamard differentiable functional of P. This method is known as *generalized empirical likelihood*, with respect to Owen's classical empirical likelihood approach for uncertainty quantification for Fréchet differentiable functionals Owen [1990]. Lam and Zhou [2017] also achieved (2.46) for particular f-divergences.

Direct Likelihood Ratio Estimation Our choice of objective is motivated by the fact that the user does not know what type of evaluation distribution Q they may observe during deployment. However, similar techniques can be applied in the setting of *domain adaptation*. This setting posits that $P \neq Q$ in some structured way, and this structure may be surmised from a potentially small number of examples from the shifted distribution. For instance, a training and test set of images may differ in distribution due to heterogeneous lighting conditions (a natural shift) or corruption of the test images through blurring (a synthetic shift). Here, we do in fact assume that $Q \ll P$, so there is a to-be-estimated population likelihood ratio $\beta : (\boldsymbol{x},\boldsymbol{y}) \mapsto \frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{x},\boldsymbol{y})$ for $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}\times\mathcal{Y}$. Given that such a population quantity exists, we may naturally attempt to estimate it, so that we may apply (2.2) to the sample.

One formal notion of structure is *covariate shift*, in which we assume that $P_{Y|X} = Q_{Y|X}$ so that $P \neq Q$ if and only if $P_X \neq Q_X$. As an abuse of notation, we consider the true β to only be a function of $\boldsymbol{x} \in \mathcal{X}$. We produce an estimate $\beta_n : \mathcal{X} \to \mathbb{R}$ which may be a function-valued estimator and not necessarily a list of n weights. Continuing within the covariate shift framework, because the shift only occurs in P_X, the statistical analysis will be *transductive*, in that we consider fixed (non-random) covariates $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^n$, random responses $Y_1,\ldots,Y_n \sim P_{Y|X}$, and fixed, unlabeled examples $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^m$ from the target domain. We discuss two seminal examples of methods in this setting. Both methods are kernel-based, in that the user first specifies a statistical model $\mathcal{P}$ such that $P_X, Q_X \in \mathcal{P}$ and a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ over $\mathbb{R}$ containing functions on $\mathcal{X}$. We denote by $\phi : \mathcal{X} \to \mathcal{H}$ the feature/lifting map and by $\mu_X : \mathcal{P} \to \mathcal{H}$ the mean map of $\mathcal{H}$, i.e.,

$\mu_X(P_X)(\boldsymbol{x}) = \mathbb{E}_{X\sim P_X}[\phi(X)(\boldsymbol{x})]$ for any $\boldsymbol{x} \in \mathcal{X}$. Consider the empirical measure of the examples from the target domain.

$$Q_{m,X} := \frac{1}{m}\sum_{i=1}^{m} \delta_{\boldsymbol{x}_i^{\mathrm{te}}}. \tag{2.48}$$

Note that while Q is absolutely continuous with respect to P, there is no guarantee that $Q_{m,X} \ll (P_n)_X$. In kernel mean matching [Gretton et al., 2008], β_n is estimated by solving

$$\min_{\beta\in\mathcal{B}(P_n)} \left\|\mu_X(Q_{m,X}) - \mathbb{E}_{(X,Y)\sim P_n}[\beta(X,Y)\phi(X)]\right\|_{\mathcal{H}}^2$$

which can be written as a finite-dimensional quadratic program. In practice, β is optimized with a slight relaxation of the constraint $\mathbb{E}_{(X,Y)\sim P_n}[\beta(X,Y)] = 1$, which is standard for a likelihood ratio. Note that this approach, while employing RKHS techniques, does not actually assume that $\beta \in \mathcal{H}$. On the other hand, kernelized unconstrained least-squares importance fitting [Kanamori et al., 2013] selects β_n as an element of the RKHS using a clever objective. Observe that for any β, the squared loss criterion against $\frac{\mathrm{d}Q}{\mathrm{d}P}$ can be written as

$$\left\|\beta - \tfrac{\mathrm{d}Q}{\mathrm{d}P}\right\|_{\mathbf{L}^2(P)}^2 = \int_{\mathcal{X}} \beta^2(\boldsymbol{x})\,\mathrm{d}P(\boldsymbol{x}) - 2\int_{\mathcal{X}} \beta(\boldsymbol{x})\,\mathrm{d}Q(\boldsymbol{x}) + \int_{\mathcal{X}} \tfrac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{x})\,\mathrm{d}Q(\boldsymbol{x})$$

with the last term independent of β. Thus, the optimization problem is written as

$$\min_{\beta\in\mathcal{H}} \mathbb{E}_{\xi\sim P_n}\left[\beta^2(Z)\right] - 2\mathbb{E}_{\xi\sim Q_{m,X}}\left[\beta(Z)\right],$$

Standard techniques to convert problems over RKHS to finite-dimensional programs via Tikhonov regularization can be applied to the problem above.

Importance-Weighted Algorithms for ERM Related to the likelihood ratio estimation literature, one may view a DRO routine as simultaneously estimating importance weights for training examples while actually solving the resulting importance-weighted empirical risk minimization (ERM) problem. The idea of combining ERM with a weighting function for generic losses is often credited to Shimodaira [2000], in the context of (weighted) maximum

likelihood estimation. In fact, this principle can be traced even further to pseudolikelihood estimation used in the analysis of survey data [Binder, 1983]. From then, importance-weighted ERM became a prominent approach in statistical learning for correcting sample bias and handling covariate shift [Huang et al., 2006, Sugiyama et al., 2007, 2008, Wen et al., 2014], with the likelihood ratio estimation problem studied in its own right [Kanamori et al., 2009]. Note that these applications of likelihood ratios for *bias reduction* methods differ from their use in stochastic simulation for *variance reduction* (e.g., in Monte Carlo estimation) [Kahn and Marshall, 1953].

To be precise, the use of importance weighting in a Monte Carlo application implies that there is a *known* target distribution Q, and we may select the distribution P from which we sample in order to generate the best estimate of a parameter of Q. In our setting, that of bias reduction, we may not change the data-generating distribution P, yet still estimate a parameter of Q. If either P or Q were unknown, this leads to an intermediate problem of estimating the likelihood function. Even if both were known, then we may devise unbiased estimators but still may wish to characterize the rates of decay in the variance [Ma et al., 2023]. Finally, in either bias or variance-reducing reweighting, if the estimand is some population risk function, there is the subsequent question of how empirical risk minimization performs on a rebalanced objective.

After the initial surge of importance weighted ERM for covariate shift in supervised learning, these methods appeared in the context of active learning [Mahmood et al., 2014, Swaminathan and Joachims, 2015, Wang et al., 2021], as a major source of bias comes from sampling from a particular policy (as opposed to "expert" demonstrations). More recently, the role of importance weighting in overparametrized statistical models (e.g., neural networks) has been studied aggressively in theory and practice [Byrd and Lipton, 2019, Xu et al., 2021]. As before, the task is typically thought of as a bilevel optimization problem consisting of 1) estimating the likelihood ratio, and 2) applying the learned likelihood ratio to importance-weighted ERM. The likelihood ratio may itself be parametrized by a neural network [Kato and Teshima, 2021] or related to an optimal transport map [Gong et al., 2016],

and even be learned in an online manner [Zhang et al., 2023]. A number of "end-to-end" approaches (i.e., learning the weighted estimator with a single optimization problem) have also been proposed [Fang et al., 2020, Zhang et al., 2020]. Finally, from the perspective of domain generalization more broadly, likelihood ratio estimation is seen as a form of *domain alignment* Zhang et al. [2013], such as moment/distribution matching between the source and target distributions.

Fairness and Subpopulation Shift DRO objectives, which are maxima over reweightings of the observed training data, are a special case of *subpopulation shift*, wherein the data-generating distribution is modeled as a mixture of subpopulations, and the distribution shift stems from changes in the mixture. In our case, the subpopulations are point masses at the observed data points. In the context of *algorithmic fairness*, the subpopulations may represent data conditioned on some protected attribute (e.g., race, gender, age range), and common notations of fairness such as *demographic/statistical parity* [Agarwal et al., 2018, 2019] impose (informally) that model performance with respect to each subpopulation should be roughly equal. As such, robustness to reweighting and algorithmic fairness are often aligned notions [Williamson and Menon, 2019], with recent research arguing that distributionally robust models are more fair [Hashimoto et al., 2018, Vu et al., 2022] and that fair models are more distributionally robust [Mukherjee et al., 2022].

2.9 Experiments

We compare Prospect against baselines in a variety of learning tasks. While we focus attention on its performance as an optimizer of its training objective, we also highlight metrics of interest on the test set in fairness and distribution shift benchmarks. To clarify terminology, we use "Prospect" and "LSVRG" to refer to the methods as described in Figure 2.3. In comparisons, we include stochastic algorithms that either are single-hyperparameter "out-of-the-box" methods, such as stochastic gradient descent and stochastic regularized dual averaging [Xiao, 2009], or multi-hyperparameter methods that converge linearly on strongly

convex SRM-based objectives such as stochastic saddle-point SAGA [Palaniappan and Bach, 2016]. The algorithm implementation and data preparation code are made publicly available online: https://github.com/ronakdm/prospect.

Setting, Baselines, Evaluation. We consider supervised learning tasks with input-label example $(\boldsymbol{x}_i, y_i)$. Losses are of the form $\ell_i(\boldsymbol{\theta}) := \text{err}(y_i, \langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}_i) \rangle)$, with a fixed feature embedding ϕ, and err measuring prediction loss. Uncertainty sets considered are the CVaR/superquantile, extremile, and ESRM. We compare against four baselines: minibatch stochastic gradient descent (SGD), stochastic regularized dual averaging (SRDA) [Xiao, 2009], Saddle-SAGA [Palaniappan and Bach, 2016], and LSVRG [Mehta et al., 2023]. For SGD and SRDA, we use a batch size of 64, and for LSVRG we use an epoch length of n. For Saddle-SAGA, we find that allowing different learning rates for the primal and dual variables improves experimental performance, so we compare against an improved heuristic (setting the dual stepsize as $10n$ times smaller than the primal stepsize). We plot

$$\text{Suboptimality}(\boldsymbol{\theta}) = \left(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^\star)\right) / \left(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^\star)\right), \tag{2.49}$$

where $\boldsymbol{\theta}^\star$ is approximated by running LBFGS [Nocedal and Wright, 1999] on the objective until convergence. The x-axis displays the number of calls to any first-order oracle $\boldsymbol{\theta} \mapsto (\ell_i(\boldsymbol{\theta}), \nabla \ell_i(\boldsymbol{\theta}))$ divided by n, i.e., the number of passes through the training set. We fix the shift cost $\nu = 1$ and regularization parameter $\mu = 1/n$. Experimental details and additional experiments with various hyperparameters are contained in Appendix A.4.

2.9.1 Tabular Least-Squares Regression

We consider five tabular regression benchmarks under square loss. The datasets used are `yacht` ($n = 244$) [Tsanas and Xifara, 2012], `energy` ($n = 614$) [Baressi Segota et al., 2020], `concrete` ($n = 824$) [Yeh, 2006], `kin8nm` ($n = 6553$) [Akujuobi and Zhang, 2017], and `power` ($n = 7654$) [Tüfekci, 2014]. The training curves are shown in Figure 2.4.

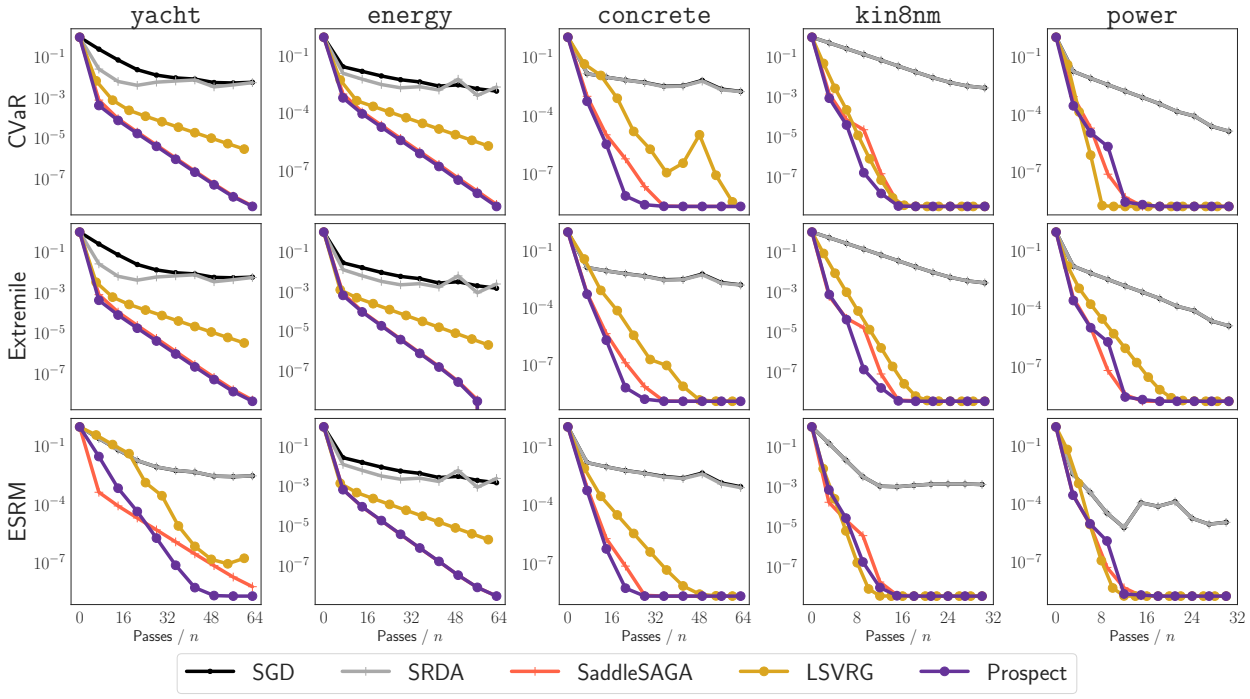

Figure 2.4: **Regression Benchmarks.** The y-axis measures suboptimality as given by (2.49), while the x-axis measures the number of calls to the function value/gradient oracle divided by n (i.e., passes through the training set). Rows indicate different SRM objectives while columns indicate datasets.

Results. Across datasets and objectives, we find that Prospect exhibits linear convergence at a rate no worse than SaddleSAGA and LSVRG, but that is often much better. For example, Prospect converges to precision 10^{-8} for the CVaR on `concrete` and the extremile on `power` within half the number of passes that LSVRG takes for the same suboptimality. Similarly, for the ESRM on `yacht`, SaddleSAGA requires 64 epochs to reach the same precision as Prospect at 40 epochs. The direct stochastic methods, SGD and SRDA, are biased and fail to converge for any learning rate.

2.9.2 Fair Classification and Regression

Inspired by Williamson and Menon [2019], we explore the relationship between distributional robustness and group fairness on 2 common tabular benchmarks. **Diabetes 130-Hospitals** (`diabetes`) is a classification task of predicting readmission for diabetes patients based on clinical data from US hospitals [Rizvi et al., 2014]. **Adult Census** (`acsincome`) is a regression task of predicting income of US adults from data compiled by the American Community Survey [Ding et al., 2021].

Evaluation. We evaluate fairness with the *statistical parity score*, which compares predictive distributions of a model given different values of a particular protected attribute Agarwal et al. [2018, 2019]. Letting $Z = (X, Y, A)$ denote a random (input, label, metadata attribute) triplet, a model g is said to satisfy statistical parity (SP) if the conditional distribution of $g(X)$ over predictions given $A = a$ is equal for any value a. Intuitively, SP scores measure the maximum deviation between these distributions for any over a, so values close to zero indicate SP-fairness. In `diabetes`, we use gender as the protected attribute A, whereas in `acsincome` we use race as the protected attribute. Note that the protected attributes are not supplied to the models. Results are given in Figure 2.5.

Results. Firstly, we note that Prospect converges rapidly on both datasets while LSVRG fails to converge on `diabetes` and SaddleSAGA fails to converge on `acsincome`. Secondly, LSVRG does not stabilize with respect to classification SP, showing a mean/std SP score of $1.38 \pm 0.25\%$ within the final ten passes on the `diabetes` CVaR, whereas Prospect gives $0.82 \pm 0.00\%$, i.e., a 40% relative improvement with greater stability. While SaddleSAGA does stabilize in SP on `diabetes`, it fails to qualitatively decrease at all on the `acsincome`. Interestingly, while suboptimality and SP-fairness are correlated for Prospect, SGD (reaching only 10^{-1} suboptimality with respect to the CVaR objectives on `acsincome`) achieves a lower fairness score. Again, across both suboptimality and fairness, Prospect is either the best or close to the best.

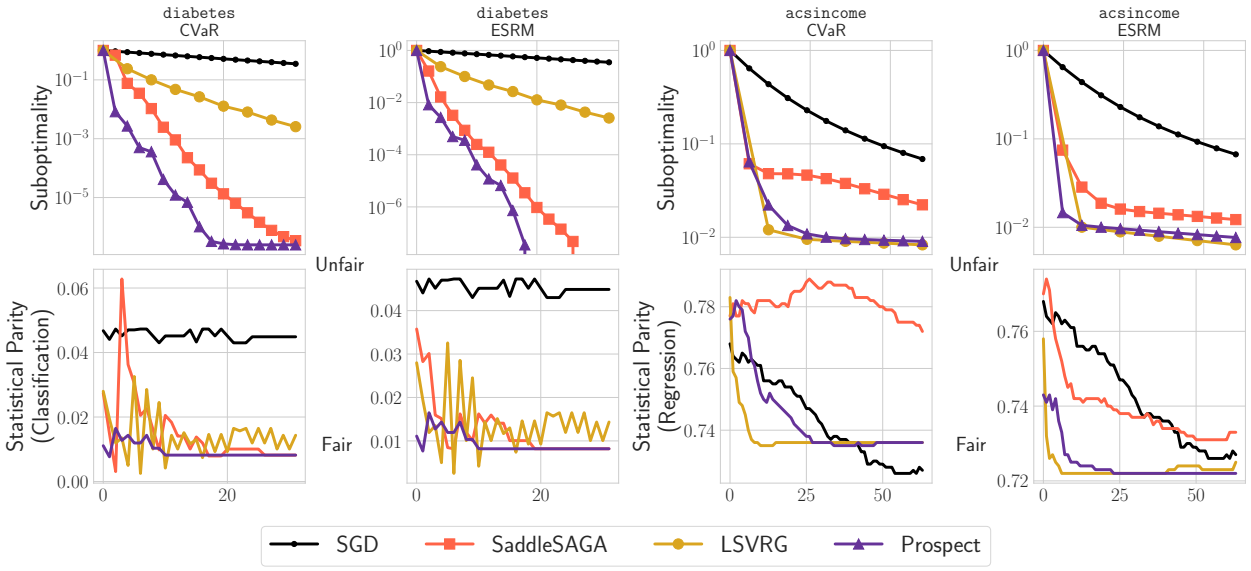

Figure 2.5: **Fairness Benchmarks. Top:** Training curves for optimizers on the CVaR and extremile for `diabetes` (**left**) and CVaR and extremile for `acsincome` (**right**). **Bottom:** Statistical parity scores for the two classification objectives on `diabetes` (**left**) and regression objectives on `acsincome`. Smaller values indicate better performance for all metrics.

2.9.3 Image and Text Classification under Distribution Shift

We consider two tasks from the WILDS distribution shift benchmark [Koh et al., 2021]. The **Amazon Reviews** (`amazon`) task [Ni et al., 2019] consists of classifying text reviews of products to a rating of 1-5, with disjoint train and test reviewers. The **iWildCam** (`iwildcam`) image classification challenge [Beery et al., 2020] contains labeled images of animals, flora, and backgrounds from cameras placed in wilderness sites. Shifts are due to changes in camera angles, locations, lighting... We use $n = 10000$ and $n = 20000$ examples respectively. For both datasets, we train a *linear probe classifier*, i.e., a linear model over a frozen deep representation. For `amazon`, we use a pretrained BERT model [Devlin et al., 2019a] fine-tuned on a held-out subset of the Amazon Reviews training set for 2 epochs. For `iwildcam`, we use a ResNet50 pretrained on ImageNet (without fine-tuning).

Evaluation. Apart from the training suboptimality, we evaluate the spectral risk objectives on their robustness to subpopulation shifts. We define each subpopulation group based on the true label. For `amazon`, we use the *worst group misclassification error* on the test set [Sagawa et al., 2020]. For `iwildcam`, we use the *median group error* owing to its larger number of classes.

Results. For both `amazon` and `iwildcam`, Prospect and SaddleSAGA (with our heuristic) outperform LSVRG in training suboptimality. We hypothesize that this phenomenon is due to checkpoints of LSVRG getting stale over the n-length epochs for these datasets with large n (leading to a slow reduction of bias). In contrast, Prospect and SaddleSAGA avoid this issue by dynamically updating the running estimates of the importance weights. For the worst group error for `amazon`, Prospect and SaddleSAGA outperform LSVRG. Prospect has a mean/std worst group error of 77.38±0.00% over the last ten passes on the extremile, whereas SaddleSAGA has a slightly worse 77.53±1.57%. Interestingly, on `iwildcam`, LSVRG and Prospect give stronger generalization performance, nearly 1pp better, than SaddleSAGA in terms of median group misclassification rate. In summary, across tasks and objectives, Prospect demonstrates best or close to best performance.

2.9.4 Scaling Laws and Shift Cost

We aim to disentangle the many effects of the shift cost parameter ν in the following experiments While ν has the statistical interpretation of penalizing values of $\boldsymbol{q}$ that stray far from the original uniform weights $\mathbf{1}/n$, we choose to control this distributional robustness property by instead using the uncertainty set $\mathcal{Q}$. This allows users of the non-smooth ($\nu = 0$) and smooth ($\nu > 0$) variants of the objective to use similar intuition for designing the uncertainty set. Accordingly, we view ν purely from an optimization lens, that is, we use it to allow for the algorithm to converge quickly. This involves trading off the per-iteration cost and the number of iterations.

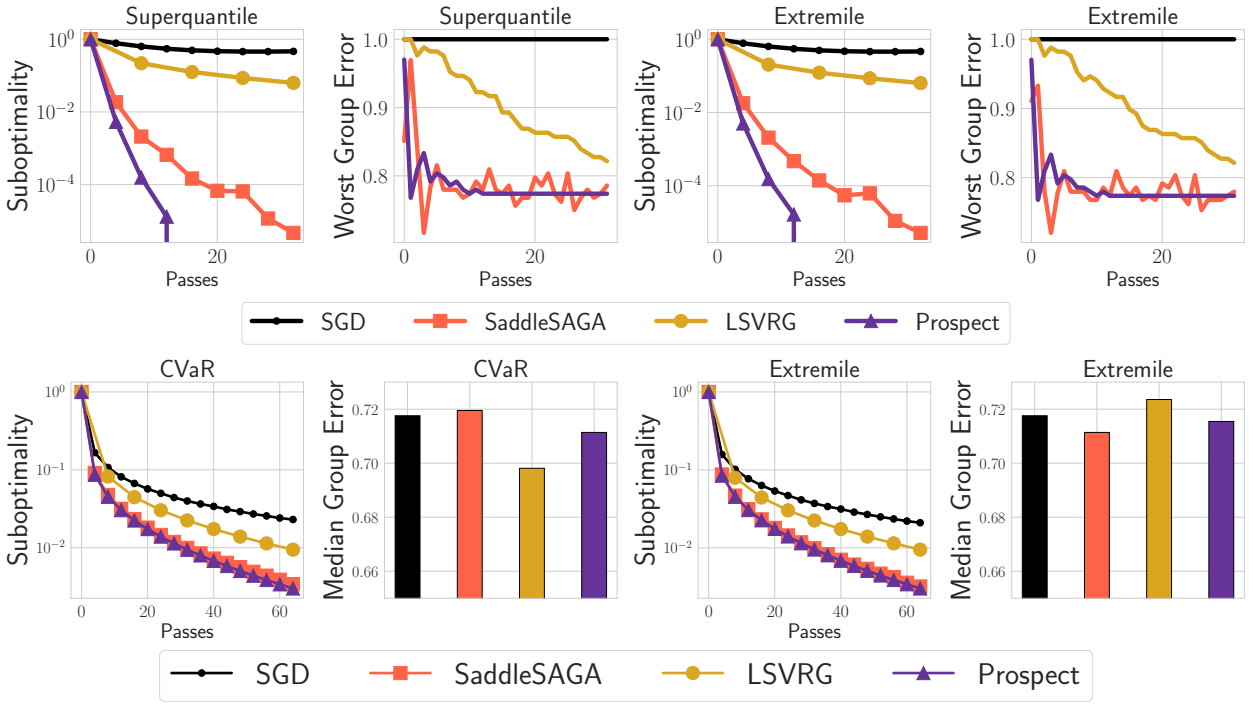

Figure 2.6: **Distribution Shift Benchmarks. Top:** Training curves and worst group misclassification error on `amazon` test. **Bottom:** Training curves and median group misclassification error on the `iwildcam` test set. Smaller values indicate better performance for all metrics.

Evaluation. Consider the following competing effects.

- **Number of Iterations:** Let $\varepsilon > 0$ by the desired suboptimality. Ignoring other parameters, the number of iterations has an $O(1/\nu)$ dependence on $\nu > 0$ (for any $\mu \geq 0$). More intuitively, ν regularizes the dual problem, leading to better conditioning of the objective. Thus, increasing ν will generally *decrease* the number of steps needed for convergence (but for a different objective).

- **Per-Iteration Cost:** The per-iteration cost has a subtle dependence on ν. Each iteration performs full maximization over q, which relies on subroutines such as sorting (e.g., the Pool Adjacent Violators algorithm in Appendix A.3). In particular, the map $\boldsymbol{l} \mapsto \max_{\boldsymbol{q} \in \mathcal{Q}} \langle \boldsymbol{q}, \boldsymbol{l} \rangle - \nu D_f(\boldsymbol{q} \| \mathbf{1}/n)$ can often be computed by fitting a monotonic function

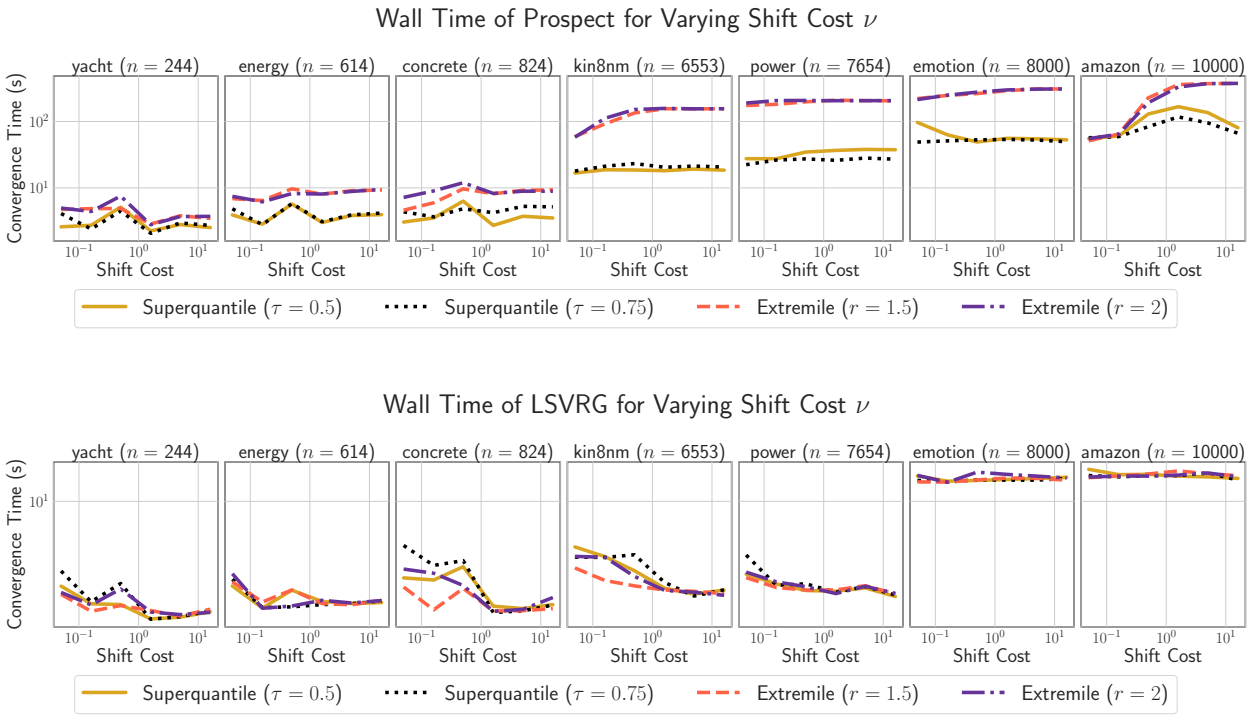

Figure 2.7: **Wall Time and Shift Cost Relationship. Top:** Wall time against shift cost for reaching convergence (squared gradient norm less than 10^{-3}) for Prospect. **Bottom:** Wall time against shift cost for reaching convergence (squared gradient norm less than 10^{-3}) for LSVRG.

to the sorted vector $(l_{(1)}, \ldots, l_{(n)})$ where $l_{(1)} \leq \ldots \leq l_{(n)}$. The parameter ν perturbs this sorted vector, perhaps breaking monotonicity and hence making it more "difficult" to fit a monotonic function. This difficulty renders as more inner loop iterations, so that the per-iteration cost may *increase* with increasing ν.

- **Altering the Objective/Solution:** Finally, changing ν also changes the objective, and in turn changes the optimal primal-dual pair $(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star)$. If the practitioner designs *Qcal* using a superquantile, for example, they imagine $\boldsymbol{q}^\star$ to be a "top-k weights" like vector. However, as we see below, increasing ν may alter the dual solution to the point of resembling the uniform weights with a few large "spikes". Thus, we would not like to bias our objective significantly.

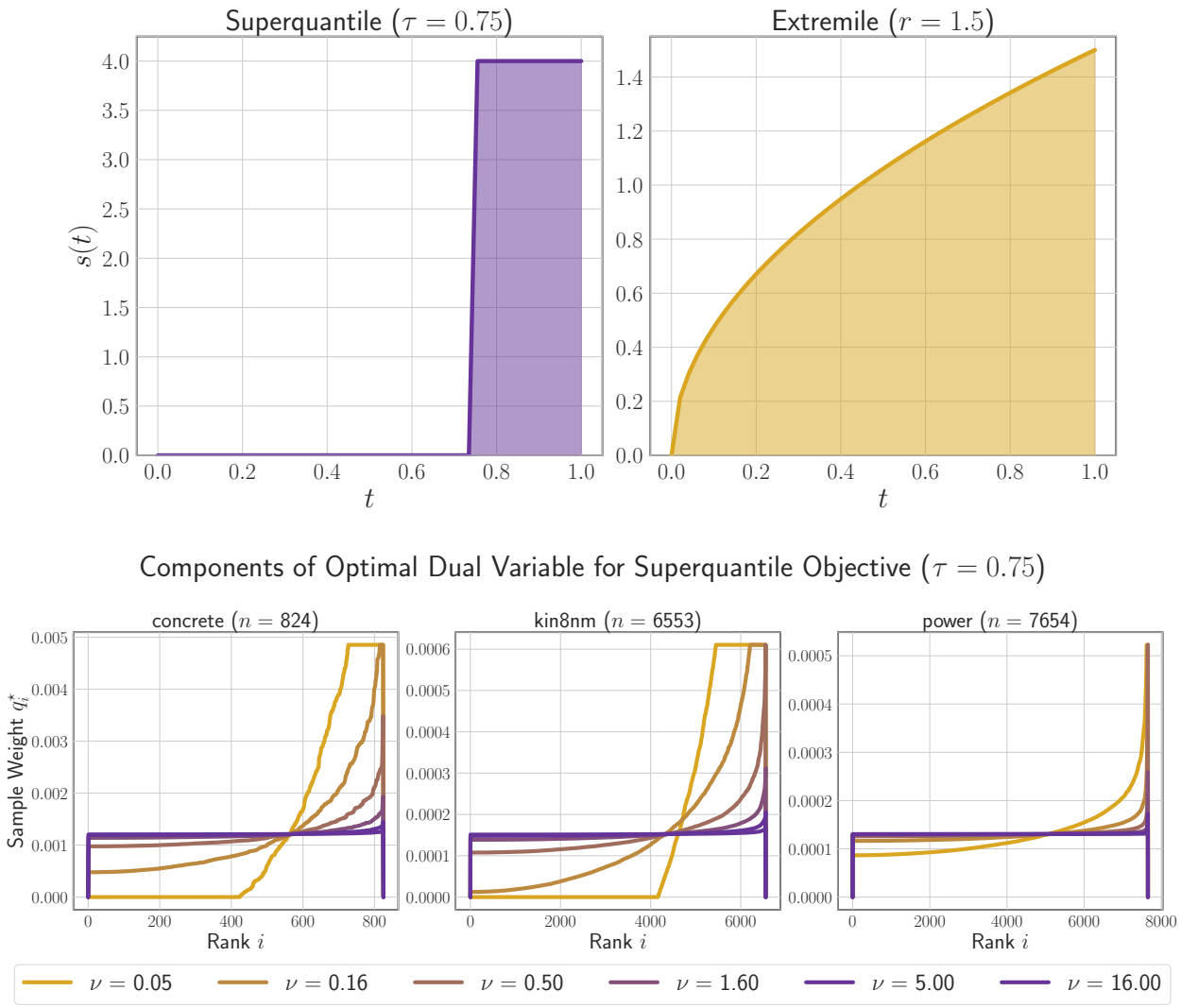

Figure 2.8: **Optimal Dual Solutions and Shift Cost Relationship. Top:** Visualization of continuous spectra for the superquantile and extremile. **Bottom:** Sorted optimal dual solution $\boldsymbol{q}^\star$ for different values of ν, meant to compare to the superquantile spectrum in the top left panel.

Results. Ultimately, we find that setting ν to be as small as possible without harming convergence (e.g., a dual value of $\nu = 0.05$) generally works well, as 1) convergence times are generally *faster* for small shift costs (Figure 2.7), and 2) the dual solution remains close to that of $\nu = 0$ and depends only on the uncertainty set (Figure 2.8).

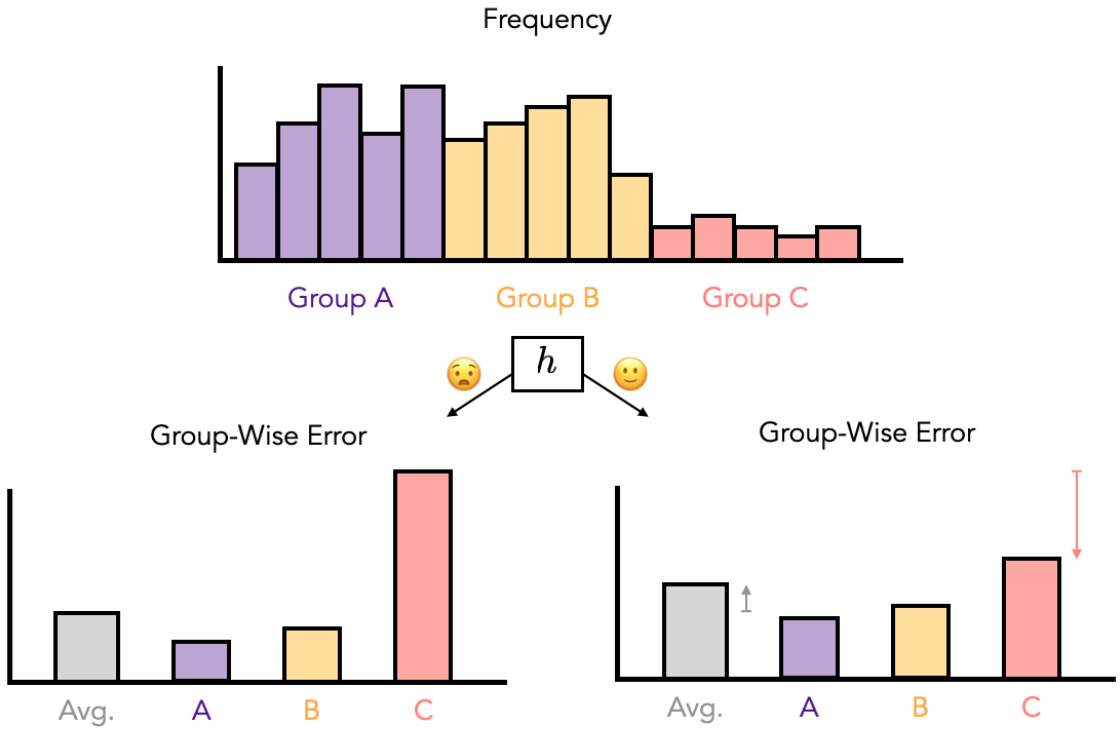

Figure 2.9: **Illustration of Group-Wise Error Evaluation.** The top panel depicts a hypothetical probability distribution, where the sample space is partitioned into three groups. When applying a statistical prediction model h to this distribution, the bottom left and right panels show non-uniform and uniform performance conditional on each group. On the right panel specifically, the arrows indicate a slight increase in the average loss across all examples, but a more dramatic decrease in the worst-case group-wise error.

We consider extensions to both the theoretical analysis and algorithmic aspects in the next section.

2.10 Incorporating Group Structure

While the stochastic algorithms introduced in this chapter make the DRO problem tractable for moderately large sample sizes, the $O(n)$ cost of each update of the dual variables becomes a major limitation when attempting to adapt methods to neural network applications such as computer vision and natural language processing. The seminal work of Sagawa et al. [2020] applied DRO at the level of groups or subpopulations within the data. That is, while different

groups may be upweighted or downweighted in the objective, the overall performance of a model on a particular group was measured by the simple average of within-group errors. In this section, we refer to the setup of Section 2.1 as *instance-level DRO*, to be contrasted with *group DRO* as described below. At a high level, the practitioner wishes to train a model that controls *worst case group-wise error* across groups by changing the training objective in a manner similar to that of Section 2.4. Unfavorable and favorable group-wise error distributions are depicted in Figure 2.9.

To describe the setting formally, we rely on the likelihood ratio-based framework introduced in Chapter 1 and Section 2.8. We assume that the data-generating distribution P and the shifted distribution Q are over the augmented domain $\Xi \times \mathcal{A}$, where Ξ denotes observable data (possibly feature-label pairs) and $\mathcal{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M\}$ denotes an additional component of the datum, often called the *group label* or *protected attribute* in various contexts. Denote by P_ξ and P_A the marginal distributions of P on Ξ and $\mathcal{A}$, respectively. Because A is discrete, we will use the notation $P_A(\boldsymbol{a}) \equiv P_A(\{\boldsymbol{a}\})$ to indicate the probability mass function in this section. To proceed, we collect the assumptions required for a well-defined empirical risk objective below.

Assumption 2.10.1. The following statements hold.

- We have that $P_A(\boldsymbol{a}) > 0$ for all $a \in \mathcal{A}$.
- An i.i.d. sample $(\xi_1, A_1), \ldots, (\xi_n, A_n) \sim P$ is observed at train time. In particular, training examples are accompanied by group labels (but test examples may not be).
- Consider the empirical measure P_n with marginals $P_{n,\xi}$ and $P_{n,A}$ on Ξ and $\mathcal{A}$, respectively. Then, $P_{n,A}(\boldsymbol{a}) > 0$ for all $\boldsymbol{a} \in \mathcal{A}$ (i.e., all groups are observed in the data).

The first part of the Assumption 2.10.1 is purely technical, so that group conditional means are well-defined. The second is methodological, implying that the user may observe the group information when training models. The third part is in fact a random event for

which it makes sense to consider an objective that depends on group-wise means. In a statistical analysis, we would return another estimator in the case that $P_{n,A}(\boldsymbol{a}) = 0$ for some $a \in \mathcal{A}$, and upper bound the probability of this occurring in terms of n. See Proposition B.3.3, used to prove the main results of Section 4.3, for a direct example of such an argument.

We first introduce the population objective. While the format will resemble the likelihood ratio-based reweighted objective (2.2) from Section 2.2, it involves structural assumptions similar to the covariate shift example from Section 2.8 and the distributional assumption from (2.3). We first construct

$$\mathcal{B}^{\text{struct}}(P) := \left\{\beta \in \mathcal{B}_0(P) : \beta(\boldsymbol{z}, \boldsymbol{a}) = \beta(\boldsymbol{z}', \boldsymbol{a}) \ \forall (\boldsymbol{z}, \boldsymbol{z}', \boldsymbol{a}) \in \Xi \times \Xi \times \mathcal{A}\right\},$$

indicating that the likelihood ratio can only be a function of the attribute input $\boldsymbol{a} \in \mathcal{A}$. This is analogous to covariate shift, in which case the likelihood ratio was only a function of the feature component of ξ. Then, using f-DRO as an example, we also introduce

$$\mathcal{B}^{\text{DRO}}(P) := \{\beta \in \mathcal{B}_0(P) : \mathbb{E}_P\left[f\left(\beta(\xi, A)\right)\right] \leq \rho\},$$

and finally set $\mathcal{B}(P) := \mathcal{B}^{\text{struct}}(P) \cap \mathcal{B}^{\text{DRO}}(P)$. In words, we consider distribution shifts that only occur on the weights of P_A (via $\mathcal{B}^{\text{struct}}(P)$), subject to a constraint on uncertainty (via $\mathcal{B}^{\text{DRO}}(P)$). The objective then reads as

$$\min_{\theta \in \mathbb{R}^d} \max_{\beta \in \mathcal{B}(P)} \left\{\mathbb{E}_{(\xi,A)\sim P}\left[\beta(\xi, A)\ell(\boldsymbol{\theta}, \xi)\right] := \mathbb{E}_{A \sim P_A}\left[\beta(\cdot, A) \cdot \mathbb{E}_P\left[\ell(\boldsymbol{\theta}, \xi)\right] | A\right]\right\}. \tag{2.50}$$

Notice that another implicit choice is made here, which is that while we observe realizations of A, they are not included as input to the prediction function or the instance-level loss function $\ell : \mathbb{R}^d \times \Xi \to \mathbb{R}$. Next, we describe the empirical counterpart of this objective along with the conceptual and practical differences between group DRO and the instance-level setting.

2.10.1 Learning Objective and Algorithm Description

Learning Objective By replacing P with P_n in (2.50), develop an objective similar to (2.18) at the group level. Given observed data $\{(\xi_i, A_i)\}_{i=1}^n$, we maintain the notation $\ell_i(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}, \xi_i)$ for parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and define groups of indices $\mathcal{I}_1, \ldots, \mathcal{I}_M$ such that $A_i = \boldsymbol{a}_m \iff i \in \mathcal{I}_m$. Letting $n_m := |\mathcal{I}_m| \geq 2$ be the group-level sample size, we denote the empirical group means as

$$\bar{\ell}_m(\boldsymbol{\theta}) := \frac{1}{n_m} \sum_{i \in \mathcal{I}_m} \ell_i(\boldsymbol{\theta}).$$

We then naturally define the objective

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left[\mathcal{L}^{\text{gr}}(\boldsymbol{\theta}) := \max_{\boldsymbol{q} \in \mathcal{Q}} \sum_{m=1}^M q_m \cdot \bar{\ell}_m(\boldsymbol{\theta}) - \nu \operatorname{Reg}(\boldsymbol{q}) + \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2 \right], \tag{2.51}$$

where $\mathcal{Q} \subseteq \Delta^{M-1}$ is an uncertainty set and $\operatorname{Reg} : \Delta^{M-1} \to \mathbb{R}$ is a regularization function that is strongly convex with respect to a particular norm. While (2.51) might seem like a direct analog of (2.18), there are a number of differences to be considered in the group DRO setting.

- There is no canonical choice of Reg based on a statistical divergence, as there are multiple choices justified for different reasons. At the instance level, the empirical measure of the data places uniform weight on all examples. While we may use $\operatorname{Reg}(\boldsymbol{q}) := D_f(\boldsymbol{q} \| \mathbf{1}_M / M)$ as a regularizer, the empirically observed group weights $\boldsymbol{p}^{\text{gr}} := (n_1/n, \ldots, n_M/n)$ could be completely different. We may consider using $\operatorname{Reg}(\boldsymbol{q}) := D_f(\boldsymbol{q} \| \boldsymbol{p}^{\text{gr}})$ to promote similarity to the observed distribution.

- Unlike in (2.18), where we may compute each $(\ell_i, \nabla \ell_i)$ at cost $\tilde{O}(d)$, we require $\tilde{O}(nd/M)$ operations to query $(\bar{\ell}_i, \nabla \bar{\ell}_i)$ on average. In a practical setting, we may only access unbiased estimates of the $(\bar{\ell}_i, \nabla \bar{\ell}_i)$ oracles, which introduces another source of statistical error to be considered in the analysis.

Algorithm 2 Group-DRO with Prospect Bias Reduction

Inputs: Initial points $\boldsymbol{\theta}^{(0)}$, stepsize $\eta > 0$, per-group batch sizes $b_1 \in [n_1], \ldots, b_M \in [n_M]$, primal regularization (weight decay) parameter μ, dual regularization (shift cost) parameter ν, and number of iterations t.

Set initialization $\boldsymbol{\theta}^{(0)}$.
Compute $\boldsymbol{l}^{(0)} = (l_1^{(0)}, \ldots, l_n^{(0)}) = \ell(\boldsymbol{\theta}^{(0)})$.
Compute $\bar{\boldsymbol{l}}^{(0)} = (\bar{l}_1^{(0)}, \ldots, \bar{l}_M^{(0)})$ where $\bar{l}_m^{(0)} = \frac{1}{n_m} \sum_{i \in \mathcal{I}_m} l_i^{(0)}$ for $m = 1, \ldots, M$.
for $k = 0, \ldots, t-1$ **do**
    Sample b_m points from each group, yielding indices $\{(i_{m,1}, \ldots, i_{m,b_m})\}_{m=1}^M$.

    Compute the group-wise average losses via

$$\ell^{(k+1)}(\boldsymbol{\theta}^{(k)}) = \left(\frac{1}{b_1} \sum_{j=1}^{b_1} \ell_{i_{1,j}}(\boldsymbol{\theta}^{(k)}), \ldots, \frac{1}{b_M} \sum_{j=1}^{b_M} \ell_{i_{M,j}}(\boldsymbol{\theta}^{(k)}) \right) \in \mathbb{R}^M.$$

    Compute the group weights $\boldsymbol{q}^{(k)} = q^{\text{opt}}\left(\bar{\boldsymbol{l}}^{(k)}\right)$, which incorporates the shift cost ν.
    Compute the gradient estimate $\boldsymbol{g}^{(k+1)} = \sum_{m=1}^M q_m^{(k)} \nabla \ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)})$. In an auto-differentiation framework, this can be accomplished by backpropagating through the computation graph of $\theta \mapsto \langle \text{StopGrad}(\boldsymbol{q}^{(k)}), \ell^{(k+1)}(\boldsymbol{\theta}) \rangle$.
    $\boldsymbol{\theta}^{(k+1)} = \text{Step}(\boldsymbol{\theta}^{(k)}, \boldsymbol{g}^{(k+1)}, \eta, \mu)$.

    Update $l_{i_{m,j}}^{(k+1)} = \ell_{i_{m,j}}(\boldsymbol{\theta}^{(k+1)})$ for $m \in [M]$ and $j \in [b_m]$.
    Set $l_i^{(k+1)} = l_i^{(k)}$ for any $i \in [n]$ such that $i \neq i_{m,j}$ for all (m, j).
Update $\bar{l}_m^{(k+1)} = \frac{1}{n_m} \sum_{i \in \mathcal{I}_m} l_i^{(k+1)}$ for $m = 1, \ldots, M$.

Output: Final point $\boldsymbol{\theta}^{(t)}$

Algorithm Description On the algorithmic side, we apply an optimization scheme in a similar spirit to Algorithm 1, described in Algorithm 2 below. In this case, we eschew the variance reduction component because such algorithms are generally used for large-scale applications with neural network models, so storing historical parameter vectors may be impractical (even at the group level). Instead, we will capture the variance of the group mean estimates described above in an SGD-style analysis. Algorithm 2 operates by maintaining an $O(n)$-sized table $\boldsymbol{l}^{(k)}$ and changing $\sum_m b_m$ coordinates at each iteraton. This table aggregates past information and is used to estimate the dual variables (the sample weights), which would otherwise cost $O(n)$ oracles to compute. However, even in the case of bias reduction, we find

Algorithm 3 Group-DRO with Stratified Sampling

Inputs: Initial points $\boldsymbol{\theta}^{(0)}$, stepsize $\eta > 0$, per-group batch sizes $b_1 \in [n_1], \ldots, b_M \in [n_M]$, primal regularization (weight decay) parameter μ, dual regularization (shift cost) parameter ν, and number of iterations t.

1: Set initialization $\boldsymbol{\theta}^{(0)}$.
2: **for** $k = 0, \ldots, t-1$ **do**
3: Sample b_m points from each group, yielding indices $\{(i_{m,1}, \ldots, i_{m,b_m})\}_{m=1}^{M}$.
4: Compute $\ell^{(k+1)}(\boldsymbol{\theta}^{(k)}) = \left(\frac{1}{b_1}\sum_{j=1}^{b_1} \ell_{i_{1,j}}(\boldsymbol{\theta}^{(k)}), \ldots, \frac{1}{b_M}\sum_{j=1}^{b_M} \ell_{i_{M,j}}(\boldsymbol{\theta}^{(k)})\right) \in \mathbb{R}^M$.
5: Compute $\boldsymbol{q}^{(k)} = q^{\text{opt}}(\ell^{(k+1)}(\boldsymbol{\theta}^{(k)}))$ and the gradient estimate $\boldsymbol{g}^{(k+1)} = \sum_{m=1}^{M} q_m^{(k)} \nabla \ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)})$ (see Algorithm 2).
6: $\boldsymbol{\theta}^{(k+1)} = \text{Step}(\boldsymbol{\theta}^{(k)}, \boldsymbol{g}^{(k+1)}, \eta, \mu)$.

Output: Final point $\boldsymbol{\theta}^{(t)}$

in experiments that this approach generally does not change optimization or classification performance, as the average within-group loss can be estimated with low bias using moderate batch sizes (between the orders of 10^1 and 10^2). Thus, we also describe a more standard stochastic gradient algorithm Algorithm 3 that can be easily plugged into any deep learning workflow with minimal additional code. Algorithm 3 still differs from previous approaches in the literature, as we may apply the techniques for handling spectral risk measure uncertainty sets and smoothing developed throughout this chapter. Finally, we allow for the choice of a general update function in the description of the algorithms, namely in line 10 of Algorithm 2 and line 6 of Algorithm 3. While we typically consider the standard stochastic gradient update

$$\boldsymbol{\theta}^{(k+1)} = \text{Step}(\boldsymbol{\theta}^{(k)}, \boldsymbol{g}^{(k+1)}, \eta, \mu) = \boldsymbol{\theta}^{(k)} - \eta \boldsymbol{v}^{(k+1)}, \text{ for } \boldsymbol{v}^{(k+1)} = \boldsymbol{g}^{(k+1)} + \mu \boldsymbol{\theta}^{(k)}, \tag{2.52}$$

the algorithm works well empirically by using update steps that are attuned to deep learning models, such as the Adam optimizer [Kingma and Ba, 2015].

Sampling Details In both Algorithm 2 and Algorithm 3, the data is sampled in a *stratified* manner; that is, we select batch sizes for each group and sample data points from each group on every iteration [Singh and Mangat, 1996, pages 102–144]. This ensures that the

user may compute estimates of the average loss on each group and control their statistical precision. While this can be implemented in practice by using multiple data loaders, one trick to approximate this with minimal changes to a typical PyTorch workflow is to sample the data using *weighted sampling with replacement*. In this scheme, one specifies sample weights $(p_1, \ldots, p_n)$ (or an unnormalized version thereof) associated with each data point, and samples from this distribution with replacement when producing a mini-batch. If the weights are selected so that for all $m \in [M]$, the conditions

$$p_i = p_j \text{ for all } i, j \in \mathcal{I}_m \text{ and } \sum_{i \in \mathcal{I}_m} p_i = \frac{1}{M},$$

then we may expect approximately b observations from each group in any mini-batch of size bM. In PyTorch, we simply supply the sample weight vector when initializing the data loader. In the code snippet below, we use `group_labels` to refer to the vector $(A_1, \ldots, A_n)$, where the $\mathcal{A} = \{0, \ldots, M-1\}$, and `batch_size` refers to bM. The reason for setting `drop_last=True` is so that if bM does not divide n, we do not get a smaller mini-batch that does not have enough samples to compute each group mean.

```
import numpy as np
from torch.utils.data import DataLoader, WeightedRandomSample
...
group_label_count = np.bincount(group_labels)
sample_weight = (1 / group_label_count)[group_labels] / len(group_label_count)


train_dataloader = DataLoader(
    train_dataset,
    sampler=WeightedRandomSampler(
        sample_weight,
        len(sample_weight),
        replacement=True
    ),
    batch_size=batch_size,
    drop_last=True
)
```

2.10.2 Convergence Analysis

The convergence analysis will employ similar tools to Section 2.6, but incorporate the noise resulting from imperfect estimates of the average loss of the parameter on each group. Using a smoothed group DRO objective will be instrumental in providing bias and variance control in terms of the per-group batch sizes. We introduce some notation that is used in the proof, which is analogous to the one used in the convergence analysis of Section 2.6. Let $\bar{r}_m(\cdot) := \bar{\ell}_m(\cdot) + \frac{\mu}{2}\|\cdot\|_2^2$ denote the regularized group-wise loss, defining the function $\bar{r} : \mathbb{R}^d \to \mathbb{R}^n$ via $\bar{r}(\boldsymbol{\theta}) := (\bar{r}_1(\boldsymbol{\theta}), \ldots, \bar{r}_M(\boldsymbol{\theta}))$. We also define the iterates $\hat{\boldsymbol{\theta}}_i^{(k)} \in \mathbb{R}^d$ as those satisfying $l_i^{(k)} = \ell_i(\hat{\boldsymbol{\theta}}_i^{(k)})$ in Algorithm 2. Due to strong convexity and strong concavity, we

have that (2.51) has a unique saddle point $(\boldsymbol{\theta}^\star, \boldsymbol{q}^\star) \in \mathbb{R}^d \times \mathcal{Q}$ (recalling that $\mathcal{Q} \subseteq \mathbb{R}^M$). We maintain the following assumption throughout.

Assumption 2.10.2. The unregularized losses ℓ_i are convex, G_m-Lipschitz, and $(\bar{M} - \mu)$-smooth for $\bar{M} \geq \mu$, $i \in \mathcal{I}_m$, and $m \in [M]$. The shift penalty $\mathrm{Reg} : \mathcal{Q} \to \mathbb{R}$ is 1-strongly convex with respect to $\|\cdot\|_2$.

We emphasize the main difference between this setting and the instance-level incremental setting is that here, we observe a noisy version *every* element of the vector $\bar{\ell}(\boldsymbol{\theta}^{(k)})$ on each iteration, whereas before, we observed an exact version of *one* element of the vector $\ell(\boldsymbol{\theta}^{(k)})$. To handle this, we will require an additional variance assumption. Recall the notation

$$\nabla \ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)}) = \frac{1}{b_m} \sum_{j=1}^{b_m} \nabla \ell_{i_{m,j}}(\boldsymbol{\theta}^{(k)})$$

from Algorithm 2. Conditioned on $\boldsymbol{\theta}^{(k)}$, the randomness in $\nabla \ell_m^{(k+1)}$ is governed by sampling the indices $i_{m,1}, \ldots, i_{m,b_m}$ without replacement from $[n_m]$, motivating the assumption below. Let $\mathbb{E}_k$ denote the conditional expectation given $\boldsymbol{\theta}^{(k)}$.

Assumption 2.10.3. There exists constants $\bar{\sigma}_1^2, \ldots, \bar{\sigma}_M^2 \geq 0$ such that for all $k \geq 0$ and $m = 1, \ldots, M$, it holds that

$$\mathbb{E}_k \left\| \nabla \ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)}) - \nabla \bar{\ell}_m(\boldsymbol{\theta}^{(k)}) \right\|_2^2 \leq \frac{(n_m - b_m)\bar{\sigma}_m^2}{(n_m - 1)b_m}$$

Crucially, this assumption is made on the gradient of the *unregularized* losses. Under Assumption 2.10.2, these are bounded random variables which allow us to satisfy Assumption 2.10.3 uniformly over $\boldsymbol{\theta} \in \mathbb{R}^d$. To see this, fix $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^d$ and compute directly (using the

finite population correction) the expression

$$\begin{aligned}\mathbb{E}_k \left\|\nabla \ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)}) - \nabla \bar{\ell}_m(\boldsymbol{\theta}^{(k)})\right\|_2^2 &= \frac{(n_m - b_m)}{(n_m - 1)b_m}\left(\frac{1}{n_m}\sum_{i \in \mathcal{I}_m} \left\|\nabla \ell_i(\boldsymbol{\theta}^{(k)}) - \nabla \bar{\ell}_m(\boldsymbol{\theta}^{(k)})\right\|_2^2\right) \\ &\leq \frac{(n_m - b_m)}{(n_m - 1)b_m}\left(\frac{1}{n_m}\sum_{i \in \mathcal{I}_m} \left\|\nabla \ell_i(\boldsymbol{\theta}^{(k)})\right\|_2^2\right) \\ &\leq \frac{(n_m - b_m)}{(n_m - 1)b_m} G_m^2,\end{aligned}$$

where the second line follows because the second moment upper bounds the variance, and the third line follows by Assumption 2.10.2. Thus, $\bar{\sigma}_m^2$ is always upper bounded by G_m^2, but could be much smaller if the gradients within a group are large but concentrated about their mean. The last assumption is on the update function.

Assumption 2.10.4. Consider $\boldsymbol{v}^{(k+1)}$ as defined in (2.52). The Step function from line 10 of Algorithm 2 and line 6 of Algorithm 3 satisfies the decomposition

$$\mathbb{E}_k \left\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\right\|_2^2 \leq \left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2 - 2\eta C_1 \left\langle \bar{r}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}, \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star \right\rangle + \eta^2 C_2 \mathbb{E}_k \left\|\boldsymbol{v}^{(k+1)}\right\|_2^2.$$

In the case of the standard stochastic gradient update (2.52), Assumption 2.10.4 is satisfied as an equality with $C_1 = C_2 = 1$.

As before, we organize the analysis into a bound on the descent term (or bias analysis) and on the noise term (or variance analysis), and derive conditions on the learning rate $\eta > 0$. The following result is similar to Proposition 2.6.1.

Lemma 2.10.1 (Bias Analysis)**.** *For any $\boldsymbol{q}, \boldsymbol{q}^\star \in \Delta^{M-1}$ and $\boldsymbol{\theta}, \boldsymbol{\theta}^\star \in \mathbb{R}^d$, it holds that*

$$\begin{aligned}-(\nabla \bar{r}(\boldsymbol{\theta})^\top \boldsymbol{q} - \nabla \bar{r}(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^\star)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \leq -(\boldsymbol{q} - \boldsymbol{q}^\star)^\top (\bar{\ell}(\boldsymbol{\theta}) - \bar{\ell}(\boldsymbol{\theta}^\star)) - \frac{\mu}{2}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right\|_2^2 \\ - \frac{1}{2(\bar{M} + \mu)}\left\|(\nabla \bar{r}(\boldsymbol{\theta}) - \nabla \bar{r}(\boldsymbol{\theta}^\star))^\top \boldsymbol{q}\right\|_2^2.\end{aligned}$$

Proof. The functions $\boldsymbol{\theta} \mapsto \bar{r}(\boldsymbol{\theta})^\top \boldsymbol{q}$ and $\boldsymbol{\theta} \mapsto \bar{r}(\boldsymbol{\theta})^\top \boldsymbol{q}^\star$ are each $\bar{M}$-smooth and μ-strongly convex, as $\boldsymbol{q}$ and $\boldsymbol{q}^\star$ are both contained in the probability simplex. By applying Lemma A.1.1,

we achieve the inequalities

$$\begin{aligned}\boldsymbol{q}^\top \bar{r}(\boldsymbol{\theta}^\star) \geq \boldsymbol{q}^\top \bar{r}(\boldsymbol{\theta}) + \boldsymbol{q}^\top \nabla \bar{r}(\boldsymbol{\theta})(\boldsymbol{\theta}^\star - \boldsymbol{\theta}) \\ + \frac{1}{2(\bar{M}+\mu)} \left\| (\nabla \bar{r}(\boldsymbol{\theta}) - \nabla \bar{r}(\boldsymbol{\theta}^\star))^\top \boldsymbol{q} \right\|_2^2 + \frac{\mu}{4} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^\star \right\|_2^2\end{aligned}$$

and

$$\begin{aligned}(\boldsymbol{q}^\star)^\top \bar{r}(\boldsymbol{\theta}) \geq (\boldsymbol{q}^\star)^\top \bar{r}(\boldsymbol{\theta}^\star) + (\boldsymbol{q}^\star)^\top \nabla \bar{r}(\boldsymbol{\theta}^\star)(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\ + \frac{1}{2(\bar{M}+\mu)} \left\| (\nabla \bar{r}(\boldsymbol{\theta}^\star) - \nabla \bar{r}(\boldsymbol{\theta}))^\top \boldsymbol{q}^\star \right\|_2^2 + \frac{\mu}{4} \left\| \boldsymbol{\theta}^\star - \boldsymbol{\theta} \right\|_2^2 .\end{aligned}$$

Summing the two inequalities gives

$$\begin{aligned}-(\nabla \bar{r}(\boldsymbol{\theta})^\top \boldsymbol{q} - \nabla \bar{r}(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^\star)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \leq -(\boldsymbol{q} - \boldsymbol{q}^\star)^\top (\bar{r}(\boldsymbol{\theta}) - \bar{r}(\boldsymbol{\theta}^\star)) - \frac{\mu}{2} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^\star \right\|_2^2 \\ - \frac{1}{2(\bar{M}+\mu)} \left\| (\nabla \bar{r}(\boldsymbol{\theta}) - \nabla \bar{r}(\boldsymbol{\theta}^\star))^\top \boldsymbol{q} \right\|_2^2 \\ - \frac{1}{2(\bar{M}+\mu)} \left\| (\nabla \bar{r}(\boldsymbol{\theta}) - \nabla \bar{r}(\boldsymbol{\theta}^\star))^\top \boldsymbol{q}^\star \right\|_2^2 .\end{aligned}$$

Drop the final non-positive term and mirror the argument of Proposition 2.6.1 to achieve

$$(\boldsymbol{q} - \boldsymbol{q}^\star)^\top (\bar{r}(\boldsymbol{\theta}) - \bar{r}(\boldsymbol{\theta}^\star)) = (\boldsymbol{q} - \boldsymbol{q}^\star)^\top (\bar{\ell}(\boldsymbol{\theta}) - \bar{\ell}(\boldsymbol{\theta}^\star)).$$

and complete the proof. □

In the variance bound, we will also use the spectral norm, or largest singular value $s_{\max}(\nabla \bar{r}(\boldsymbol{\theta}^\star))$ of the Jacobian of the regularized loss at the optimum. We only use this constant at the optimum, as the quantity can be unbounded over all of $\boldsymbol{\theta} \in \mathbb{R}^d$ due to regularization.

Lemma 2.10.2 (Variance Analysis)**.** *It holds that*

$$\begin{aligned}\mathbb{E}_k \left\| \boldsymbol{v}^{(k+1)} \right\|_2^2 \leq \max_{\boldsymbol{q} \in \mathcal{Q}} \left\| \boldsymbol{q} \right\|_2^2 \sum_{m=1}^{M} \frac{(n_m - b_m)\bar{\sigma}_m^2}{(n_m - 1)b_m} + 2 \left\| (\nabla \bar{r}(\boldsymbol{\theta}^{(k)}) - \nabla \bar{r}(\boldsymbol{\theta}^\star))^\top \boldsymbol{q}^{(k)} \right\|_2^2 \\ + 2 s_{\max}(\nabla \bar{r}(\boldsymbol{\theta}^\star))^2 \left\| \boldsymbol{q}^{(k)} - \boldsymbol{q}^\star \right\|_2^2 .\end{aligned}$$

Proof. Using the variance identity $\mathbb{E}\left\|X\right\|_2^2 = \mathbb{E}\left\|X - \mathbb{E}[X]\right\|_2^2 + \left\|\mathbb{E}[X]\right\|_2^2$, we first write

$$\begin{aligned}
&\mathbb{E}_k \left\|\boldsymbol{v}^{(k+1)}\right\|_2^2 \\
&= \mathbb{E}_k \left\|\boldsymbol{g}^{(k+1)} + \mu\boldsymbol{\theta}^{(k)}\right\|_2^2 \\
&= \mathbb{E}_k \left\|\boldsymbol{g}^{(k+1)} + \mu\boldsymbol{\theta}^{(k)} - \nabla\bar{r}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}\right\|_2^2 + \left\|\nabla\bar{r}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}\right\|_2^2 \\
&= \mathbb{E}_k \left\|\boldsymbol{g}^{(k+1)} - \nabla\bar{\ell}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}\right\|_2^2 + \left\|\nabla\bar{r}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}\right\|_2^2 \\
&\leq \mathbb{E}_k \left\|\boldsymbol{g}^{(k+1)} - \nabla\bar{\ell}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}\right\|_2^2 + 2\left\|(\nabla\bar{r}(\boldsymbol{\theta}^{(k)}) - \nabla\bar{r}(\boldsymbol{\theta}^\star))^\top \boldsymbol{q}^{(k)}\right\|_2^2 + 2\left\|\nabla\bar{r}(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^{(k)}\right\|_2^2.
\end{aligned}$$

We control the first and the third term above. For the first term, using the notation of Algorithm 2, use the vector-valued Cauchy-Schwarz inequality to achieve

$$\begin{aligned}
\mathbb{E}_k \left\|\boldsymbol{g}^{(k+1)} - \nabla\bar{\ell}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{q}^{(k)}\right\|_2^2 &= \mathbb{E}_k \left\|\textstyle\sum_{m=1}^M q_m^{(k)}(\nabla\ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)}) - \nabla\bar{\ell}_m(\boldsymbol{\theta}^{(k)}))\right\|_2^2 \\
&\leq \left\|\boldsymbol{q}^{(k)}\right\|_2^2 \sum_{m=1}^m \mathbb{E}_k \left\|\nabla\ell_m^{(k+1)}(\boldsymbol{\theta}^{(k)}) - \nabla\bar{\ell}_m(\boldsymbol{\theta}^{(k)})\right\|_2^2 \\
&\leq \left\|\boldsymbol{q}^{(k)}\right\|_2^2 \sum_{m=1}^M \frac{(n_m - b_m)\bar{\sigma}_m^2}{(n_m - 1)b_m},
\end{aligned}$$

where the last line follows from Assumption 2.10.3. For the third term, using that $\nabla\bar{r}(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^\star = \mathbf{0}$ at the optimum, we have that

$$\begin{aligned}
\left\|\nabla\bar{r}(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^{(k)}\right\|_2^2 &= \left\|\nabla\bar{r}(\boldsymbol{\theta}^\star)^\top (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)\right\|_2^2 \\
&\leq \left\|\nabla\bar{r}(\boldsymbol{\theta}^\star)^\top\right\|_{2,2}^2 \left\|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\right\|_2^2,
\end{aligned}$$

and by using $\left\|\nabla\bar{r}(\boldsymbol{\theta}^\star)^\top\right\|_{2,2} = \left\|\nabla\bar{r}(\boldsymbol{\theta}^\star)\right\|_{2,2} = s_{\max}\left(\bar{r}(\boldsymbol{\theta}^\star)\right)$, we complete the proof. □

Having established all the required bounds, we return to the overall analysis. We will adopt a Lyapunov stability argument for large shift costs, as in Section 2.6. The Lyapunov function will be

$$V^{(k)} = \left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2 + \sum_{m=1}^M c_m \sum_{i \in \mathcal{I}_m} \left\|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2, \tag{2.53}$$

where $c_1, \ldots, c_M$ are to-be-specified constants. We can easily bound the evolution of the M

additional terms for a given batch size sequence, as for any $i \in \mathcal{I}_m$,

$$\mathbb{E}_k\left[\|\hat{\boldsymbol{\theta}}_i^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2\right] = \frac{b_m}{n_m}\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 + \left(\frac{n_m - b_m}{n_m}\right)\|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\|_2^2$$

which then implies that

$$\mathbb{E}_k\left[\sum_{i\in\mathcal{I}_m}\|\hat{\boldsymbol{\theta}}_i^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2\right] = b_m\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 + \left(\frac{n_m - b_m}{n_m}\right)\sum_{i\in\mathcal{I}_m}\|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\|_2^2. \tag{2.54}$$

The result below is a "large shift cost" regime analysis, where the condition on the smoothing parameter can in fact be optimized with respect to the per-group batch size parameter. The convergence will not be exact, but will have a convergence radius that scales with the learning rate.

Proposition 2.10.1. *Recalling the constants C_1 and C_2 from Assumption 2.10.4, assume that*

$$\eta \le \min\left\{\frac{C_1}{2C_2(\bar{M}+\mu)}, \frac{2C_1^2\sum_{m=1}^M G_m^2}{C_2 s_\star^2 \mu}\right\} \text{ and } \nu \ge \frac{8\sum_{m'=1}^M G_{m'}^2}{C_1\mu}\sqrt{\max_m n_m}.$$

Define the half-life

$$\tau := 2\max\left\{\frac{1}{\eta\mu}, \frac{n_1}{b_1}, \ldots, \frac{n_M}{b_M}\right\}. \tag{2.55}$$

Then, by setting $c_m = \frac{\eta\mu n_m}{4M}$ for $m = 1, \ldots, M$ in (2.53), it holds that

$$\mathbb{E}_k\left[V^{(k+1)}\right] \le (1-\tau^{-1})V^{(k)} + \eta^2\left(\max_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{q}\|_2^2\right)\sum_{m=1}^M \frac{(n_m - b_m)\bar{\sigma}_m^2}{(n_m - 1)b_m}.$$

Proof. By combining Assumption 2.10.4, Lemma 2.10.1, Lemma 2.10.2, and (2.54), we have

that

$$\begin{aligned}
\mathbb{E}_k\left[V^{(k+1)}\right] \leq & \left(1-\eta\mu+\sum_{m=1}^{M} c_m b_m\right)\left\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{\star}\right\|_2^2 \\
& -2\eta C_1(\boldsymbol{q}^{(k)}-\boldsymbol{q}^{\star})^{\top}(\bar{\ell}(\boldsymbol{\theta}^{(k)})-\bar{\ell}(\boldsymbol{\theta}^{\star})) && (2.56)\\
& -\eta\left(\frac{C_1}{\bar{M}+\mu}-2\eta C_2\right)\left\|(\nabla\bar{r}(\boldsymbol{\theta}^{(k)})-\nabla\bar{r}(\boldsymbol{\theta}^{\star}))^{\top}\boldsymbol{q}^{(k)}\right\|_2^2 && (2.57)\\
& +\eta^2 C_2\left(\max_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{q}\|_2^2\right)\sum_{m=1}^{M}\frac{(n_m-b_m)\bar{\sigma}_m^2}{(n_m-1)b_m} \\
& +2\eta^2 C_2 s_{\max}\left(\nabla\bar{r}(\boldsymbol{\theta}^{\star})\right)^2\left\|\boldsymbol{q}^{(k)}-\boldsymbol{q}^{\star}\right\|_2^2 && (2.58)\\
& +\sum_{m=1}^{M} c_m\left(\frac{n_m-b_m}{n_m}\right)\sum_{i\in\mathcal{I}_m}\|\hat{\boldsymbol{\theta}}_i^{(k)}-\boldsymbol{\theta}^{\star}\|_2^2.
\end{aligned}$$

The condition $\eta \leq \frac{C_1}{2C_2(\bar{M}+\mu)}$ implies that the term (2.57) is non-positive. To handle (2.56) and (2.58), we use the smoothness of $q^{\mathrm{opt}}(\cdot)$ (via Lemma 2.4.1) and Assumption 2.10.2. First, apply Young's inequality with parameter $\alpha > 0$ to achieve

$$\begin{aligned}
2\eta(\boldsymbol{q}^{(k)}-\boldsymbol{q}^{\star})^{\top}(\bar{\ell}(\boldsymbol{\theta}^{(k)})-\bar{\ell}(\boldsymbol{\theta}^{\star})) &\leq \eta\alpha\left\|\boldsymbol{q}^{(k)}-\boldsymbol{q}^{\star}\right\|_2^2+\frac{\eta}{\alpha}\left\|\bar{\ell}(\boldsymbol{\theta}^{(k)})-\bar{\ell}(\boldsymbol{\theta}^{\star})\right\|_2^2 \\
&\leq \eta\alpha\left\|\boldsymbol{q}^{(k)}-\boldsymbol{q}^{\star}\right\|_2^2+\frac{\eta}{\alpha}\left(\textstyle\sum_m G_m^2\right)\left\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{\star}\right\|_2^2.
\end{aligned}$$

We also have that

$$\left\|\boldsymbol{q}^{(k)}-\boldsymbol{q}^{\star}\right\|_2^2 \leq \frac{1}{\nu^2}\sum_{m=1}^{M} G_m^2\sum_{i\in\mathcal{I}_m}\|\hat{\boldsymbol{\theta}}_i^{(k)}-\boldsymbol{\theta}^{\star}\|_2^2.$$

Combining these bounds together and letting $s_{\max}(\nabla \bar{r}(\boldsymbol{\theta}^\star)) = s_\star$, we have ultimately that

$$\begin{aligned}\mathbb{E}_k\left[V^{(k+1)}\right] \leq &\left[1 - \eta\mu + \sum_{m=1}^M \left(c_m b_m + \frac{\eta C_1 G_m^2}{\alpha}\right)\right] \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\ &+ \sum_{m=1}^M c_m \left[1 + \frac{\eta G_m^2}{c_m \nu^2}\left(C_1\alpha + 2\eta C_2 s_\star^2\right) - \frac{b_m}{n_m}\right] \sum_{i \in \mathcal{I}_m} \|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\ &+ \eta^2 C_2 \left(\max_{\boldsymbol{q} \in \mathcal{Q}} \|\boldsymbol{q}\|_2^2\right) \sum_{m=1}^M \frac{(n_m - b_m)\bar{\sigma}_m^2}{(n_m - 1) b_m}.\end{aligned}$$

The convergence rate will be determined by setting the constants in the square brackets. When setting $c_m = \frac{\eta G_m^2 \mu}{4 b_m \sum_{m'} G_{m'}^2}$ and $\alpha = \frac{4 C_1 \sum_m G_m^2}{\mu}$, we have that

$$1 - \eta\mu + \sum_{m=1}^M \left(c_m b_m + \frac{\eta C_1 G_m^2}{\alpha}\right) = 1 - \frac{\eta\mu}{2}$$

and

$$\frac{\eta G_m^2}{c_m \nu^2}\left(C_1\alpha + 2\eta C_2 s_\star^2\right) = \frac{8 b_m \sum_{m'} G_{m'}^2}{\mu \nu^2}\left(\frac{2 C_1^2 \sum_{m'} G_{m'}^2}{\mu} + \eta C_2 s_\star^2\right) \overset{\text{want}}{\leq} \frac{b_m}{2 n_m}.$$

The inequality is accomplished when the conditions

$$\eta \leq \frac{2 C_1^2 \sum_m G_m^2}{C_2 s_\star^2 \mu} \text{ and } \nu \geq \frac{8 \sum_{m'=1}^M G_{m'}^2}{C_1 \mu}\sqrt{\max_m n_m}$$

hold, which achieves the desired rate and completes the proof. □

We highlight the smoothness condition, which, by leveraging non-uniformity of $G_1, \ldots, G_M$, scales as $\sqrt{\max_m n_m}$ times the sum of squared Lipschitz constants (as opposed to the total sample size times maximum Lipschitz constant squared seen in the instance-level analysis earlier in this chapter). We will observe a similar theme in Chapter 3, where we group coordinates of the loss vector and are able to sample entire blocks of coordinates on each iteration. To simplify the upcoming discussion, we assume that $C_1 = C_2 = 1$.

Next, iterating the result of Proposition 2.10.1 for a total iteration bound of t, we have

that

$$\mathbb{E}_0\left[V^{(t)}\right] \leq (1-\tau^{-1})^t V^{(0)} + \left(\sum_{k=0}^{t-1}(1-\tau^{-1})^k\right)\eta^2\left(\max_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{q}\|_2^2\right)\sum_{m=1}^{M}\frac{(n_m-b_m)\bar{\sigma}_m^2}{(n_m-1)b_m}$$
$$\leq (1-\tau^{-1})^t V^{(0)} + \tau\eta^2\left(\max_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{q}\|_2^2\right)\sum_{m=1}^{M}\frac{(n_m-b_m)\bar{\sigma}_m^2}{(n_m-1)b_m}.$$

While τ depends on the per-group batch sizes, if the dominant condition in (2.55) is the first, then the radius scales as

$$\frac{\eta}{\mu}\left(\max_{\boldsymbol{q}\in\mathcal{Q}}\|\boldsymbol{q}\|_2^2\right)\sum_{m=1}^{M}\frac{(n_m-b_m)\bar{\sigma}_m^2}{(n_m-1)b_m},$$

which, in terms of the $O(\eta/\mu)$ dependence, resembles state-of-the-art stochastic gradient descent analyses in similar settings [Cutler et al., 2023]. At a high level, the notion of training examples partitioned into logical groups and adapting to their non-uniformity is a theme that will be discussed further in Chapter 3. Here, we leveraged the per-group batch sizes and operated in a primal-only SGD-like framework, pursuing a practical and simple algorithm. As alluded to above, in Chapter 3, we pursue other ideas such as adaptive sampling and block coordinate-wise updates to achieve improved complexities in the primal-dual setting.

2.10.3 Experiments

We provide a numerical benchmark to apply Algorithm 3 in practice. As mentioned before, while Algorithm 2 and Algorithm 3 perform almost equivalently in practice, Algorithm 3 is much easier to implement for practitioners operating within an existing PyTorch workflow, as the indices of data points may not be provided by the sampling mechanism. Consider the following setup, which resembles the instance-level distribution shift benchmark from Figure 2.6 (Section 2.9).

- **Data:** A subset of $100,000$ points from the Amazon Reviews dataset from the WILD Distribution Shift Benchmark. They are split into $50,000$ train and $50,000$ test examples, respectively. Each data point represents a review of an Amazon product written

in natural language, along with a categorical label indicating the number of stars given from 1 to 5. The review is augmented with metadata, including the product index, product category, and year. The text is discretized using the `bert-base-uncased` tokenizer and truncated to 100 tokens. The group labels are given by the product category, and may shift in proportion between the train and test sets.

- **Model:** The model architecture is a 2-layer transformer network, where each transformer block uses 8 heads, an embedding dimension of 512, and pre-normalization.

- **Optimizer:** Algorithm 3 with the base optimizer Adam with default moment and variance parameters (β_1, β_2) are used without weight decay and a fixed learning rate of 3×10^{-4}. The spectrum is chosen to be either the 1.5 or 2-extremile compared to a group-wise empirical risk minimization baseline (or $\mathcal{Q} = \{\mathbf{1}/M\}$).

The results are shown in Figure 2.10. We observe that the average performance over the test set is indistinguishable between the various objectives, which include group-wise empirical risk minimization, as well as the 1.5-extremile and 2-extremile. However, the worst group-wise accuracy for each of the spectral risk measure-based group DRO models achieves a 5% increase over the ERM baseline.

To motivate the upcoming discussion on the relationship between worst group-wise error and fairness, compare the experiment above to Figure 2.5. We used two datasets, and in `diabetes`, we use gender as the protected attribute A, whereas in `acsincome` we use race as the protected attribute. It is natural to consider the protected attribute to be a group label, and for statistical parity (SP) to be a measure of worst-case group-wise performance. However, the instance-level DRO algorithms, while evaluated on SP, were never supplied the protected attribute (i.e., the group label) at training time, which is a common constraint in fairness benchmarks. This differs from the models explored in this section, for which observing the group label is an essential component of the algorithm. If the protected attribute is available to the model, then group DRO presents an alternative formulation that

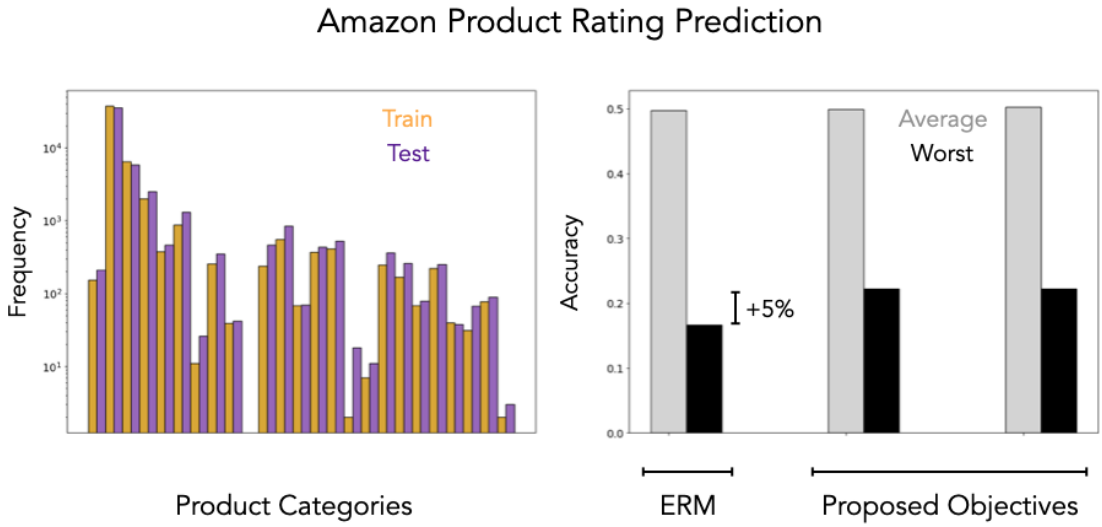

Figure 2.10: **Amazon Reviews Data.** The left panel shows the group-wise proportions in the train and test sets, respectively. The right panel shows the average and worst-case group-wise test accuracy. The proposed objectives are the 1.5 or 2-extremile. Even though the group proportions have a minor shift in most product categories, a moderate worst-case accuracy increase is observed for the distributionally robust variants.

explicitly promotes fairness via the group reweighting structure. We provide details on these group-wise metrics and others in the next subsection.

2.10.4 Other Group-Wise Performance Metrics

We conclude this section by commenting on our group DRO formulation in the context of other group-level performance measures, such as class-specific performance and algorithmic fairness. Consider the setting of supervised learning, where $\Xi = \mathcal{X} \times \mathcal{Y}$ for feature space $\mathcal{X}$ and label space $\mathcal{Y}$. Furthermore, the sake of exposition, assume that $\mathcal{Y} = \{0, 1\}$, so that for a parametrized function $f_{\boldsymbol{\theta}} : \mathcal{X} \to (0, 1)$ and $\xi = (X, Y)$, we define

$$\ell(\boldsymbol{\theta}, \xi) := Y \log(f_{\boldsymbol{\theta}}(X)) + (1 - Y) \log(1 - f_{\boldsymbol{\theta}}(X)).$$

We interpret $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as the predicted probability that $Y = 1$ given $X = \boldsymbol{x}$. Accordingly, define the predicted value $\widehat{Y}_{\boldsymbol{\theta}}(X) = \mathbb{1}\{f_{\boldsymbol{\theta}}(X) \geq 0.5\}$. One could, in principle, set $A_1 =$

$Y_1, \ldots, A_n = Y_n$, i.e., the group labels are themselves the class labels. Thus, if a class y is harder under θ, or

$$\mathbb{P}_{(X,Y)\sim P}\left[\widehat{Y}_{\boldsymbol{\theta}}(X) = y | Y = y\right] \leq \mathbb{P}_{(X,Y)\sim P}\left[\widehat{Y}_{\boldsymbol{\theta}}(X) = 1 - y | Y = 1 - y\right],$$

then reweighting the group labels toward this label would balance performance across classes. The quantities on the left and right-hand side above are, in fact, the *recall* for classes y and $1 - y$, respectively. Typically, the reweighting parameter might be selected using a held-out validation set. However, to adapt to unknown class imbalances, the group DRO formulation of (2.51) offers a principled solution that simultaneously incorporates many possible class distributions (or *label shifts*) that may be observed during deployment.

Relatedly, in the fairness literature, a common claim of success is that some notion of error is approximately equal across groups. As mentioned above, the group label is more commonly referred to as the "protected attribute". Under $(X, Y) \sim P$, two examples of parameter choice $\theta \in \mathbb{R}^d$ that are fair under *equalized odds* are those that satisfy either false positive or false negative invariance to the attribute A:

$$\mathbb{P}\left[\widehat{Y}_{\boldsymbol{\theta}}(X) = 1 | Y = 0, A = \boldsymbol{a}_1\right] \approx \ldots \approx \mathbb{P}\left[\widehat{Y}_{\boldsymbol{\theta}}(X) = 1 | Y = 0, A = \boldsymbol{a}_M\right]$$
$$\mathbb{P}\left[\widehat{Y}_{\boldsymbol{\theta}}(X) = 0 | Y = 1, A = \boldsymbol{a}_1\right] \approx \ldots \approx \mathbb{P}\left[\widehat{Y}_{\boldsymbol{\theta}}(X) = 0 | Y = Y, A = \boldsymbol{a}_M\right].$$

This motivates the use of (2.51), as a parameter with approximately equal performance on the group conditional risk $\mathbb{E}_P\left[\ell(\boldsymbol{\theta}, \xi) | A = \boldsymbol{a}\right]$ across all $\boldsymbol{a} \in \mathcal{A}$ cannot be largely increased by reweighing. Note that we explored this connection experimentally at the instance-level in Section 2.9.

Finally, there is an alternative utilization of group information which is similar to (2.51) algorithmically but distinct in terms of the underlying statistical problem. Observe that at an instance level, (2.51) applies equal weight to all examples in the group, where the distribution over groups is learned. The assumption that the within-group distribution on Ξ is uniform regardless of the distributional shift. Consider instead having *known* target

weights $q_1^\star, \ldots, q_M^\star$ over groups, but no reasonable justification that the per-group distribution over Ξ is uniform. This setting can be formulated as an instance-level DRO problem with uncertainty set

$$\mathcal{B} := \left\{ \beta \in \mathcal{B}^{\mathrm{DRO}} : \int_{\Xi \times \{\boldsymbol{a}_m\}} \beta(\boldsymbol{x}, \boldsymbol{a}) \, \mathrm{d}P_n(\boldsymbol{x}, \boldsymbol{a}) = q_m^\star \ \forall m \in [M] \right\}.$$

This viewpoint is a combination of the "known" distribution framework direct likelihood estimation (see Section 2.8) and the "unknown" target distribution. We may know our target distribution up to a partition of the space Ξ of resolution M with probabilities $\{q_i^\star\}_{i=1}^M$, but wish to account for many possible distributions within each cell of the partition.

2.11 Possible Extensions

2.11.1 Changes in the Support

We consider here a version of the objective in which the learned parameter can account for changes in the support of the input distribution, instead of just changes in the relative weights on each training example. Consider the squared loss

$$\ell(\boldsymbol{\theta}, (\boldsymbol{x}, y)) := \frac{1}{2} \left(y - \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle \right)^2,$$

which is not Lipschitz continuous unless $\boldsymbol{\theta}$ is restricted to a compact domain. We perturb $\boldsymbol{x}$ by a random vector $\boldsymbol{\varepsilon}$ realized in $\mathbb{R}^d$, whose mean is denoted $\boldsymbol{\mu}$ and whose covariance matrix is denoted $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$. We consider the maximum expected loss achievable by perturbation distributions whose mean and covariance are bounded in norm-balls of radius r and Δ, respectively. The $\boldsymbol{\mu}$ is bounded in the Euclidean norm $\|\cdot\|_2$ and the covariance is bounded in spectral norm $\|\cdot\|_{2,2}$. While the specific distribution of $\boldsymbol{\varepsilon}$ does not affect the objective beyond the first-two moments, we let $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for simplicity. Given data points $\{\xi_i\}_{i=1}^n$

for $\xi_i = (\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, this gives the objective

$$\mathcal{L}_{r,\Delta}(\boldsymbol{\theta}) := \max_{\boldsymbol{q} \in \mathcal{Q}} \max_{\substack{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2 \le r, \\ \boldsymbol{\Sigma}:\|\boldsymbol{\Sigma}\|_{2,2} \le \Delta}} \sum_{i=1}^{n} \frac{q_i}{2} \mathbb{E}_{\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i \rangle)^2 \right], \tag{2.59}$$

which is still convex in $\boldsymbol{\theta}$. By expanding this term, we can in fact remove the optimization over $\boldsymbol{\Sigma}$, whereas the dependence on $\boldsymbol{\mu}$ can be simplified using a univariate optimization problem.

Proposition 2.11.1. *For any $r, \Delta > 0$, it holds that*

$$\mathcal{L}_{r,\Delta}(\boldsymbol{\theta}) := \max_{\boldsymbol{q} \in \mathcal{Q}} \left[\langle \boldsymbol{q}, \ell(\boldsymbol{\theta}) \rangle + \max \left\{ 0, r \left(\langle \boldsymbol{\theta}, \bar{\boldsymbol{x}}_{\boldsymbol{q}} \rangle - \bar{y}_{\boldsymbol{q}} \right) \|\boldsymbol{\theta}\|_2 + \frac{r^2}{2} \|\boldsymbol{\theta}\|_2^2 \right\} \right] + \frac{\Delta}{2} \|\boldsymbol{\theta}\|_2^2 \tag{2.60}$$

for $\ell_i(\boldsymbol{\theta}) = \frac{1}{2}(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)$, $\bar{y}_{\boldsymbol{q}} = \sum_{i=1}^n q_i y_i$ and $\bar{\boldsymbol{x}}_{\boldsymbol{q}} = \sum_{i=1}^n q_i \boldsymbol{x}_i$. Furthermore, the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ achieving (2.60) for $\boldsymbol{\theta} \neq \boldsymbol{0}$ are given by

$$\begin{aligned} \boldsymbol{\mu}^\star(\boldsymbol{\theta}) &= \left(\arg\max_{a \in \{0, r\}} a \left(\langle \boldsymbol{\theta}, \bar{\boldsymbol{x}}_{\boldsymbol{q}} \rangle - \bar{y}_{\boldsymbol{q}} \right) \|\boldsymbol{\theta}\|_2 + \frac{a^2}{2} \|\boldsymbol{\theta}\|_2^2 \right) \cdot \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} \\ \boldsymbol{\Sigma}^\star(\boldsymbol{\theta}) &= \frac{\Delta}{\|\boldsymbol{\theta}\|_2^2} \boldsymbol{\theta}\boldsymbol{\theta}^\top + \boldsymbol{\Sigma}_0, \end{aligned}$$

where $\boldsymbol{\Sigma}_0$ is any positive semidefinite matrix satisfying $\operatorname{range}(\boldsymbol{\Sigma}_0) = \operatorname{span}\{\boldsymbol{\theta}\}^\perp$.

Proof. First, expand the summands of the objective (2.59) to achieve

$$\begin{aligned} &\frac{1}{2} \mathbb{E}_{\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i \rangle)^2 \right] \\ &= \frac{1}{2} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)^2 - (y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) \mathbb{E}_{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[\langle \boldsymbol{\theta}, \boldsymbol{\varepsilon}_i \rangle \right] + \frac{1}{2} \mathbb{E}_{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \left[\langle \boldsymbol{\theta}, \boldsymbol{\varepsilon}_i \rangle^2 \right] \\ &= \frac{1}{2} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)^2 - (y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} + \frac{1}{2} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle^2 . \end{aligned}$$

Thus, it holds that $\mathcal{L}_{r,\Delta}(\boldsymbol{\theta})$ is equal to

$$\max_{\boldsymbol{q} \in \mathcal{Q}} \left\{ \langle \boldsymbol{q}, \ell(\boldsymbol{\theta}) \rangle + \max_{\boldsymbol{\mu}:\|\boldsymbol{\mu}\|_2 \le r} \left(\langle \boldsymbol{\theta}, \bar{\boldsymbol{x}}_{\boldsymbol{q}} \rangle - \bar{y}_{\boldsymbol{q}} \right) \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle + \frac{1}{2} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle^2 \right\} + \frac{1}{2} \max_{\boldsymbol{\Sigma}:\|\boldsymbol{\Sigma}\|_{2,2} \le \Delta} \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta}$$

The final term above can be computed in closed form, as the objective depends on $\boldsymbol{\Sigma}$ only based on evaluation on the span of $\boldsymbol{\theta}$. Thus, one solution (in the case that $\boldsymbol{\theta} \neq \boldsymbol{0}$) of the

maximization problem is $\boldsymbol{\Sigma} := \frac{\Delta}{\|\boldsymbol{\theta}\|_2^2}\boldsymbol{\theta}\boldsymbol{\theta}^\top$, yielding

$$\frac{1}{2}\max_{\boldsymbol{\Sigma}:\|\boldsymbol{\Sigma}\|_{2,2}\leq\Delta} \boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = \frac{\Delta}{2}\|\boldsymbol{\theta}\|_2^2,$$

which is simply a regularization term. The equality above also holds when $\boldsymbol{\theta} = \mathbf{0}$. Note that even though the expectation was computed under a Gaussian distribution (which requires $\boldsymbol{\Sigma}$ to be full rank), we may make it full rank by simply completing the singular value decomposition with arbitrary positive singular values that are strictly less than Δ. On the other hand, we see that the objective also depends on $\boldsymbol{\mu}$ only through $\langle\boldsymbol{\theta}, \boldsymbol{\mu}\rangle$ (also assuming that $\boldsymbol{\theta} \neq 0$), we may reparametrize the optimization problem by considering

$$\boldsymbol{\mu} = a \cdot \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}$$

for $a \in [0, r]$. To compute $\boldsymbol{\mu}$, we must then solve the univariate problem

$$\max_{0\leq a\leq r} a\left(\langle\boldsymbol{\theta}, \bar{\boldsymbol{x}}_{\boldsymbol{q}}\rangle - \bar{y}_{\boldsymbol{q}}\right)\|\boldsymbol{\theta}\|_2 + \frac{a^2}{2}\|\boldsymbol{\theta}\|_2^2 = \max\left\{0, r\left(\langle\boldsymbol{\theta}, \bar{\boldsymbol{x}}_{\boldsymbol{q}}\rangle - \bar{y}_{\boldsymbol{q}}\right)\|\boldsymbol{\theta}\|_2 + \frac{r^2}{2}\|\boldsymbol{\theta}\|_2^2\right\},$$

where the solution follows by simply considering the two endpoints, as the objective it is the maximum of a strongly convex function over an interval. □

In order to preserve the dual linearity of the saddle-point objective in $\boldsymbol{q}$, one could alternatively maximize over $\boldsymbol{q}$ and $\boldsymbol{\mu}$ in the inner-most objective by considering $\mathcal{L}_{r,\Delta}(\boldsymbol{\theta}) = \max_{\boldsymbol{q},\boldsymbol{\mu}} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{q}, \boldsymbol{\mu})$ for

$$\begin{aligned}\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{q}, \boldsymbol{\mu}) &= \langle\boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle + \left(\langle\boldsymbol{\theta}, \bar{\boldsymbol{x}}_{\boldsymbol{q}}\rangle - \bar{y}_{\boldsymbol{q}}\right)\langle\boldsymbol{\theta}, \boldsymbol{\mu}\rangle + \frac{1}{2}\langle\boldsymbol{\theta}, \boldsymbol{\mu}\rangle^2 + \frac{\Delta}{2}\|\boldsymbol{\theta}\|_2^2 \\ &= \langle\boldsymbol{q}, \ell(\boldsymbol{\theta})\rangle - \bar{y}_{\boldsymbol{q}}\boldsymbol{\mu}^\top\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^\top\left(\left(2\bar{\boldsymbol{x}}_{\boldsymbol{q}} + \boldsymbol{\mu}\right)\boldsymbol{\mu}^\top + \Delta\boldsymbol{I}\right)\boldsymbol{\theta}\end{aligned}$$

as the solution for $\boldsymbol{\mu}$ is computable in closed form. Thus, by modifying the optimization problem only slightly, we can account for some degree of robustness to input noise, effectively changing the support of the input distribution.

2.11.2 A Shrinking Uncertainty Set

As mentioned in Section 2.8, there are several applications (such as empirical likelihood) in which the size of the uncertainty set is dependent on the sample size n. If one wishes to use an estimator learned from a distributionally robust objective that also decreases the amount of uncertainty as more data arrives, it is worthwhile to know the necessary conditions on this decay to match various rates.

To do so, we change notation slightly to better agree with standard generalization bounds (e.g., Wainwright [2019]) and consider a generic class $f \in \mathcal{F}$ of measurable functions $f : \Xi \to \mathbb{R}$ (indicating losses) and consider an uncertainty set $\mathcal{Q}_n \subseteq \Delta^{n-1}$ dependent on n. Finally, consider an empirical DR risk minimizer

$$\hat{f}_n \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \left[R_n(f) := \max_{\boldsymbol{q} \in \mathcal{Q}_n} \sum_{i=1}^{n} q_i f(\xi_i) - \nu \operatorname{Reg}(\boldsymbol{q}) \right].$$

We also employ the standard empirical process notation $P(f) := \mathbb{E}_{\xi \sim P}[f(\xi)]$ and let P_n be the empirical measure of the data. We have the following guarantee.

Lemma 2.11.1. *Assume that $f(\xi) \leq B$ with P-probability one for every $f \in \mathcal{F}$ and that $\operatorname{Reg}(\mathbf{1}/n) = 0$. Then, for any $f \in \mathcal{F}$, it holds that*

$$P(\hat{f}_n) - P(f) \leq (P - P_n)(\hat{f}_n - f) + B \max_{\boldsymbol{q} \in \mathcal{Q}_n} \|\boldsymbol{q} - \mathbf{1}/n\|_1 . \tag{2.61}$$

Proof. First, write the decomposition

$$\begin{aligned} P(\hat{f}_n) - P(f) &= (P - P_n)(\hat{f}_n) + \underbrace{(P_n - R_n)(\hat{f}_n)}_{\leq 0} + \underbrace{R_n(\hat{f}_n) - R_n(f)}_{\leq 0} \\ &\quad + (R_n - P_n)(f) + (P_n - P)(f) \\ &\leq (P - P_n)(\hat{f}_n - f) + (R_n - P_n)(f), \end{aligned}$$

where we used in the first line that $(P_n - R_n)(f) \leq 0$ for all $f \in \mathcal{F}$ because $\operatorname{Reg}(\mathbf{1}/n) = 0$,

whereas $R_n(\hat{f}_n) - R_n(f) \leq 0$ due to the optimality of $\hat{f}_n$. Next, write

$$\begin{aligned}(R_n - P_n)(f) &= \max_{\boldsymbol{q}\in\mathcal{Q}_n} \sum_{i=1}^{n} (q_i - 1/n) f(\xi_i) - \nu \operatorname{Reg}(\boldsymbol{q}) \\ &\leq \max_{\boldsymbol{q}\in\mathcal{Q}_n} \sum_{i=1}^{n} (q_i - 1/n) f(\xi_i) \\ &\leq B \max_{\boldsymbol{q}\in\mathcal{Q}_n} \|\boldsymbol{q} - \mathbf{1}/n\|_1 ,\end{aligned}$$

completing the proof. □

Notice that (2.61) contains the standard empirical process term, which can be bounded in a variety of ways, which do not necessarily depend on the fact that $\hat{f}_n$ is the minimizer of the average loss. For example, the standard "slow rate" approach would immediately apply

$$(P - P_n)(\hat{f}_n - f) \leq 2 \sup_{f\in\mathcal{F}} |(P_n - P)f| ,$$

where we assume that the right-hand side remains measurable. For a localization-based approach, first assume that there exists a unique minimizer

$$f_\star := \arg\min_{f\in\mathcal{F}} P(f).$$

Define the following *excess risk* term as

$$\mathcal{E}_n(f) := (P - P_n)(f - f_\star).$$

This term differs from the usual excess risk process $f \mapsto P(f - f_\star)$, but is defined in this way to apply the arguments of, for instance, Ohn and Kim [2025, Section 2] to control it. Note that these do not depend on the fact that $\hat{f}_n$ was generated by optimizing the average loss, so they can be applied for this setting. In various settings, this can be shown to be bounded by an $O(1/n)$ term with high probability. Thus, it only remains to control the second term of (2.61). In the case of spectral risk measures, we may achieve the following bounds in terms of Kullback-Liebler (KL) or χ^2-divergence.

Lemma 2.11.2. *Let $\mathcal{Q}_n$ be an ambiguity set determined by a spectral risk measure with spectrum s_n on $(0,1)$ (see Section 2.3). Then, it holds that*

$$\max_{\boldsymbol{q}\in\mathcal{Q}_n} \|\boldsymbol{q} - \mathbf{1}/n\|_1 \leq \sqrt{\frac{1}{2}\int_0^1 s_n(t)\log s_n(t)\,dt}.$$

Proof. First, the maximizer of the strongly convex function $\boldsymbol{q} \mapsto \|\boldsymbol{q} - \mathbf{1}/n\|_1^2$ occurs on the boundary of the closed, convex permutahedron $\mathcal{Q}_n \equiv \mathcal{Q}(\sigma)$. Furthermore, because this objective is permutation invariant, it holds that

$$\max_{\boldsymbol{q}\in\mathcal{Q}_n} \|\boldsymbol{q} - \mathbf{1}/n\|_1 = \|\sigma - \mathbf{1}/n\|_1 .$$

Then, it holds that

$$\begin{aligned}
\|\sigma - \mathbf{1}/n\|_1 &= \sum_{i=1}^n \left| \int_{(i-1)/n}^{i/n} s_n(t)\,\mathrm{d}t - 1/n \right| \\
&= \sum_{i=1}^n \left| \int_{(i-1)/n}^{i/n} (s_n(t) - 1)\,\mathrm{d}t \right| \\
&\leq \sum_{i=1}^n \int_{(i-1)/n}^{i/n} |s_n(t) - 1|\,\mathrm{d}t \\
&= \|s_n - 1\|_1 .
\end{aligned}$$

By an immediate application of Pinkser's inequality,

$$\|s_n - 1\|_1 \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(s_n\|1)},$$

where the number 1 indicates the uniform distribution on $(0,1)$. Then, compute the KL-divergence term and use that $\log(1) = 0$ to complete the proof. □

We may use the upper bound in the lemma above to compute the decay of the term to zero as $n \to \infty$, and ensure that it matches the first term of (2.61). To see this, let us recall the examples from Section 2.3 and determine what values of the risk parameter are necessary to enforce $\max_{\boldsymbol{q}\in\mathcal{Q}_n} \|\boldsymbol{q} - \mathbf{1}/n\|_1 = O(n^{-\alpha})$.

- **Superquantile:** For $\tau_n \in (0, 1]$, the τ_n-*superquantile* [Rockafellar and Royset, 2013], is specified by $s_n(t) = \frac{1}{1-\tau_n} \mathbb{1}\{\tau_n \leq t \leq 1\}$. Then, it holds that

$$\sqrt{\int_0^1 s_n(t) \log s_n(t) \, \mathrm{d}t} = \sqrt{\log\left(\frac{1}{1-\tau_n}\right)} \implies \tau_n = O\left(1 - e^{-\left(1/n^{2\alpha}\right)}\right).$$

- **Extremile:** For $r_n > 1$, the r_n-*extremile* [Daouia et al., 2019] is specified by $s_n(t) = r_n t^{r_n - 1}$. Then, it holds that

$$\sqrt{\int_0^1 s_n(t) \log s_n(t) \, \mathrm{d}t} = \sqrt{\log r_n - \left(\frac{r_n - 1}{r_n}\right)} \implies r_n = O\left(e^{\left(1/n^{2\alpha}\right)}\right).$$

For $\alpha = 1$, these indicate extremely fast convergence of $\tau_n \to 0$ or $r_n \to 1$ in order to match the "fast" rate of $O(1/n)$ for the excess risk term.

One avenue for future work in this area is to establish fast rates for a fixed uncertainty set, describing convergence to the minimizer of the population (distributionally robust) risk. There is work in this area showing matching upper and lower bounds in the case of a particular family of f-divergences [Duchi and Namkoong, 2021], but similar results for spectral risk measure uncertainty sets remain an open question.

2.11.3 Concentration under Heavy Tails

In risk-averse applications, a common consideration is the heavy-tailedness of the distribution being analyzed. As commented in Mikosch [1999], while the term "heavy-tailed" does not have a universal definition, two classes of distributions are ubiquitous cases of interest: random variables with regularly varying tails, and sub-exponential random variables. They are defined below.

Definition 2.11.1 (Regularly Varying Tails)**.** For a real-valued random variable X with CDF F, we say that X has *regularly varying tails* with index $\alpha \in \mathbb{R}$ if

$$\lim_{x \to \infty} \frac{1 - F(tx)}{1 - F(x)} = t^\alpha \text{ for all } t > 0.$$

Definition 2.11.2 (Sub-Exponential R.V.)**.** We say that a real-valued, non-negative random variable X is *sub-exponential* if there exist positive constants a and b such that

$$\log \mathbb{P}\left[X \geq x\right] \leq \log a - bx.$$

The relationship between the two classes of random variables is the following. Let $X_1, \ldots, X_n$ be independent and identically distributed copies of a real-valued, non-negative random variable X. For function f, g, we write $f(x) \sim g(x)$ to mean that $\lim_{x\to\infty} f(x)/g(x) = 1$. It holds that both random variables with regularly varying tails and sub-exponential random variables satisfy

$$\mathbb{P}\left[\sum_{i=1}^n X_i > x\right] \sim n\mathbb{P}\left[X > x\right] \sim \mathbb{P}\left[\max_i X_i > x\right] \text{ as } x \to \infty. \tag{2.62}$$

In fact, (2.62) is an equivalent (albeit qualitative) condition to being sub-exponential [Mikosch, 1999]. In words, (2.62) captures the intuition that the maximum of a collection of i.i.d. heavy-tailed random variables contributes nearly all of the value in the sum. As such, the sub-exponential distribution is seen as a *boundary* between light- and heavy-tailed distributions.

Recently, Vladimirova et al. [2020] and Kuchibhotla and Chakrabortty [2022] independently proposed a class of random variables with Weibull-like tails, which has connections to both notions of heavy-tailed distributions given above.

Definition 2.11.3 (Sub-Weibull R.V.)**.** We say that a real-valued, non-negative random variable X is *sub-exponential* if there exist positive constants a, b, and ζ such that

$$\log \mathbb{P}\left[X \geq x\right] \leq \log a - bx^{\zeta}.$$

Equivalently, we may have parameters (K, ζ) such that

$$\|X\|_{\mathbf{L}^p(P)} \leq Kp^{1/\zeta}.$$

It can clearly be seen that by setting $\zeta = 1$, we recover sub-exponentiality. On the other

hand, when X is bounded by B, we have that if

$$\lim_{x\to\infty} \frac{1 - F(B - (tx)^{-1})}{1 - F(B - x^{-1})} = t^{-\alpha} \text{ for all } t > 0.$$

for $\alpha \geq 0$, then we have that the distribution of X is in the maximum domain of attraction of the Weibull distribution with parameter α (often denoted Ψ_α in the extreme value theory literature). Thus, the random variables satisfying Definition 2.11.3 for $\zeta \geq 1$ are relevant representatives of heavy-tailed distributions.

We use two steps for concentration results for spectral risk measures applied to collections of sub-Weibull random variables. To do so, we first relate spectral risk measures to scalings of standard uniform averages. Then, we appeal to concentration bounds for sums of independent sub-Weibull random variables.

Lemma 2.11.3. *Let $x_1, \ldots, x_n \in [0, +\infty)$ define empirical CDF $F_n(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty, x_i]}(x)$ and empirical quantile function $F_n^{-1}(p) := \inf\{x : F_n(x) \geq p\}$. Then, for any integrable spectrum $s : [0,1] \to [0, +\infty)$, we have that*

$$\mathbb{L}_s[F_n] = \sum_{i=1}^n \sigma_i x_{(i)} \leq \frac{s(1)}{n} \sum_{i=1}^n x_i,$$

where $\sigma_i := \int_{(i-1)/n}^{i/n} s(p)\, dp$ and $x_{(1)} \leq \ldots \leq x_{(n)}$ are the order statistics of $(x_1, \ldots, x_n)$.

Proof. Note that we may express the quantile function of an empirical measure using the order statistics of the sample. Then,

$$\mathbb{L}_s[F_n] = \int_0^1 s(p) F_n^{-1}(p) = \sum_{i=1}^n \left(\int_{(i-1)/n}^{i/n} s(p) x_{(\lceil np \rceil)}\, \mathrm{d}p \right) = \sum_{i=1}^n \left(\int_{(i-1)/n}^{i/n} s(t)\, \mathrm{d}p \right) x_{(i)}.$$

. To prove the first claim, we compute the integral

$$\sum_{i=1}^n \left(\int_{(i-1)/n}^{i/n} s(t)\, \mathrm{d}p \right) x_{(i)} = \sum_{i=1}^n \sigma_i x_{(i)}.$$

For the second claim, we apply monotonicity of s and non-negativity of $x_1, \ldots, x_n$ to upper

bound the integral via

$$\sum_{i=1}^{n}\left(\int_{(i-1)/n}^{i/n} s(t)\,\mathrm{d}p\right)x_{(i)} \leq \frac{s(1)}{n}\sum_{i=1}^{n} x_{(i)} = \frac{s(1)}{n}\sum_{i=1}^{n} x_i.$$

□

Thus, spectral risk measures are at most a constant factor larger than the simple average. For the superquantile, we have that $s(1) = \frac{1}{1-\tau}$, whereas for the extremile, we have that $s(1) = r$. From Lemma 2.11.3, we can derive a tail bound for sub-Weibull random variables.

Next, we present a concentration inequality for sub-Weibull random variables proved in Vladimirova et al. [2020]. Before stating the result, we note that tail bounds proved for sums of sub-Weibull random variables in Vladimirova et al. [2020] does not yield concentration of averages, while that of Kuchibhotla and Chakrabortty [2022] does with confidence band of size $O(n^{-1/2})$. In fact, Zhang and Wei [2022] eventually proved sharper concentration rates in terms of problem constants (and the same dependence on n), but with less interpretable constant. Consider the following claim.

Proposition 2.11.2. *Let $X_1, \ldots, X_n \sim P$ be i.i.d. non-negative sub-Weibull random variables with parameters K and $\zeta < 1$. Then, for any $t \geq s(1)\max\{eK, 1\}$, it holds that*

$$\mathbb{P}\left[\mathbb{L}_s[F_n] \geq t\right] \leq Ke\exp\left(-\frac{n}{eK}\left(\frac{t}{s(1)}\right)^{\min\{2,1/\zeta\}}\right).$$

Proof. First, apply Lemma 2.11.3 so that for any $t \geq 0$, we have that

$$\mathbb{P}\left[\mathbb{L}_s[F_n] \geq t\right] \leq \mathbb{P}\left[\sum_{i=1}^{n} X_i \geq nt/s(1)\right].$$

We may then apply Vladimirova et al. [2020, Eq. (7)] (note that K_ζ in their notation is equal to eK in ours) and the condition that $t \geq eKs(1)$ to achieve

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq nt/s(1)\right] \leq Ke\exp\left(-\frac{n}{eK}\min\left\{\frac{t^2}{s^2(1)}, \frac{t^{1/\zeta}}{s^{1/\zeta}(1)}\right\}\right).$$

We use that $t \geq s(1)$ to pass the minimum to the exponent to complete the proof. □

From Proposition 2.11.2, we see that the spectrum can scale the confidence region by a factor of $s(1)$. The restriction $\zeta < 1$ is made to rule out sub-Gaussian and sub-exponential random variables.

2.12 Perspectives & Future Work

In this chapter, we proposed a class of practical algorithms for the distributionally robust optimization (DRO) and proved their linear convergence guarantee for smoothed DRO problems. We paid particular attention to spectral risk measure objectives by first formulating them as DRO objectives and using their quantile-based representation to bound their bias under general conditions (e.g. the loss may be unbounded). Unlike previous DRO formulations, which were typically based on closed balls in KL-divergence [Kumar et al., 2024] or χ^2-divergence [Duchi and Namkoong, 2019], the dual problem for spectral risk measures is easy to solve by subroutines such as sorting or isotonic regression. Furthermore, the hyperparameter $\sigma = (\sigma_1, \ldots, \sigma_n)$ gives the user both visibility and control concerning the optimal dual solution (i.e. the solution will resemble a permutation of σ).

Several directions for future work were outlined in Section 2.11. One direct path would be to generalize the results on squared error loss and changes in the data support from Proposition 2.11.1 to generalized linear model (GLM) loss functions, i.e., those that depend on the parameter $\boldsymbol{\theta}$ through convex functions of the quantities $\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle$ for $i = 1, \ldots, n$. Importantly, this addresses a major limitation of the likelihood ratio-based approaches in this chapter, that is, the absolute continuity constraint of the shifted distributions. The main competing DRO approach is based on Wasserstein metric-based uncertainty sets, which allow for changes in the support but are much less tractable computationally and may only be formulated into finite-dimensional programs under strict conditions [Kuhn et al., 2019]. Incorporating input noise into the likelihood ratio-based approaches of this chapter might yield a "best of both worlds" method in this regard.

Secondly, while the perturbations discussed in this chapter are based on rewighting the observed training data, another interpretation of "changing data" is direct transformation/cor-

ruption of data instances. While this idea is explored using Gaussian noise for vector-valued inputs in Section 2.11.1, formulating the problem using a finite set of naturally-occurring distortions (such as image blur) would be an interesting line of work for applications in vision and language.

In the next chapter, we expand on the ideas mentioned in Section 2.6.4. We recognize that the structure of the DRO objective applies to multiple problems in the data sciences and aim to derive theoretically optimal algorithms using a primal-dual approach. Note that generalizing in this manner does not supplant the work of this chapter; instead, questions such as the large-sample properties of the objective (Section 2.3), uncertainty set selection and computing the maximization map (Section 2.7), and viability on realistic distribution shift scenarios (Section 2.9 and Section 2.11), remain exclusive subjects of Chapter 2.

Chapter 3

ALGORITHMIC EXTENSIONS OF DISTRIBUTIONALLY ROBUST OPTIMIZATION

3.1 Introduction

In Chapter 2, we introduced the distributionally robust optimization problem (2.18) and provided an algorithm and convergence analysis from a "primal-only" viewpoint. Concretely, this refers to the insistence on having a single hyperparameter $\eta > 0$ as a learning rate and using a proof technique called the Lyapunov stability argument (see Section 2.6), which is largely inspired by variance-reduced stochastic algorithms for finite-sum minimization [Johnson and Zhang, 2013, Defazio et al., 2014]. Moreover, the class of algorithms studied in Chapter 2 was motivated primarily by their excellent performance in numerical benchmarks, as observed in Section 2.9. Nevertheless, we would be remiss not to recognize the primal-dual structure of this saddle-point problem, leading to a possibly sharper theoretical convergence guarantee and applicability to more general problems.

Indeed, the DRO problem is an instance of a more general one, which has been a subject of fundamental research in optimization for decades. Consider the saddle point (or min-max) problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \; c(\boldsymbol{x},\boldsymbol{y}) - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x}), \tag{3.1}$$

with respect to primal variable $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and dual variable $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^n$. We refer to $c(\boldsymbol{x},\boldsymbol{y})$ as the *coupled* component, whereas $\phi(\boldsymbol{x})$ and $\psi(\boldsymbol{y})$ are the *individual* components of the objective. Such problems are further categorized as *bilinearly coupled* if $c(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{x}$ for a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, and are called *nonbilinearly coupled* otherwise. While primal-dual algorithms for saddle point problems are still an active area of modern machine learning and

optimization research, nonbilinearly coupled objectives have received much less attention than their bilinearly coupled counterparts. This is due in part to bilinearly coupled objectives being amenable to various innovations such as Chambolle-Pock-style [Chambolle and Pock, 2011] and/or stochastic coordinate-wise updates [Song et al., 2021] that yield an improved runtime. Interestingly, the objective (2.18) has more structure than in a general nonbilinearly coupled problem.

The explorations in this chapter are motivated by the observation that numerous nonbilinearly coupled objectives, including DRO, are in fact linear with respect to one of the two decision variables. Accordingly, insights from the bilinear setting can be used to design efficient algorithms for these "dual-linear" min-max problems[1]. Formally, we introduce the optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \left[\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) := \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})\right], \tag{3.2}$$

in which $f = (f_1, \ldots, f_n)$ has convex components, ϕ is μ-strongly convex ($\mu \geq 0$), and ψ is ν-strongly convex ($\nu \geq 0$). We further require that $\mathcal{Y} \subseteq \{\boldsymbol{y} \in \mathbb{R}^n : y_j \geq 0 \text{ if } f_j \text{ is non-linear}\}$, so that (3.2) constitutes a legitimate convex-concave saddle point problem. Beyond DRO, we consider the following examples, which appear frequently in statistical learning applications.

Example 3.1.1 (Generalized Linear Models (GLMs))**.** Consider a design matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, where each row $\boldsymbol{A}_{i\cdot}$ contains a d-dimensional feature vector. Fitting a generalized linear model (GLM) via the maximum likelihood principle results in a problem of the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^{n} \psi_i^*(\langle \boldsymbol{A}_{i\cdot}, \boldsymbol{x}\rangle) + \phi(\boldsymbol{x}),$$

where $\mathcal{X}$ is a parameter space, $\psi_i^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a convex loss function (e.g., the negative log-likelihood of the i-th output conditioned on $\boldsymbol{A}_{i\cdot}$), and ϕ is a regularizer such as the ℓ_2-norm squared or elastic net penalty. When taking the Fenchel conjugate of ψ_i^*

[1]While we fix the convention that the coupled term is linear in the dual variables, our methods extend by analogy to primal-linear objectives.

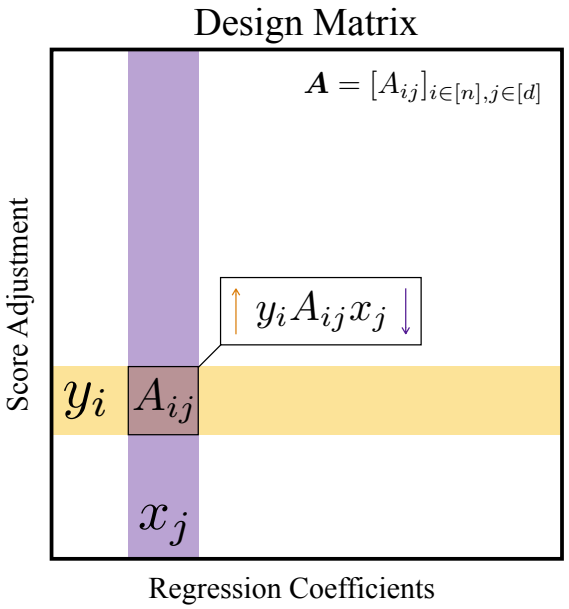

Figure 3.1: **Generalized Linear Models as Matrix Games.** Visualization to accompany Example 3.1.1, where $\boldsymbol{A}$ denotes the design matrix, $\boldsymbol{x}$ is parameter vector, and $\boldsymbol{y}$ multiplier to adjust the predicted scores $\boldsymbol{Ax}$.

(denoted ψ_i), we uncover the min-max problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathbb{R}^n} \langle \boldsymbol{y}, \boldsymbol{Ax}\rangle - \underbrace{\sum_{i=1}^{n} \psi_i(y_i)}_{\psi(\boldsymbol{y})} + \phi(\boldsymbol{x}), \tag{3.3}$$

which is an example of (3.2) with $f_i(\boldsymbol{x}) = \langle \boldsymbol{A}_{i\cdot}, \boldsymbol{x}\rangle$. This bilinearly coupled min-max problem, depicted in Figure 3.1, has interesting features beyond the classical setup. Notice that in terms of the dual variables, the objective of (3.3) is the sum of functions that depend only on each $y_i \in \mathbb{R}$ individually. This "separable" structure inspired recent methods for bilinearly coupled min-max problems with applications to statistical learning [Song et al., 2021], and is one that we pay specific attention to in Section 3.5.

Example 3.1.2 (Fully Composite Optimization)**.** The fully composite optimization problem (see Cui et al. [2018], Drusvyatskiy and Paquette [2019], Doikov and Nesterov [2022],

Vladarean et al. [2023] and references therein) writes as

$$\min_{\boldsymbol{x}\in\mathcal{X}} F(h(\boldsymbol{x}),\boldsymbol{x}),$$

where $h:\mathcal{X}\to\mathbb{R}^{n-d}$ (for $n>d$) is component-wise convex and $F:\mathbb{R}^n\times\mathcal{X}\to\mathbb{R}$ is closed and convex. It is assumed that h is smooth and "hard" to compute, whereas F may be non-differentiable but "easy" to compute. This formulation can be viewed as a generalization of the typical additive composite problem in which $n-d=1$ and $F(h(\boldsymbol{x}),\boldsymbol{x})=h(\boldsymbol{x})+g(\boldsymbol{x})$ for a smooth component h and non-smooth component g. By taking the Fenchel conjugate $F^*(\boldsymbol{y}):=\sup_{\boldsymbol{z}\in\mathbb{R}^n\times\mathcal{X}}\langle\boldsymbol{z},\boldsymbol{y}\rangle-F(\boldsymbol{z})$ of F, we achieve the formulation (3.2) with $f(\boldsymbol{x}):=(h(\boldsymbol{x}),\boldsymbol{x})\in\mathbb{R}^n$, $\psi(\boldsymbol{y})=F^*(\boldsymbol{y})$, and $\phi(\boldsymbol{x})=0$. To ensure that the overall problem is convex-concave we also assume that $F(\cdot,\boldsymbol{x})$ is monotone in that $\boldsymbol{u}\leq\boldsymbol{v}$ element-wise $\implies$ $F(\boldsymbol{u},\boldsymbol{x})\leq F(\boldsymbol{v},\boldsymbol{x})$.

Example 3.1.3 (Problems with Functional Constraints)**.** Consider the classical convex minimization problem with constraints defined by sublevel sets of convex functions

$$\min_{\boldsymbol{x}\in\mathcal{X}}\phi(\boldsymbol{x}) \text{ s.t. } f_j(\boldsymbol{x})\leq 0 \text{ for all } j=1,\dots,n.$$

Then, the Lagrangian formulation yields the expression (3.2) by letting $\boldsymbol{y}\in\mathcal{Y}=\mathbb{R}^n_+$ denote the Lagrange multipliers and setting $\psi\equiv 0$. The objective (3.2) also encompasses the related setting of "soft" functional constraints, where we set ψ as any ν-strongly convex function ψ with $\nu>0$ to produce a faster convergence rate at an approximation cost governed by the parameter ν. In this case, the primal solution resulting from this smoothed problem may only approximately satisfy the functional constraints. A classical example in statistical learning is the support vector machine problem [Cortes and Vapnik, 1995].

Example 3.1.4 (Maximal eigenvalue minimization)**.** Given a collection of d symmetric matrices $\mathbf{A}_1,\dots,\mathbf{A}_d\in\mathbb{R}^{m\times m}$, the classical problem of minimizing the maximal eigenvalue $\lambda_{\max}(\sum_{i=1}^d x_i\mathbf{A}_i)$ in $\boldsymbol{x}=(x_1,\dots,x_d)\in\mathbb{R}^d$ can be formulated [Nesterov, 2007b, Baes et al.,

2013] as the saddle-point problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \sum_{i=1}^{d} \operatorname{tr}\left((\mathbf{A}_i x_i)\boldsymbol{y}\right),$$

where $\mathcal{Y}$ is the set of positive semi-definite matrices satisfying $\operatorname{tr}(\boldsymbol{y}) = 1$ and $\mathcal{X}$ is any convex, compact subset of $\mathbb{R}^d$. Here, $n = m^2$ depends quadratically on the height/width of the matrices $(\mathbf{A}_i)_{i=1}^d$, and blocks may naturally correspond to matrix structure, such as rows or columns. This constitutes a non-separable matrix game, a specific problem class discussed in Section 3.6.

Even though some of the examples above move from statistical learning problems to more general optimization problems in applied mathematics, the field of statistics is not simply a source of example problems that fit our setting; instead, ideas such as using adaptive sampling and carefully handling complex dependence structures that arise will be essential for our improved complexity guarantees. In order to make statements of complexity, we also note that we will use a different convergence criterion in this chapter than in the previous one (another concrete facet of the "primal-dual" mindset). This *primal-dual gap* criterion will be amenable to an analysis that does not require a Lyapunov function and will be able to handle cases of non-strong convexity in a unified manner. Furthermore, the analysis of the primal-dual gap done in this chapter is of mathematical value; our proof techniques (outlined Section 3.3) are in line with prior work providing *constructive* arguments for the analysis of optimization methods [Diakonikolas and Orecchia, 2019, Mehta et al., 2024a, Li et al., 2024b, Diakonikolas, 2025]. In other words, the theoretical analysis provides guiding principles for deriving optimization algorithms that may otherwise be unsuspected. Applying this method, we present an algorithm that employs a unique combination of randomized updates and a "historical regularization" technique that is conceptually novel and interesting in its own right. This contrasts with the retrospective, empirically motivated approach of Chapter 2. By furnishing both perspectives, we argue that a deeper understanding of the problem of interest is developed. Let us outline the rest of the chapter.

In Section 3.2, we state the assumptions that fully specify the problem class and review the complexities of classical methods for nonbilinearly coupled saddle point problems. We also define *block separable* problems, whose structure we will exploit in later sections. Section 3.3 we introduce the high-level steps of the analysis and technical lemmas used throughout the proofs in various settings. In Section 3.4 and Section 3.5, we realize the template from the previous section for general and block separable objectives, respectively. Extensive comparisons are made to recent literature in Section 3.6. Extensions such as online convergence certificates and lower bounds are discussed in Section 3.7, and future work is commented on in Section 3.8.

3.2 Preliminaries

To initiate the discussion of computational complexity, we first make precise our (strong) convexity assumption, where we interpret values of the strong convexity modulus being zero ($\mu = 0$ or $\nu = 0$) as the corresponding function being simply convex. Let $\mathrm{ri}(\cdot)$ denote the relative interior of a set, and let $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ be norms on $\mathbb{R}^d$ and $\mathbb{R}^n$, respectively.

Assumption 3.2.1. Assume that ϕ is proper with $\mathcal{X} \subseteq \mathrm{dom}(\phi) := \left\{\boldsymbol{x} \in \mathbb{R}^d : \phi(\boldsymbol{x}) < +\infty\right\}$, closed (i.e., has a closed epigraph in $\mathbb{R}^{d+1}$), and μ-strongly convex ($\mu \geq 0$) with respect to $\|\cdot\|_{\mathcal{X}}$, that is, for any $\boldsymbol{s} \in \partial\phi(\boldsymbol{u})$, we have that $\phi(\boldsymbol{z}) \geq \phi(\boldsymbol{u}) + \langle \boldsymbol{s}, \boldsymbol{z} - \boldsymbol{u}\rangle + \frac{\mu}{2}\|\boldsymbol{z} - \boldsymbol{u}\|_{\mathcal{X}}^2$. Assume in addition that $\mathcal{X} \cap \mathrm{ri}(\mathrm{dom}(\phi))$ is non-empty. Similarly, assume that ψ is proper (with $\mathcal{Y} \subseteq \mathrm{dom}(\psi)$ and $\mathcal{Y} \cap \mathrm{ri}(\mathrm{dom}(\psi))$ non-empty), closed, and ν-strongly convex ($\nu \geq 0$) with respect to $\|\cdot\|_{\mathcal{Y}}$.

Recall that when $\mu > 0$, we aimed to control $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2$ in the results of Section 2.5.3, where the $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ was the unique optimum of the objective. That analog of this criterion for (3.2) would clearly be $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{X}}^2$ with $\boldsymbol{x}^\star \in \mathcal{X}$ defined analogously. Other than the possibly non-Euclidean geometry induced by $\|\cdot\|_{\mathcal{X}}$, note that $\boldsymbol{x}^\star$ may not exist or be unique for $\mu = 0$, which is also true of the similarly defined $\boldsymbol{y}^\star \in \mathcal{Y}$. Instead, as is standard for primal-dual algorithm of the same type [Chambolle and Pock, 2011, Alacaoglu et al., 2020,

Song et al., 2021], the optimality criterion will be the *primal-dual gap* of $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ at $\boldsymbol{u} \in \mathcal{X}$ and $\boldsymbol{v} \in \mathcal{Y}$:

$$\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y}) := \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) - \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}). \tag{3.4}$$

To relate this criterion to our previous one in the strongly convex case (when $\mu > 0$ and $\nu > 0$), we set $\boldsymbol{u} = \boldsymbol{x}_\star$ and $\boldsymbol{v} = \boldsymbol{y}_\star$, where $(\boldsymbol{x}_\star, \boldsymbol{y}_\star)$ also forms the unique saddle point of the objective (3.2), satisfying $\mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{y}) \leq \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{y}_\star) \leq \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_\star)$. In fact, by strong convexity, we have that

$$\mathrm{Gap}^{\boldsymbol{x}^\star,\boldsymbol{y}^\star}(\boldsymbol{x}, \boldsymbol{y}) - \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{X}}^2 - \frac{\nu}{2} \|\boldsymbol{y} - \boldsymbol{y}^\star\|_{\mathcal{Y}}^2 \geq 0,$$

which can be seen by adding the two non-negative terms $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star) - \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{X}}^2 \geq 0$ and $\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}) - \frac{\nu}{2} \|\boldsymbol{y} - \boldsymbol{y}^\star\|_{\mathcal{Y}}^2, \geq 0$. Thus, a quantitative convergence guarantee on (3.4) immediately yields such a guarantee on the distance-to-optimum terms $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{X}}^2$ and $\|\boldsymbol{y} - \boldsymbol{y}^\star\|_{\mathcal{Y}}^2$. Importantly, the primal-dual gap can be used when $\mu = 0$ or $\nu = 0$ by taking a supremum of (3.4) over $\boldsymbol{u} \in \mathcal{U}$ or $\boldsymbol{v} \in \mathcal{V}$, where $\mathcal{U} \subseteq \mathcal{X}$ and $\mathcal{V} \subseteq \mathcal{Y}$ are compact sets in which the iterates are contained.

Having described the convergence criterion formally, we may now motivate the usefulness of specialized algorithms for (3.2). We will compare various approaches in terms of the total arithmetic complexity or runtime needed to guarantee that (3.4) is smaller than a suboptimality parameter $\varepsilon > 0$. Consider treating (3.2) generically as a nonbilinearly coupled saddle-point problem (or more generally still as a variational inequality (VI) problem). Let λ be the Lipschitz parameter of $(\boldsymbol{x}, \boldsymbol{y}) \mapsto (\nabla_{\boldsymbol{x}} \mathcal{L}, -\nabla_{\boldsymbol{y}} \mathcal{L})$, when it exists. Classical approaches such as Korpelevich's extragradient [Korpelevich, 1976], Popov's method [Popov., 1980], Nemirovski's mirror-prox [Nemirovski, 2004], Nesterov's dual extrapolation [Nesterov, 2007a] will achieve the complexity $O(nd\lambda/\varepsilon)$. when $\mu = \nu = 0$, whereas their linearly convergent variants [Nesterov and Scrimali, 2006, Marcotte, 1991, Tseng, 1995, Mokhtari et al., 2020, Le Thi Thanh Hai and Vuong, 2025] achieve $O\left(nd(\lambda/\min\{\mu, \nu\}) \ln(1/\varepsilon)\right)$ in the strongly

convex-strongly concave setting.

Furthermore, using generic acceleration Catalyst schemes [Lin et al., 2018] adapted for min-max problems, improved complexity bounds can be obtained. Yang et al. [2020] achieve an $\tilde{O}(nd\lambda/\sqrt{\mu\varepsilon})$ runtime in the strongly convex-concave regime. Recently Lan and Li [2023] obtained a complexity of $\tilde{O}(nd\lambda/\sqrt{\mu\nu}\ln\left(\frac{1}{\varepsilon}\right))$ for strongly convex-strongly concave problems. A disadvantage of classical approaches that is overcome by the Catalyst approach is the dependence on the minimum of the two strong convexity constants. In the examples above, ψ often plays the role of a strongly convex smoothing penalty whose strong convexity constant may be near-zero (e.g., $O(\varepsilon)$ for some applications). Thus, guarantees depending on $(\mu \wedge \nu)$ are undesirable. A second disadvantage incurred by all of the approaches above is the dependence on a coarse Lipschitz constant λ for the entire vector field. When it is only known that each component function f_j is $G_{\max}$-Lipschitz continuous and $M_{\max}$-smooth, and that $|y_j| \leq D_{\max}$ for any $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathcal{Y}$, then λ can be estimated as $n(G_{\max} + D_{\max} M_{\max})$. Combined with the order-nd per-iteration cost, the total arithmetic complexity can have an $O(n^2 d)$ dependence on the dimensions (n, d). These two issues highlight the price to pay when using generic acceleration schemes: the loss of adaptation to the structure of the problem and the related constants. In contrast, by leveraging non-uniformity in the various Lipschitz/smoothness constants for each f_j and randomized updates, we may achieve complexities that are linear in $(n + d)$.

In contrast, the (possibly non-uniform) Lipschitz continuity and smoothness parameters of the component functions represent quantities of practical interest in applications. In Example 3.1.1, the Lipschitz continuity properties of each f_j will depend on the norm of the corresponding row $\boldsymbol{A}_{j\cdot}$. In Example 3.1.2, we have that $f(\boldsymbol{x}) = (h(\boldsymbol{x}), \boldsymbol{x})$ so that $f_{n-d+1}, \ldots, f_n$ are 1-Lipschitz continuous and 0-smooth; a complexity result that only depends on the n times the maxima of these constants over the components of f could be overly pessimistic.

In our analysis, we handle every combination of $\mu = 0$ versus $\mu > 0$ and $\nu = 0$ versus $\nu > 0$ to achieve the same dependence on ε in a unified way with a constant that enjoys a

transparent dependence on these component-wise Lipschitz and smoothness constants. The complexity benefits of using component-wise problem constants have been shown in recent works on stochastic variance-reduced methods for variational inequality problems [Alacaoglu and Malitsky, 2022, Cai et al., 2024, Pichugin et al., 2024, Alizadeh et al., 2024, Diakonikolas, 2025].

We now introduce the mathematical objects and assumptions used in this chapter. As before, we let $\|\cdot\|_{\mathcal{X}}$ denote a norm on $\mathbb{R}^d$. Let $\|\cdot\|_{\mathcal{Y}}$ denote an ℓ_p-norm on $\mathbb{R}^n$ for $p \in [1, 2]$. The associated dual norms are denoted by $\|\cdot\|_{\mathcal{X}^*}$ and $\|\cdot\|_{\mathcal{Y}^*}$ and defined in the usual way as $\|\boldsymbol{w}\|_{\mathcal{X}^*} = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_{\mathcal{X}}\leq 1} \langle \boldsymbol{w}, \boldsymbol{x}\rangle$, $\|\boldsymbol{z}\|_{\mathcal{Y}^*} = \sup_{\boldsymbol{y}:\|\boldsymbol{y}\|_{\mathcal{Y}}\leq 1} \langle \boldsymbol{z}, \boldsymbol{y}\rangle$. The following assumptions about the objective in (3.2) are made throughout the chapter. We employ block coordinate-wise updates in the upcoming stochastic algorithms, of which coordinate-wise updates are a special case with block size one. To do so, we introduce a partitioning of the n components of f into N blocks and define the relevant Lipschitz constants for each one. In discussions of arithmetic complexity, we may assume a uniform block size n/N, but the analysis natively handles blocks of possibly non-uniform size.

Assumption 3.2.2. Assume that each $f_j : \mathbb{R}^d \to \mathbb{R}$ is convex and the indices $[n] = \{1, \ldots, n\}$ are partitioned into *blocks* $(B_1, \ldots, B_N)$. There exist constants $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_N \geq 0$ and $\boldsymbol{L}_1, \ldots \boldsymbol{L}_N \geq 0$ such that for each $J = 1, \ldots, N$:

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\textstyle\sum_{j\in B_J} z_j \nabla f_j(\boldsymbol{x})\right\|_{\mathcal{X}^*} \leq \boldsymbol{G}_J \|\boldsymbol{z}\|_2, \qquad \forall \boldsymbol{z} \in \mathbb{R}^{|B_J|}, \tag{3.5}$$

$$\sup_{\boldsymbol{y}\in\mathcal{Y}} \left\|\textstyle\sum_{j\in B_J} y_j(\nabla f_j(\boldsymbol{x}) - \nabla f_j(\boldsymbol{x}'))\right\|_{\mathcal{X}^*} \leq \boldsymbol{L}_J \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}. \tag{3.6}$$

In addition, for $\nabla f(\boldsymbol{x}) := (\nabla f_1(\boldsymbol{x}), \ldots, \nabla f_n(\boldsymbol{x})) \in \mathbb{R}^{n\times d}$, there exist $G \geq 0$ and $L \geq 0$ such that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\nabla f(\boldsymbol{x})^\top \boldsymbol{z}\right\|_{\mathcal{X}^*} \leq G \|\boldsymbol{z}\|_{\mathcal{Y}}, \qquad \forall \boldsymbol{z} \in \mathbb{R}^n, \tag{3.7}$$

$$\sup_{\boldsymbol{y}\in\mathcal{Y}} \left\|(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}'))^\top \boldsymbol{y}\right\|_{\mathcal{X}^*} \leq L \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}. \tag{3.8}$$

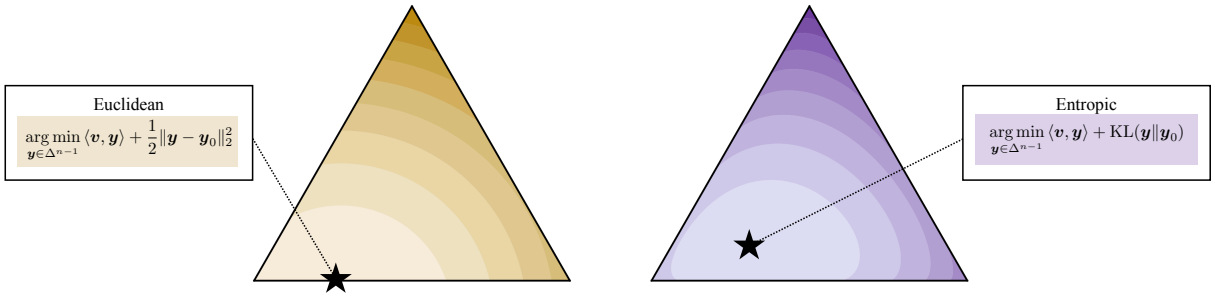

Figure 3.2: **Moreau Envelope of Bregman Divergences on the Unit Simplex.** Visualization of the objective of the proximal operator in the case of the standard ℓ_2-norm-squared Bregman divergence (left) and the negative entropy-base Bregman divergence (right).

Next, we address the possible non-Euclidean structure of $\mathcal{X}$ and $\mathcal{Y}$. A classic example of such a feasible set is the probability simplex $\Delta^{n-1} = \left\{\boldsymbol{y} \in \mathbb{R}^n_{\geq 0} : \langle \boldsymbol{y}, \mathbf{1}\rangle = 1\right\}$, visualized in Figure 3.2. Notice that using the ℓ_2-geometry results in a constrained optimization problem (Euclidean projection on the simplex), whereas the entropic proximal step objective has a unique minimizer contained within the set. Arguing analytically, consider the case in which $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_2$ and we wish to upper bound G from (3.7) in terms of the individual Lipschitz constants of the component functions $f_1, \ldots, f_n$. In the case of $\|\cdot\|_{\mathcal{Y}} = \|\cdot\|_2$ we have that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\nabla f(\boldsymbol{x})^\top \boldsymbol{y}\right\|_{\mathcal{X}^*} \leq \sup_{\boldsymbol{x}\in\mathcal{X}} \underbrace{\sqrt{\textstyle\sum_{j=1}^n \|\nabla f_j(\boldsymbol{x})\|_2^2}}_{\|\nabla f_j(\boldsymbol{x})\|_{\mathrm{Fro}}} \|\boldsymbol{y}\|_2$$

whereas for $\|\cdot\|_{\mathcal{Y}} = \|\cdot\|_1$, we have

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left\|\nabla f(\boldsymbol{x})^\top \boldsymbol{y}\right\|_{\mathcal{X}^*} \leq \sup_{\boldsymbol{x}\in\mathcal{X}} \underbrace{\max_{j\in[n]} \|\nabla f(\boldsymbol{x})\|_2}_{\left\|\nabla f(\boldsymbol{x})^\top\right\|_{1,2}} \|\boldsymbol{y}\|_1 .$$

For the second inequality, the resulting bound is up to $\sqrt{n}$ smaller than for the first. To take advantage of non-Euclidean norms, we rely upon the following definition of Bregman divergences for possibly non-differentiable distance-generating functions.

Consider a convex subset $\mathcal{Z} \subseteq \mathbb{R}^m$, and let φ be a proper, closed, and 1-strongly convex

function satisfying $\mathcal{Z} \subseteq \mathrm{dom}(\varphi)$. Define the *Bregman divergence* $\Delta_\varphi : \mathrm{dom}(\varphi) \times \mathrm{ri}(\mathrm{dom}(\varphi)) \to \mathbb{R}$ as

$$\Delta_\varphi(\boldsymbol{z}, \boldsymbol{z}') := \varphi(\boldsymbol{z}) - \varphi(\boldsymbol{z}') - \langle \nabla\varphi(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \rangle .$$

The notation $\nabla\varphi(\boldsymbol{z}') \in \partial\varphi(\boldsymbol{z}')$ denotes an arbitrary, but consistently chosen subgradient at $\boldsymbol{z}'$ when applied to a convex but possibly non-differentiable function. This slight modification is made for purely technical reasons, as we perform a mirror descent-style analysis with Bregman divergences generated by the (possibly non-smooth) component functions ϕ and ψ. Lemma 3.2.1 provides a modified three-point inequality used in both the upper and lower bounds on the objective used in bounding the initial gap estimates.

Lemma 3.2.1. *Let h, g, and φ be proper, closed, and convex functions whose domains contain $\mathcal{Z}$ and that map to $\mathbb{R} \cup \{+\infty\}$. Assume that g is relatively γ-strongly convex with respect to φ on $\mathcal{Z}$, i.e., $\Delta_g(\boldsymbol{u}, \boldsymbol{z}) \geq \gamma\Delta_\varphi(\boldsymbol{u}, \boldsymbol{z})$ for $\gamma \geq 0$ and $\boldsymbol{u}, \boldsymbol{z} \in \mathcal{Z}$. Let $A \geq 0$, $a > 0$, $\gamma_0 > 0$ be constants, and let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r \in \mathrm{ri}(\mathrm{dom}(\varphi))$. Let*

$$\boldsymbol{z}^+ = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathcal{Z}} \left\{ m(\boldsymbol{u}) := h(\boldsymbol{u}) + ag(\boldsymbol{u}) + \tfrac{A\gamma+\gamma_0}{2} \textstyle\sum_{i=1}^r w_i \Delta_\varphi(\boldsymbol{u}, \boldsymbol{z}_i) \right\}, \tag{3.9}$$

where each $w_i \geq 0$ and $\sum_{i=1}^r w_i = 1$. Then, for any $\boldsymbol{u} \in \mathcal{Z}$,

$$m(\boldsymbol{u}) \geq m(\boldsymbol{z}^+) + \left(\tfrac{(A+a)\gamma+\gamma_0}{2} \right) \Delta_\varphi(\boldsymbol{u}, \boldsymbol{z}^+) + \tfrac{a\gamma}{2} \Delta_\varphi(\boldsymbol{u}, \boldsymbol{z}^+).$$

Proof. By the definition of the Bregman divergence generated by m, we have that

$$\begin{aligned} m(\boldsymbol{u}) &= m(\boldsymbol{z}^+) + \left\langle \nabla m(\boldsymbol{z}^+), \boldsymbol{u} - \boldsymbol{z}^+ \right\rangle + \Delta_m(\boldsymbol{u}, \boldsymbol{z}^+), \\ &\geq m(\boldsymbol{z}^+) + \Delta_m(\boldsymbol{u}, \boldsymbol{z}^+), \end{aligned}$$

where we use that $\langle \nabla m(\boldsymbol{z}^+), \boldsymbol{u} - \boldsymbol{z}^+ \rangle \geq 0$ for any subgradient $\nabla m(\boldsymbol{z}^+)$ as $\boldsymbol{z}^+ \in \arg\min_{\boldsymbol{u} \in \mathcal{Z}} m(\boldsymbol{u})$. Then, by using the definition of m, and that $\Delta_{\Delta_\varphi(\cdot, \boldsymbol{z})} = \Delta_\varphi$ for any fixed $\boldsymbol{z} \in \mathrm{ri}(\mathrm{dom}(\varphi))$, we

have that

$$\begin{aligned} m(\boldsymbol{u}) &\geq m(\boldsymbol{z}^+) + \Delta_m(\boldsymbol{u}, \boldsymbol{z}^+) \\ &= m(\boldsymbol{z}^+) + \Delta_h(\boldsymbol{u}, \boldsymbol{z}^+) + a\Delta_g(\boldsymbol{u}, \boldsymbol{z}^+) + \frac{A\gamma + \gamma_0}{2} \sum_{i=1}^{r} w_i \Delta_{\Delta_\varphi(\cdot, \boldsymbol{z}_i)}(\boldsymbol{u}, \boldsymbol{z}^+) \\ &= m(\boldsymbol{z}^+) + \Delta_h(\boldsymbol{u}, \boldsymbol{z}^+) + a\Delta_g(\boldsymbol{u}, \boldsymbol{z}^+) + \frac{A\gamma + \gamma_0}{2} \Delta_\varphi(\boldsymbol{u}, \boldsymbol{z}^+). \end{aligned}$$

Use $\Delta_h(\boldsymbol{u}, \boldsymbol{z}^+) \geq 0$ and then relative strong convexity $a\Delta_g(\boldsymbol{u}, \boldsymbol{z}^+) \geq a\gamma\Delta_\varphi(\boldsymbol{u}, \boldsymbol{z}^+)$ to prove the desired result. □

Henceforth, we use the notation $\Delta_\mathcal{X}(\cdot, \cdot)$ to denote the Bregman divergence on $\mathcal{X}$ that is both 1-strongly convex with respect to $\|\cdot\|_\mathcal{X}$ and that satisfies $\Delta_\phi \geq \mu\Delta_\mathcal{X}$ (i.e., ϕ is relatively μ-strongly convex with respect to $\Delta_\mathcal{X}$). As an example, this is satisfied by $\Delta_\mathcal{X} := \Delta_{\phi/\mu}$ for $\mu > 0$. We define $\Delta_\mathcal{Y}(\cdot, \cdot)$ analogously. As a technical consideration, we also assume that the unique solution of (3.9) lies in $\mathrm{ri}(\mathrm{dom}(\varphi))$, which is satisfied for common choices of the generator φ. Finally, we introduce additional structure on the problem (3.2) which can be exploited when available.

Definition 3.2.1. We call $\mathcal{L}$ a *dual-separable objective* if the dual component decomposes as $\psi(\boldsymbol{y}) = \sum_{J=1}^{N} \psi_J(\boldsymbol{y}_J)$ where $\boldsymbol{y}_J$ denotes the components of $\boldsymbol{y} \in \mathcal{Y}$ corresponding to the indices in block B_J. We call $\mathcal{X} \times \mathcal{Y}$ a *dual-separable feasible set* if $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_N$ and $\boldsymbol{y}_J \in \mathcal{Y}_J$ for $J = 1, \ldots, N$. We call the problem (3.2) a *dual-separable problem* if its objective and feasible set are both dual-separable.

Dual-separability of the objective is commonly satisfied, such as when ψ represents an ℓ_2 or negative entropy penalty. In this case, we have that $\Delta_\mathcal{Y}(\boldsymbol{y}, \boldsymbol{y}') := \sum_{J=1}^{N} \Delta_J(\boldsymbol{y}_J, \boldsymbol{y}'_J)$, where each $\Delta_J(\cdot, \cdot)$ is a Bregman divergence on $\mathrm{dom}(\psi_J) \times \mathrm{ri}(\mathrm{dom}(\psi_J))$. As before, ψ is relatively ν-strongly convex with respect to $\Delta_\mathcal{Y}$. In later sections, we may also use the subscript on $\boldsymbol{y}_k \in \mathcal{Y}$ to denote a particular time index of an algorithm, as opposed to the block index J on $\boldsymbol{y}_J \in \mathcal{Y}_J$; the difference will be clear from context. Dual-separability of the feasible set is a less common assumption. It is not satisfied, for instance, on simplicial domains such as

the one in Example 1. Based on this observation, Section 3.4 and Section 3.5 are dedicated to the proposed algorithms for non-separable and separable problems, respectively.

3.3 Method and General Analysis Template

We consider an algorithm to be a sequence of primal-dual iterates $(\boldsymbol{x}_k, \boldsymbol{y}_k)_{k\geq 0}$, with fixed initial point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. This sequence may be random, in which case the relevant probabilistic information is introduced when we analyze stochastic algorithms.

Recalling the gap function (3.4), we first fix $\boldsymbol{u}, \boldsymbol{v}$ and aim to show that $\limsup_{t\to\infty} \mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_t, \boldsymbol{y}_t) \leq 0$ (possibly in expectation), with a convergence rate in terms of the problem constants from Section 3.2 and iteration count $t \geq 0$. To this end, we introduce an averaging sequence $(a_k)_{k\geq 1}$ of positive constants with $a_0 = 0$, and their aggregation $A_t := \sum_{k=0}^{t} a_k$, and then pursue an upper bound of the form

$$\sum_{k=1}^{t} a_k \,\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k) \leq G_0(\boldsymbol{u}, \boldsymbol{v}), \tag{3.10}$$

where $G_0(\boldsymbol{u}, \boldsymbol{v})$ is a constant independent of t, so that when dividing by A_t, the average expected gap decays at rate A_t^{-1}. Accordingly, we wish for A_t to grow as fast as possible with t. By way of convexity, we have that $\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \leq G_0(\boldsymbol{u}, \boldsymbol{v})/A_t$ for $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) := A_t^{-1} \sum_{k=1}^{t} a_k(\boldsymbol{x}_k, \boldsymbol{y}_k)$, which can be returned by the algorithm to realize the gap bound (3.10). We may conclude by taking the supremum of $G_0(\boldsymbol{u}, \boldsymbol{v})$. In the randomized scenarios, we also discuss in Section 3.6 how the analysis can be adapted to an even stronger convergence criterion for which the supremum is taken prior to the expectation. Because many of the technical ideas remain similar when proving convergence with respect to this stronger criterion, we describe only the parts that change in Section 3.6.

For any algorithm we consider, the analysis will proceed by constructing an upper bound on $a_k \,\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k)$ containing telescoping and non-positive terms, by first lower bounding $a_k\mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k)$ and upper bounding $a_k\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{v})$. As we will see, the update for $\boldsymbol{x}_k$ will be used to produce the lower bound while the update for $\boldsymbol{y}_k$ will be used to produce the upper bound.

Algorithm 4 Template Method

Input: Initial point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, averaging sequence $(a_k)_{k=0}^t$, non-negative weights $(\gamma_I)_{I=1}^N$ that sum to one, balancing sequences $(w_k^{\mathrm{P}})_{k=1}^t$ and $(w_k^{\mathrm{D}})_{k=1}^t$, functions SUBROUTINE$_1$, SUBROUTINE$_2$, and SUBROUTINE$_3$.

Initialize the comparison points $\hat{\boldsymbol{x}}_{0,I} = \boldsymbol{x}_0$ for all $I \in [N]$ and $\hat{\boldsymbol{y}}_0 = \boldsymbol{y}_0$.

for $k = 1$ **to** t **do**

SUBROUTINE$_1$: Compute $\bar{\boldsymbol{g}}_{k-1}$ using stored information and oracle calls to $\nabla f_i(\boldsymbol{x}_{k-1})$, $i \in [n]$

Perform the primal update

$$\boldsymbol{x}_k = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \Big\{ a_k \left\langle \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{x} \right\rangle + a_k \phi(\boldsymbol{x}) + (A_{k-1}\mu + \mu_0)\Big(\underbrace{\tfrac{1-w_k^{\mathrm{P}}}{2} \Delta_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}_{k-1})}_{\text{standard proximity term}} + \underbrace{\tfrac{w_k^{\mathrm{P}}}{2} \textstyle\sum_{I=1}^N \gamma_I \Delta_{\mathcal{X}}(\boldsymbol{x}, \hat{\boldsymbol{x}}_{k-1,I})}_{\text{primal historical regularization}} \Big) \Big\}. \tag{3.11}$$

SUBROUTINE$_2$: Compute $\bar{\boldsymbol{f}}_{k-1/2}$ using stored information and some calls to $f_1(\boldsymbol{x}_k), \ldots, f_n(\boldsymbol{x}_k)$.

Perform the dual update

$$\boldsymbol{y}_k = \operatorname*{arg\,max}_{\boldsymbol{y}\in\mathcal{Y}} \Big\{ a_k \left\langle \boldsymbol{y}, \bar{\boldsymbol{f}}_{k-1/2} \right\rangle - a_k \psi(\boldsymbol{y}) - (A_{k-1}\nu + \nu_0)\Big(\underbrace{\tfrac{1-w_k^{\mathrm{D}}}{2} \Delta_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}_{k-1})}_{\text{standard proximity term}} + \underbrace{\tfrac{w_k^{\mathrm{D}}}{2} \Delta_{\mathcal{Y}}(\boldsymbol{y}, \hat{\boldsymbol{y}}_{k-1})}_{\text{dual historical regularization}} \Big) \Big\}. \tag{3.12}$$

SUBROUTINE$_3$: Update comparison points $(\hat{\boldsymbol{x}}_{k,I})_{I=1}^N$ and $\hat{\boldsymbol{y}}_k$.

return $A_t^{-1} \sum_{k=1}^t a_k(\boldsymbol{x}_k, \boldsymbol{y}_k)$.

The update rules are motivated directly by the analysis. Similar steps will take place in the stochastic setting, except using the expectation of (3.4) under algorithmic randomness. We start with an arbitrary point $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathrm{ri}(\mathrm{dom}(\phi)) \times \mathrm{ri}(\mathrm{dom}(\psi))$. The parameters $\mu_0 > 0$ and $\nu_0 > 0$ appearing below are employed in order to handle the strongly convex and non-strongly convex settings in a unified manner.

This general, high-level idea and analysis template are in line with prior work providing constructive arguments for the analysis of optimization methods [Diakonikolas and Orecchia,

2019, Mehta et al., 2024a, Li et al., 2024b, Diakonikolas, 2025]. As such, it provides a clear guiding principle for the analysis and motivation for the algorithmic choices. It is of note, however, that while the general principle is common to all these works, the specifics of the analysis and associated algorithms differ significantly, as the technical obstacles they need to address are problem-specific. For instance, the stochastic algorithms in the present work have a unique combination of randomized updates and historical regularization that are conceptually novel and interesting in their own right. It is the combination of *both* non-uniform sampling and non-uniform regularization that leads to complexity improvements even in more specific problem classes such as bilinearly coupled problems. For the separable case, our proof technique relies on an auxiliary sequence of dual variables that offers an elegant extension of previous meta-analyses when using coordinate-wise updates.

3.3.1 Algorithm Template and a First Gap Bound

We fix the convention that the update for $\boldsymbol{x}_k$ occurs before the update for $\boldsymbol{y}_k$. Thus, we require that only information available up to and including time $k-1$ is used in the update. In the stochastic setting, this requirement will be formalized in the language of measurability. Both the primal and dual updates will resemble those of a proximal gradient-type algorithm, wherein $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are defined by minimizing or maximizing an approximation of (3.2) with a proximity term (i.e., a Bregman divergence). In the primal update, the proximity term promotes $\boldsymbol{x}_k$ being close to not only $\boldsymbol{x}_{k-1}$, but several additional to-be-specified comparison points $\hat{\boldsymbol{x}}_{k-1,1}, \ldots, \hat{\boldsymbol{x}}_{k-1,N}$. Similarly, $\boldsymbol{y}_k$ will be made close to $\boldsymbol{y}_{k-1}$ along with a single comparison point $\hat{\boldsymbol{y}}_{k-1}$. The use of multiple comparison points is in fact the motivation for the modified three-point inequality (Lemma 3.2.1). This "historical regularization" is visualized in Figure 3.3. The remaining components to specify are $\bar{\boldsymbol{g}}_{k-1} \in \mathbb{R}^d$, a vector which will be used to linearize the objective (3.2) in the primal update, and $\bar{\boldsymbol{f}}_{k-1/2} \in \mathbb{R}^n$, an analogous vector used in the dual update. The subscript $k-1/2$ indicates that "half" of the information in iteration k (namely, the value of $\boldsymbol{x}_k$) can be used in the update. The components introduced so far generate a template algorithm which can combine them in various ways;

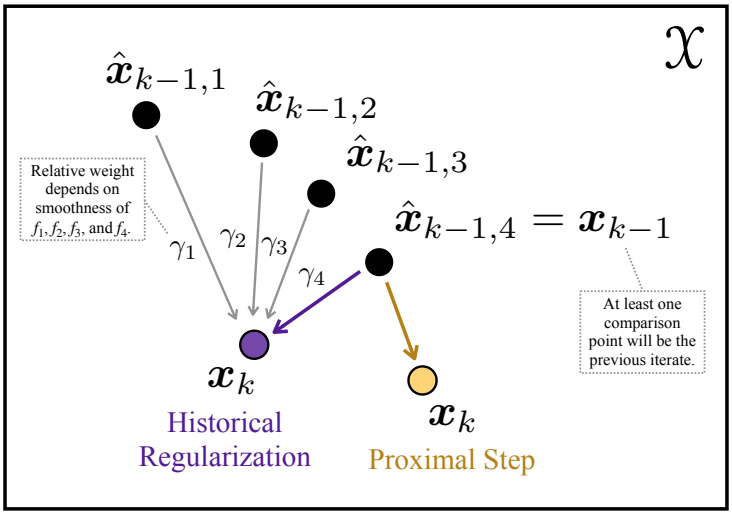

Figure 3.3: **Modified Proximal Step with Historical Regularization.** Geometric illustration of the historical regularization penalty applied in the modified primal update (3.11).

the pseudocode for this method is shown in Algorithm 4, in which the algorithm-specific content is abstracted into three subroutines.

The hyperparameters are shown explicitly to better accompany the theoretical analysis. They are set to specific values over the course of the proofs. We motivate the updates (3.11) and (3.12) using a lower bound on $a_k\mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k)$ and an upper bound on $a_k\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{v})$. Several terms will appear that either telescope when summed or are used to cancel errors incurred at each iteration. For the reader's convenience, we summarize this notation below.

Notation: Throughout the analyses for each algorithm, the primal and dual "distance-to-opt" terms will be written as

$$\mathcal{T}_k^{\mathrm{P}} := (A_k\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_k), \quad \mathcal{T}_k^{\mathrm{D}} := (A_k\nu + \nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{u}, \boldsymbol{y}_k), \tag{3.13}$$

indicating that they telescope when summed. The bounds also produce the negation of the terms

$$\mathcal{C}_k^{\mathrm{P}} := (A_{k-1}\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}_k, \boldsymbol{x}_{k-1}), \quad \mathcal{C}_k^{\mathrm{D}} := (A_{k-1}\nu + \nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{y}_k, \boldsymbol{y}_{k-1}), \tag{3.14}$$

which appear in the gap function bound and are used for canceling errors that appear when controlling the primal-dual gap. Analogous terms appear based on the comparison points instead of the iterates. That is, consider for each $I \in [N]$ the additional telescoping terms

$$\hat{\mathcal{T}}_{k,I}^{\mathrm{P}} := (A_k\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{u}, \hat{\boldsymbol{x}}_{k,I}), \quad \hat{\mathcal{T}}_k^{\mathrm{D}} := (A_k\nu + \nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{v}, \hat{\boldsymbol{y}}_k), \tag{3.15}$$

along with similar cancellation terms

$$\hat{\mathcal{C}}_{k,I}^{\mathrm{P}} := (A_{k-1}\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}_k, \hat{\boldsymbol{x}}_{k-1,I}), \quad \hat{\mathcal{C}}_k^{\mathrm{D}} := (A_{k-1}\nu + \nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{y}_k, \hat{\boldsymbol{y}}_{k-1}). \tag{3.16}$$

Finally, we will also define the inner product terms

$$\mathcal{I}_k^{\mathrm{P}} = a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle, \quad \mathcal{I}_k^{\mathrm{D}} = a_k \left\langle f(\boldsymbol{x}_k) - \bar{\boldsymbol{f}}_{k-1/2}, \boldsymbol{v} - \boldsymbol{y}_k \right\rangle, \tag{3.17}$$

which will comprise the errors that are cancelled by the terms above.

To achieve the lower bound, we compare the objective to the one minimized by $\boldsymbol{x}_k$ and use properties of Bregman divergences to produce telescoping terms akin to a mirror descent-style analysis.

Lemma 3.3.1. *For any $k \geq 1$, let $\bar{\boldsymbol{g}}_{k-1} \in \mathbb{R}^d$ and $w_k^{\mathrm{P}}, w_{k-1}^{\mathrm{P}} \in [0, 1)$ such that $w_{k-1}^{\mathrm{P}} \leq w_k^{\mathrm{P}}$.*

For $k \geq 1$, if $\boldsymbol{x}_k$ is defined via the update (3.11), then it holds that

$$a_k \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k) \geq a_k \mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}_k) + \mathcal{I}_k^{\mathrm{P}} \tag{3.18}$$

$$+ \left(\tfrac{1-w_k^{\mathrm{P}}}{2} \mathcal{T}_k^{\mathrm{P}} - \tfrac{1-w_{k-1}^{\mathrm{P}}}{2} \mathcal{T}_{k-1}^{\mathrm{P}} \right) + \tfrac{w_k^{\mathrm{P}}}{2} \left(\mathcal{T}_k^{\mathrm{P}} - \textstyle\sum_{I=1}^N \gamma_I \hat{\mathcal{T}}_{k-1,I}^{\mathrm{P}} \right)$$

$$+ \tfrac{1-w_k^{\mathrm{P}}}{2} \mathcal{C}_k^{\mathrm{P}} + \tfrac{w_k^{\mathrm{P}}}{2} \textstyle\sum_{I=1}^N \gamma_I \hat{\mathcal{C}}_{k,I}^{\mathrm{P}} + \tfrac{a_k \mu}{2} \Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_k). \tag{3.19}$$

Proof. Because $\boldsymbol{y}_k$ is observed in $\mathcal{Y}$, we have that $\boldsymbol{x} \mapsto \langle \boldsymbol{y}_k, f(\boldsymbol{x}) \rangle$ is convex and differentiable. As a result,

$$\begin{aligned} a_k \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k) &= a_k (\boldsymbol{y}_k^\top f(\boldsymbol{u}) - \psi(\boldsymbol{y}_k) + \phi(\boldsymbol{u})) \\ &\geq a_k \left\langle \boldsymbol{y}_k, f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)(\boldsymbol{u} - \boldsymbol{x}_k) \right\rangle - a_k \psi(\boldsymbol{y}_k) + a_k \phi(\boldsymbol{u}). \end{aligned} \tag{3.20}$$

Then, add and subtract terms from the objective defining (3.11) and apply Lemma 3.2.1 with $A = A_{k-1}$, $a = a_k$, $\gamma = \mu$, and $\gamma_0 = \mu_0$ to achieve

$$\begin{aligned} a_k \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k) \geq\ & a_k \left\langle \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle + a_k \phi(\boldsymbol{u}) + a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\ & + \tfrac{1-w_k^{\mathrm{P}}}{2} (A_{k-1}\mu + \mu_0) \Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_{k-1}) - \underbrace{\tfrac{1-w_k^{\mathrm{P}}}{2} (A_{k-1}\mu + \mu_0) \Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_{k-1})}_{\mathcal{T}_{k-1}^{\mathrm{P}} \text{ from (3.13)}} \\ & + \tfrac{w_k^{\mathrm{P}}}{2} (A_{k-1}\mu + \mu_0) \sum_{I=1}^N \gamma_I \Delta_{\mathfrak{X}}(\boldsymbol{u}, \hat{\boldsymbol{x}}_{k-1,I}) - \underbrace{\tfrac{w_k^{\mathrm{P}}}{2} (A_{k-1}\mu + \mu_0) \sum_{I=1}^N \gamma_I \Delta_{\mathfrak{X}}(\boldsymbol{u}, \hat{\boldsymbol{x}}_{k-1,I})}_{\sum_I \gamma_I \hat{\mathcal{T}}_{k-1,I}^{\mathrm{P}} \text{ from (3.15)}} \\ & + a_k \left\langle \boldsymbol{y}_k, f(\boldsymbol{x}_k) \right\rangle - a_k \psi(\boldsymbol{y}_k). \end{aligned}$$

When applying Lemma 3.2.1 and using the definitions of $\mathcal{C}_k^{\mathrm{P}}$ from (3.14) and $\hat{\mathcal{C}}_{k,I}^{\mathrm{P}}$ from (3.16),

we achieve the inequality

$$\begin{aligned} a_k \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k) \geq a_k \phi(\boldsymbol{x}_k) + a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\ + \tfrac{1-w_k^{\mathrm{P}}}{2} \mathcal{C}_k^{\mathrm{P}} - \tfrac{1-w_k^{\mathrm{P}}}{2} \mathcal{T}_{k-1}^{\mathrm{P}} \\ + \tfrac{w_k^{\mathrm{P}}}{2} \sum_{I=1}^{N} \gamma_I \hat{\mathcal{C}}_{k,I}^{\mathrm{P}} - \tfrac{w_k^{\mathrm{P}}}{2} \sum_{I} \gamma_I \hat{\mathcal{T}}_{k-1,I}^{\mathrm{P}} \\ + a_k \left\langle \boldsymbol{y}_k, f(\boldsymbol{x}_k) \right\rangle - a_k \psi(\boldsymbol{y}_k). \end{aligned}$$

Finally, use $-\frac{1-w_k^{\mathrm{P}}}{2} \mathcal{T}_{k-1}^{\mathrm{P}} \geq -\frac{1-w_{k-1}^{\mathrm{P}}}{2} \mathcal{T}_{k-1}^{\mathrm{P}}$, the substitution $\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}_k) = \langle \boldsymbol{y}_k, f(\boldsymbol{x}_k) \rangle - \psi(\boldsymbol{y}_k) + \phi(\boldsymbol{x}_k)$, and the definition of $\mathcal{I}_k^{\mathrm{P}}$ from (3.17) to complete the proof. □

The upper bound is proved in a nearly identical fashion, except the step that employs convexity in (3.20) is not used; the proof is omitted for brevity.

Lemma 3.3.2. *For any $k \geq 1$, let $\bar{\boldsymbol{f}}_{k-1/2} \in \mathbb{R}^n$ and $w_k^{\mathrm{D}}, w_{k-1}^{\mathrm{D}} \in (0,1)$ such that $w_{k-1}^{\mathrm{D}} \leq w_k^{\mathrm{D}}$. For $k \geq 1$, if $\boldsymbol{y}_k$ is defined via the update (3.12), then it holds that*

$$a_k \mathbb{E}[\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{v})] \leq a_k \mathcal{L}(\boldsymbol{x}_k, \boldsymbol{y}_k) + \mathcal{I}_k^{\mathrm{D}} \tag{3.21}$$

$$+ \left(\tfrac{1-w_{k-1}^{\mathrm{D}}}{2} \mathcal{T}_{k-1}^{\mathrm{D}} - \tfrac{1-w_k^{\mathrm{D}}}{2} \mathcal{T}_k^{\mathrm{D}} \right) + \tfrac{w_k^{\mathrm{D}}}{2} \left(\hat{\mathcal{T}}_{k-1}^{\mathrm{D}} - \mathcal{T}_k^{\mathrm{D}} \right) \tag{3.22}$$

$$- \tfrac{1-w_k^{\mathrm{D}}}{2} \mathcal{C}_k^{\mathrm{D}} - \tfrac{w_k^{\mathrm{D}}}{2} \hat{\mathcal{C}}_k^{\mathrm{D}} - \tfrac{a_k \nu}{2} \Delta_{\mathrm{y}}(\boldsymbol{v}, \boldsymbol{y}_k). \tag{3.23}$$

Combining the derived upper and lower bounds and canceling matching terms, we claim the following result. This bound will be the starting point for every subsequent result in the chapter, so we reference it heavily in the sequel.

Claim 3.3.1. *Using the notation from (3.13), (3.14), (3.15), (3.16), and (3.17), we have*

$$\sum_{k=1}^{t} a_k \operatorname{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k) \leq \left(\tfrac{1-w_0^{\mathrm{P}}}{2}\mathcal{T}_0^{\mathrm{P}} - \tfrac{1-w_t^{\mathrm{P}}}{2}\mathcal{T}_t^{\mathrm{P}}\right) + \left(\tfrac{1-w_0^{\mathrm{D}}}{2}\mathcal{T}_0^{\mathrm{D}} - \tfrac{1-w_t^{\mathrm{D}}}{2}\mathcal{T}_t^{\mathrm{D}}\right)$$

$$+ \sum_{k=1}^{t} \left[\mathcal{I}_k^{\mathrm{D}} - \mathcal{I}_k^{\mathrm{P}}\right] - \left[\tfrac{1-w_k^{\mathrm{P}}}{2}\mathcal{C}_k^{\mathrm{P}} + \tfrac{1-w_k^{\mathrm{D}}}{2}\mathcal{C}_k^{\mathrm{D}} + \tfrac{w_k^{\mathrm{P}}}{2}\textstyle\sum_J \gamma_I \hat{\mathcal{C}}_{k,J}^{\mathrm{P}} + \tfrac{w_k^{\mathrm{D}}}{2}\hat{\mathcal{C}}_k^{\mathrm{D}}\right] \tag{3.24}$$

$$+ \tfrac{1}{2}\sum_{k=1}^{t} \left[w_k^{\mathrm{P}}\left(\textstyle\sum_{I=1}^{N} \gamma_I \hat{\mathcal{T}}_{k-1,I}^{\mathrm{P}} - \mathcal{T}_k^{\mathrm{P}}\right) - a_k\mu\Delta_{\mathrm{x}}(\boldsymbol{u}, \boldsymbol{x}_k)\right] \tag{3.25}$$

$$+ \tfrac{1}{2}\sum_{k=1}^{t} \left[w_k^{\mathrm{D}}\left(\hat{\mathcal{T}}_{k-1}^{\mathrm{D}} - \mathcal{T}_k^{\mathrm{D}}\right) - a_k\nu\Delta_{\mathrm{y}}(\boldsymbol{v}, \boldsymbol{y}_k)\right] \tag{3.26}$$

The remaining work will be to bound the inner product terms $\mathcal{I}_k^{\mathrm{P}}$ and $\mathcal{I}_k^{\mathrm{D}}$ by quantities that either telescope or can be cancelled by the remaining terms within (3.24). Then, if $w_k^{\mathrm{P}} > 0$ and $w_k^{\mathrm{D}} > 0$ for any k, we will also control (3.25) and (3.26). For these two lines, we will bound the entire sum over k by a term that does not grow with t. For all three lines, this will rely both on the specific form of $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$ (which we call the primal and dual gradient estimates) and the growth conditions placed on the sequence $(a_k)_{k\geq 1}$. While these steps may also be modified slightly in the separable setting (see Definition 3.2.1), the format of the analysis remains the same. The growth of the $(a_k)_{k\geq 1}$ sequence (e.g., constant, polynomial, exponential) is determined if the user knows whether the objective is convex or strongly convex and concave or strongly concave in the primal and dual variables, respectively. Regarding hyperparameters, we consider some variants that do not use the historical regularization by setting w_k^{P} and w_k^{D} to zero in (3.11) and (3.12), meaning that $(\gamma_I)_{I=1}^N$ is no longer a hyperparameter that needs to be set. Thus, the number of hyperparameters decreases considerably for each of the cases in Section 3.4 and Section 3.5.

The comparison points (when used) are reflective of SAGA-style variance reduction methods [Defazio et al., 2014, Palaniappan and Bach, 2016]. In general, they are snapshots of previous iterates and may be used not only to define proximity terms but in the definitions of $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$ as well. The gradient estimates may be computed with adaptive sam-

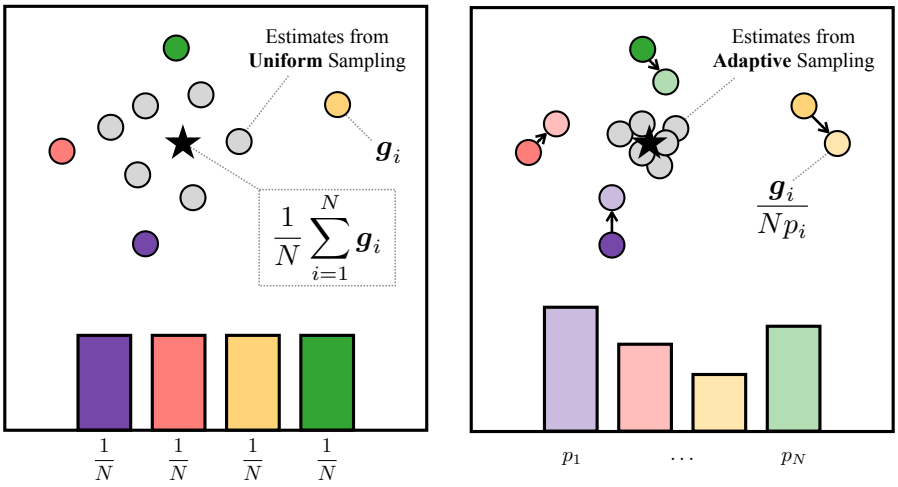

Figure 3.4: **Adapative Sampling for Gradient Estimation** Geometric illustration of adaptive sampling for computing the mean of vectors $\bar{\boldsymbol{g}}_1, \ldots, \bar{\boldsymbol{g}}_N$.

pling, as mentioned in Section 3.1 and visualized in Figure 3.4. By convention, we take any time-dependent element at a negative index to be equal to its initial value (indexed by zero). The number of points stored for the primal updates is equal to the number of blocks N, which may be much smaller than n (avoiding the $O(nd)$ complexity of SAGA). There is only a single additional comparison point in the dual update, incurring a storage cost of $O(n)$. Finally, although the updates include the strong convexity parameters μ and ν, this choice is for readability when proceeding through the analysis. Practically, the growth of the sequence $(A_k)_{k\geq 1}$ will be derived in terms of k with an unknown constant scaling. This constant is a hyperparameter to be searched by the algorithm. For example, when $\mu > 0$ and $\nu > 0$, we have that $a_{k+1} = \alpha A_k = \alpha \sum_{i=1}^{k} a_i$ for $k \geq 1$, where $\alpha > 0$ will be a tunable hyperparameter whose optimal value depends on μ and ν. We conclude this section with a deterministic full vector update method to provide intuition and work with Claim 3.3.1, whereas more advanced algorithms are presented in Section 3.4 and Section 3.5.

3.3.2 Warm-up: deterministic algorithm

Here, we allow the algorithm to access all first-order oracles $\{(f_i, \nabla f_i)\}_{i=1}^n$ in each iteration, for a total per-iteration cost of $O(nd)$. To observe how the gradient estimates can be set to control (3.17), first consider

$$\bar{\boldsymbol{g}}_{k-1} = \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} + \frac{a_{k-1}}{a_k} \left(\nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \nabla f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2} \right). \tag{3.27}$$

Then, by substituting (3.27) into $\mathcal{I}_k^{\mathrm{P}}$, we have that

$$\begin{aligned}
\mathcal{I}_k^{\mathrm{P}} &= a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\
&\quad - a_{k-1} \left\langle \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \nabla f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\
&= a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\
&\quad - a_{k-1} \left\langle \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \nabla f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}, \boldsymbol{u} - \boldsymbol{x}_{k-1} \right\rangle - \mathcal{E}_k^{\mathrm{P}}
\end{aligned} \tag{3.28}$$

where we define the error term

$$\mathcal{E}_k^{\mathrm{P}} = a_{k-1} \left\langle \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \nabla f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}, \boldsymbol{x}_{k-1} - \boldsymbol{x}_k \right\rangle. \tag{3.29}$$

We further set $w_k^{\mathrm{P}} = w_k^{\mathrm{D}} = 0$ to simplify the result of Claim 3.3.1 significantly to become

$$\begin{aligned}
&\sum_{k=1}^{t} a_k \operatorname{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k) \le \mathcal{T}_0^{\mathrm{P}} - \mathcal{T}_t^{\mathrm{P}} + \mathcal{T}_0^{\mathrm{D}} - \mathcal{T}_t^{\mathrm{D}} \\
&\quad + \sum_{k=1}^{t-1} \mathcal{E}_k^{\mathrm{P}} - a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle - \sum_{k=1}^{t} \left(\mathcal{C}_k^{\mathrm{P}} + \mathcal{C}_k^{\mathrm{D}} \right).
\end{aligned} \tag{3.30}$$

Evidently, the goal is for the sum of terms in (3.30) to be a quantity that does not grow with t. A key step in this proof (and the proofs of upcoming results) will be to bound above $\mathcal{E}_k^{\mathrm{P}}$ such that the resulting terms can be canceled by $\mathcal{C}_k^{\mathrm{P}}$ and $\mathcal{C}_k^{\mathrm{D}}$ from (3.14). While we may encounter a similar term when bounding $\mathcal{I}_k^{\mathrm{D}}$, because $\boldsymbol{x}_k$ can be used in the definition of $\bar{\boldsymbol{f}}_{k-1/2}$ and $O(nd)$ operations are permitted, we may simply set $\bar{\boldsymbol{f}}_{k-1/2} = f(\boldsymbol{x}_k)$ to make $\mathcal{I}_k^{\mathrm{D}}$ vanish. Because the terms $\mathcal{E}_k^{\mathrm{P}}$ and $\mathcal{E}_k^{\mathrm{D}}$ appear in some form for every algorithm, we can use

their definitions (and the definitions of subroutines from Algorithm 4) as an identity card for each algorithm, in the form below.

Identity Card 1: Full vector update method

SUBROUTINE$_1$: $\bar{\boldsymbol{g}}_{k-1} = \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} + \frac{a_{k-1}}{a_k}\left(\nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \nabla f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}\right)$

SUBROUTINE$_2$: $\bar{\boldsymbol{f}}_{k-1/2} = f(\boldsymbol{x}_k)$

SUBROUTINE$_3$: None

Primal error: $\mathcal{E}_k^{\mathrm{P}} = a_{k-1}\left\langle \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \nabla f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}, \boldsymbol{x}_{k-1} - \boldsymbol{x}_k \right\rangle$

Dual error: $\mathcal{E}_k^{\mathrm{D}} = 0$

The necessary steps are performed in the following proposition, which also establishes the growth rate of $(a_k)_{k\geq 1}$.

Proposition 3.3.1. *Let $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathrm{ri}(\mathrm{dom}(\phi)) \times \mathrm{ri}(\mathrm{dom}(\psi))$ and $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k\geq 1}$ be generated by the updates (3.11) and (3.12), with $\bar{\boldsymbol{g}}_{k-1}$ given by (3.27) and $\bar{\boldsymbol{f}}_{k-1/2} = f(\boldsymbol{x}_k)$. Select $(a_k)_{k\geq 1}$ to satisfy*

$$a_k \leq \min\left\{ \frac{\sqrt{(A_k\mu + \mu_0)(A_{k-1}\nu + \nu_0)}}{\sqrt{2}G}, \frac{\sqrt{(A_k\mu + \mu_0)(A_{k-1}\mu + \mu_0)}}{2L} \right\}, \tag{3.31}$$

Recalling the notation of (3.13), we have that for any $\boldsymbol{u} \in \mathcal{X}$ and $\boldsymbol{v} \in \mathcal{Y}$,

$$\sum_{k=1}^{t} a_k \,\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k) + \frac{1}{2}\mathcal{T}_t^{\mathrm{P}} + \mathcal{T}_t^{\mathrm{D}} \leq \mathcal{T}_0^{\mathrm{P}} + \mathcal{T}_0^{\mathrm{D}}. \tag{3.32}$$

Proof. We proceed from steps leading up to the gap bound (3.30). Consider $k \geq 2$ (as $\mathcal{E}_k^{\mathrm{P}} = 0$ for $k \leq 1$). Apply Young's inequality with parameter $(A_{k-1}\mu + \mu_0)/2$ and the strong convexity of Bregman divergences to write

$$\mathcal{E}_k^{\mathrm{P}} \leq \frac{1}{2}\mathcal{C}_k^{\mathrm{P}} + \frac{a_{k-1}^2}{A_{k-1}\mu + \mu_0}\left\|\nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}\right\|_{\mathcal{X}^*}^2.$$

To bound the second term above, we first decompose it via Young's inequality. Write

$$\begin{aligned}
&\left\|\nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - f(\boldsymbol{x}_{k-2})^\top \boldsymbol{y}_{k-2}\right\|_{\mathcal{X}^*}^2 \\
&\quad \le 2\left\|(\nabla f(\boldsymbol{x}_{k-1}) - \nabla f(\boldsymbol{x}_{k-2}))^\top \boldsymbol{y}_{k-1}\right\|_{\mathcal{X}^*}^2 \\
&\qquad + 2\left\|\nabla f(\boldsymbol{x}_{k-1})^\top (\boldsymbol{y}_{k-1} - \boldsymbol{y}_{k-2})\right\|_{\mathcal{X}^*}^2 \\
&\quad \le 4L^2 \Delta_{\mathcal{X}}(\boldsymbol{x}_{k-1}, \boldsymbol{x}_{k-2}) + 4G^2 \Delta_{\mathcal{Y}}(\boldsymbol{y}_{k-1}, \boldsymbol{y}_{k-2}),
\end{aligned} \tag{3.33}$$

where the last inequality follows by (3.7) and (3.8) from Assumption 3.2.2. Recall that $\mathcal{C}_k^{\mathrm{P}} = (A_{k-1}\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}_k, \boldsymbol{x}_{k-1})$ and $\mathcal{C}_k^{\mathrm{D}} = (A_{k-1}\nu + \nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{y}_k, \boldsymbol{y}_{k-1})$ (see (3.14)). Combining the steps above for $k \ge 2$ and the condition (3.31) yields

$$\begin{aligned}
\mathcal{E}_k^{\mathrm{P}} &\le \frac{1}{2}\mathcal{C}_k^{\mathrm{P}} + \frac{4a_{k-1}^2}{(A_{k-1}\mu + \mu_0)}\left[\frac{L^2 \mathcal{C}_{k-1}^{\mathrm{P}}}{A_{k-2}\mu + \mu_0} + \frac{G^2 \mathcal{C}_{k-1}^{\mathrm{D}}}{A_{k-2}\nu + \nu_0}\right] \\
&\le \frac{1}{2}\mathcal{C}_k^{\mathrm{P}} + \frac{1}{2}\mathcal{C}_{k-1}^{\mathrm{P}} + \mathcal{C}_{k-1}^{\mathrm{D}}.
\end{aligned} \tag{3.34}$$

Note that for the case of $k = 2$, our choice of a_1 satisfies (3.31) as well. Summing up the current gap bound over $k = 1, \ldots, t$ and dropping non-positive terms yields

$$\begin{aligned}
\sum_{k=1}^{t} \mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k) &\le \mathcal{T}_0^{\mathrm{P}} - \mathcal{T}_t^{\mathrm{P}} + \mathcal{T}_0^{\mathrm{D}} - \mathcal{T}_t^{\mathrm{D}} \\
&\quad - \left(\tfrac{1}{2}\mathcal{C}_t^{\mathrm{P}} + \mathcal{C}_t^{\mathrm{D}}\right) - a_t\left\langle \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{y}_t - \nabla f(\boldsymbol{x}_{t-1})^\top \boldsymbol{y}_{t-1}, \boldsymbol{u} - \boldsymbol{x}_t \right\rangle.
\end{aligned} \tag{3.35}$$

For the remaining inner product term, we apply Young's inequality with parameter $(A_t\mu + \mu_0)/2$ and apply a similar argument as for (3.34):

$$\begin{aligned}
&a_t\left\langle \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{y}_t - \nabla f(\boldsymbol{x}_{t-1})^\top \boldsymbol{y}_{t-1}, \boldsymbol{u} - \boldsymbol{x}_t \right\rangle \\
&\quad = -a_t\left\langle \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{y}_t - \nabla f(\boldsymbol{x}_{t-1})^\top \boldsymbol{y}_{t-1}, \boldsymbol{u} - \boldsymbol{x}_t \right\rangle \\
&\quad \le \frac{a_t^2}{A_t\mu + \mu_0}\left\|\nabla f(\boldsymbol{x}_t)^\top \boldsymbol{y}_t - \nabla f(\boldsymbol{x}_{t-1})^\top \boldsymbol{y}_{t-1}\right\|_{\mathcal{X}^*}^2 + \frac{1}{2}\mathcal{T}_t^{\mathrm{P}} \\
&\quad \le \frac{1}{2}\mathcal{C}_t^{\mathrm{P}} + \mathcal{C}_t^{\mathrm{D}} + \frac{1}{2}\mathcal{T}_t^{\mathrm{P}},
\end{aligned}$$

which will each be cancelled by terms in (3.35), leading to the claimed bound. □

To convert the convergence guarantee in Proposition 3.3.1 into a complexity result, we consider the four possible cases for whether $\mu = 0$ and/or $\nu = 0$, which proves Theorem 3.3.1.

Theorem 3.3.1. *Under Assumption 3.2.2 and Assumption 3.2.1, consider any* $\boldsymbol{u} \in \mathcal{X}$, $\boldsymbol{v} \in \mathcal{Y}$, *and precision* $\varepsilon > 0$. *Define the initial distance term*

$$D_0 = \sqrt{\frac{\mu_0}{\nu_0}}\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + \sqrt{\frac{\nu_0}{\mu_0}}\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_0). \tag{3.36}$$

There exists a choice of the sequence $(a_k)_{k=1}^t$ *such that Algorithm 4 with Identity Card 1 produces an output point* $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \in \mathcal{X} \times \mathcal{Y}$ *satisfying* $\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \leq \varepsilon$ *for* t *that depends on* ε *according to the following iteration complexities.*

Case	Iteration Complexity
$\mu > 0$ *and* $\nu > 0$	$O\left(\left(\frac{L}{\mu} + \frac{G}{\sqrt{\mu\nu}}\right)\ln\left(\frac{\left(L\sqrt{\nu_0/\mu_0}+G\right)D_0}{\varepsilon}\right)\right)$
$\mu > 0$ *and* $\nu = 0$	$O\left(\frac{L}{\mu}\ln\left(\frac{\left(L\sqrt{\nu_0/\mu_0}+G\right)D_0}{\varepsilon}\right) + G\sqrt{\frac{\sqrt{\mu_0/\nu_0}D_0}{\mu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu > 0$	$O\left(\frac{L\sqrt{\nu_0/\mu_0}D_0}{\varepsilon} + G\sqrt{\frac{\sqrt{\nu_0/\mu_0}D_0}{\nu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu = 0$	$O\left(\frac{\left(L\sqrt{\nu_0/\mu_0}+G\right)D_0}{\varepsilon}\right)$

Proof. We first determine the growth of the sequence A_t, so that A_t^{-1} gives the convergence rate in terms of the number of iterations. The growth rate can be derived by providing a sequence $(a_k)_{k\geq 0}$ such that (3.31) is satisfied. For the dependence of the required number of iterations on the suboptimality parameter ε, we write $G_0 A_t^{-1} \leq \varepsilon$ and solve for t, noting that $G_0 = \mu_0\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + \nu_0\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_0)$. In all cases, set $a_1 = \min\left\{\sqrt{\mu_0\nu_0}/\left(\sqrt{2}G\right), \mu_0/\left(2L\right)\right\}$ so that the condition (3.31) is satisfied, and

$$\frac{G_0}{a_1} = \max\left\{2L\sqrt{\frac{\nu_0}{\mu_0}}, \sqrt{2}G\right\}D_0 = O\left(\left(L\sqrt{\tfrac{\nu_0}{\mu_0}} + G\right)D_0\right).$$

Although the $(a_k)_{k\geq 1}$ values are unitless, the relative quantity can help determine the optimal values for μ_0 and ν_0.

Case 1: $\mu > 0$, $\nu > 0$. Let $\alpha = \min\left\{\frac{\sqrt{\mu\nu}}{\sqrt{2}G}, \frac{\mu}{2L}\right\}$. For $k \geq 2$, write $A_k - A_{k-1} = a_k = \alpha\sqrt{A_k A_{k-1}} \geq \alpha A_{k-1}$, which implies that $A_t \geq (1+\alpha)^t a_1$. Then,

$$\frac{G_0}{A_t} \leq \frac{G_0}{a_1(1+\alpha)^t} \lesssim (1+\alpha)^{-t}\left(L\sqrt{\tfrac{\nu_0}{\mu_0}} + G\right) D_0 \overset{\text{want}}{\leq} \varepsilon,$$

which is satisfied for t at the given big-O order.

Case 2: $\mu > 0$, $\nu = 0$. We have that $a_k = \min\left\{\frac{\mu\nu_0}{4G^2}k, \frac{\mu}{2L}A_{k-1}\right\}$ for $k \geq 2$ satisfies the rate condition. Then, there exists a $k^\star \geq 0$ such that $A_t \geq (1+\frac{\mu}{2L})^t a_1$ for all $t < k^\star$ and $A_t \geq \frac{\mu\nu_0}{4G^2}\sum_{k=k^\star+1}^{t} k + (1+\frac{\mu}{2L})^{k^\star} a_1$ for $t \geq k^\star$. To compute the complexity, we consider when either term is dominant. For $(1+\frac{\mu}{2L})^{k^\star} a_1$, we apply the same argument as Case 1. Otherwise, we have that

$$\frac{G_0}{A_t} \lesssim \frac{G^2}{\mu\nu_0 t^2}\left(\mu_0\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + \nu_0\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_0)\right) = \frac{G^2}{\mu t^2}\sqrt{\frac{\mu_0}{\nu_0}} D_0 \overset{\text{want}}{\leq} \varepsilon.$$

Case 3: $\mu = 0$, $\nu > 0$. For $k \geq 2$, we have that $a_k = \min\left\{\frac{\mu_0\nu}{4G^2}k, \frac{\mu_0}{2L}\right\}$ satisfies the rate condition by direct computation. Thus, for $k \leq k^\star = \frac{G^2}{2\nu L}$, we have that $A_t \geq \frac{\mu_0\nu t(t+1)}{8G^2}$ (for which we argue similarly to Case 2 above), and otherwise, $A_t \geq \frac{\mu_0(t-k^\star)}{2L} + \frac{\mu_0\nu k^\star(k^\star+1)}{8G^2}$ (for which we argue similarly to Case 4 below).

Case 4: $\mu = 0$, $\nu = 0$. Here, a_k is equal to a constant, so $A_t = a_1 t$. Then, arguing similarly to Case 1,

$$\frac{G_0}{A_t} \leq \frac{G_0}{a_1 t} \lesssim t^{-1}\left(L\sqrt{\tfrac{\nu_0}{\mu_0}} + G\right) D_0 \overset{\text{want}}{\leq} \varepsilon,$$

which is satisfied for t at the given big-O order, completing the proof. □

We assume that the cost of querying the first-order oracle $\boldsymbol{x} \mapsto (f_j(\boldsymbol{x}), \nabla f_j(\boldsymbol{x}))$ is $O(d)$ for any $j = 1, \ldots, n$, and that the optimization problems (3.11) defining the primal update

and (3.12) defining the dual update can be solved at $\tilde{O}(d)$ and $\tilde{O}(n)$ cost, respectively. Each iteration requires querying all first-order oracles and furthermore computes matrix-vector products using $n \times d$ matrices, thus the overall arithmetic complexity of each iteration of Algorithm 4 with Identity Card 1 is $\tilde{O}(nd)$.

3.4 Stochastic Algorithms for General Objectives

In the case of a randomized algorithm, we allow the vectors $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$ to be defined based on a randomly chosen subset of the indices in $[n]$. This amounts to accessing first-order information $(f_j, \nabla f_j)$ for only some $j \in [n]$. The expressions for $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$ will depend on historical values of the primal and dual iterations, in the spirit of variance reduction or random extrapolation for convex minimization (see, e.g., Gower et al. [2020]). We first describe precisely which comparison points are stored by the algorithm and then specify how they are used to compute $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$. We store N previous primal *iterates* $\hat{\boldsymbol{x}}_{k,1}, \ldots, \hat{\boldsymbol{x}}_{k,N} \in \mathcal{X}$ and a collection of past dual *coordinate blocks* $\hat{\boldsymbol{y}}_k = (\hat{\boldsymbol{y}}_{k,1}, \ldots, \hat{\boldsymbol{y}}_{k,N}) \in \mathcal{Y}$ associated to each block index. We define a primal gradient table $\hat{\boldsymbol{g}}_k = (\hat{\boldsymbol{g}}_{k,1}, \ldots, \hat{\boldsymbol{g}}_{k,n}) \in \mathbb{R}^{n \times d}$ and dual gradient table $\hat{\boldsymbol{f}}_k = (\hat{f}_{k,1}, \ldots, \hat{f}_{k,n}) \in \mathbb{R}^n$ constructed via

$$(\hat{f}_{k,i}, \hat{\boldsymbol{g}}_{k,i}) = (f_i(\hat{\boldsymbol{x}}_{k,I}), \nabla f_i(\hat{\boldsymbol{x}}_{k,I})) \text{ for all } i \in B_I. \tag{3.37}$$

In other words, the tables contain the first-order information of each f_i in block B_I at $\hat{\boldsymbol{x}}_{k,I}$. Note that we do not necessarily need to store the $O(nd)$-sized gradient table, and need only store the $O(Nd)$-sized table of comparison points, which can be much smaller. We update these tables randomly at the end of each iteration by independently sampling block indices R_k and S_k (possibly non-uniformly) and setting each block to

$$\hat{\boldsymbol{x}}_{k,I} = \begin{cases} \boldsymbol{x}_k & \text{if } I = R_k \\ \hat{\boldsymbol{x}}_{k-1,I} & \text{otherwise} \end{cases} \text{ and } \hat{\boldsymbol{y}}_{k,I} = \begin{cases} \boldsymbol{y}_{k,I} & \text{if } I = S_k \\ \hat{\boldsymbol{y}}_{k-1,I} & \text{otherwise} \end{cases}. \tag{3.38}$$

As mentioned in Section 3.3, we define $(\hat{\boldsymbol{x}}_{k,I}, \hat{\boldsymbol{y}}_{k,I}) = (\boldsymbol{x}_0, \boldsymbol{y}_{0,I})$ for any block B_I and iteration $k < 0$. The probability weights that govern the randomness in R_k and S_k are denoted as $\boldsymbol{r} = (r_1, \ldots, r_N)$, and $\boldsymbol{s} = (s_1, \ldots, s_N)$, respectively.

Next, for computing the primal and dual gradient estimates, we sample two more block indices P_k and Q_k with associated probability mass vectors $\boldsymbol{p} = (p_1, \ldots, p_N)$ and $\boldsymbol{q} = (q_1, \ldots, q_N)$, respectively. Letting $\boldsymbol{e}_j$ denote the j-th standard basis vector in $\mathbb{R}^n$, we construct

$$\bar{\boldsymbol{g}}_{k-1} = \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1} + \frac{a_{k-1}}{p_{P_k} a_k} \sum_{i \in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i}) \tag{3.39}$$

$$\bar{\boldsymbol{f}}_{k-1/2} = \hat{\boldsymbol{f}}_k + \frac{a_{k-1}}{q_{Q_k} a_k} \sum_{j \in B_{Q_k}} (f_j(\boldsymbol{x}_{k-1}) - \hat{f}_{k-1,j}) \boldsymbol{e}_j. \tag{3.40}$$

Even though $\boldsymbol{x}_k$ is known during the update of $\boldsymbol{y}_k$, notice that we do not set $\bar{\boldsymbol{f}}_{k-1/2} = f(\boldsymbol{x}_k)$ in this setting to avoid an $\tilde{O}(nd)$ per-iteration complexity. Also notice that $\hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1}$ can be maintained at an $O(nd/N)$ cost per iteration on average (as opposed to $O(nd)$), because both $\hat{\boldsymbol{y}}_k$ and $\hat{\boldsymbol{g}}_k$ only change within a single coordinate block each (R_k and S_k, respectively). This is discussed in detail alongside the per-iteration complexities of specific algorithms.

We collect here the probabilistic notation used in this chapter. We have introduced four random variables for any iteration k: $P_k \sim \boldsymbol{p}$, $Q_k \sim \boldsymbol{q}$, $R_k \sim \boldsymbol{r}$, and $S_k \sim \boldsymbol{s}$. To formally analyze the resulting algorithm, consider a filtered probability space $\mathcal{P} = (\Omega, (\mathcal{F}_k)_{k \geq 0}, \mathbb{P})$, where we use the natural filtration $(\mathcal{F}_k)_{k \geq 0}$ (with $\mathcal{F}_0 = \{\varnothing, \Omega\}$ and $\mathcal{F}_k$ being the σ-algebra generated by the collection $\{(P_\kappa, Q_\kappa, R_\kappa, S_\kappa)\}_{\kappa=1}^k$). We also denote by $\mathcal{F}_{k-1/2}$ the σ-algebra generated by $\{(P_\kappa, Q_\kappa, R_\kappa, S_\kappa)\}_{\kappa=1}^{k-1} \cup \{P_k\}$, which captures information up until and including the computation of $\boldsymbol{x}_k$ (but not $\boldsymbol{y}_k$). Thus, in the language of probability theory, we will say that $\boldsymbol{x}_k$ is $\mathcal{F}_{k-1/2}$-measurable and $\boldsymbol{y}_k$ is $\mathcal{F}_k$-measurable. The full or marginal expectation on $\mathcal{P}$ is given by $\mathbb{E}$, whereas the conditional expectation given $\mathcal{F}_k$ is denoted by $\mathbb{E}_k$. We let $\boldsymbol{z}_{k,J} = (z_{k,j})_{j \in B_J}$ be the block B_J coordinates of a time-indexed vector $\boldsymbol{z}_k \in \mathbb{R}^n$. In the arguments below, we will always consider $(\boldsymbol{u}, \boldsymbol{v})$ that is independent of $\{(P_\kappa, Q_\kappa, R_\kappa, S_\kappa)\}_{\kappa \geq 1}$.

In Section 3.6.1, we then precisely describe how this assumption can be relaxed to achieve the same complexity guarantees for a stronger convergence criterion than the one in Theorem 3.4.1 and Theorem 3.4.2.

We proceed to the convergence analysis. The primal-dual sequence $(\boldsymbol{x}_k, \boldsymbol{y}_k)_{k\geq 0}$ is now a stochastic process; we do not distinguish random variables and realizations when clear from context. We will use the following fact throughout this section: because $\|\cdot\|_{\mathcal{Y}}$ is an ℓ_p-norm with $p \in [1, 2]$ (see Section 3.2), it holds that $\|\cdot\|_{\mathcal{Y}} \geq \|\cdot\|_2$ and $\|\cdot\|_{\mathcal{Y}^*} \leq \|\cdot\|_2$. We will use similar techniques as before to upper bound $\sum_{k=1}^{t} a_k \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k)]$. First, either by using Lemma 3.3.1 and Lemma 3.3.2, we produce a lower bound for $a_k \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}_k)$ and an upper bound for $a_k \mathcal{L}(\boldsymbol{x}_k, \boldsymbol{v})$. Before stating these results, we describe the aspects of the analysis that are similar to the algorithm from Section 3.3.2. Recall the notation box from Section 3.3.

As in Section 3.3.2, we need to use the structure of $\bar{\boldsymbol{g}}_{k-1}$ (defined in (3.39)) and $\bar{\boldsymbol{f}}_{k-1/2}$ (defined in (3.40)) and conditions on $(a_k)_{k\geq 1}$ to control (the expected value of) the inner product terms that appear in (3.42) and (3.43) below. We describe the analogous argument to the one leading to (3.29), except for the stochastic setting. Using $\bar{\boldsymbol{g}}_{k-1}$ as an example, we take the conditional expectation of $\mathcal{I}_k^{\mathrm{P}}$ from (3.17) given $\mathcal{F}_{k-1}$ for $k \geq 1$ (recalling that $\boldsymbol{u}$ is independent of the algorithm randomness) to write

$$\begin{aligned}\mathbb{E}_{k-1}[\mathcal{I}_k^{\mathrm{P}}] = a_k \mathbb{E}_{k-1} \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\ - a_{k-1} \left\langle \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \hat{\boldsymbol{g}}_{k-2}^\top \hat{\boldsymbol{y}}_{k-2}, \boldsymbol{u} - \boldsymbol{x}_{k-1} \right\rangle - \mathbb{E}_{k-1}[\mathcal{E}_k^{\mathrm{P}}] \end{aligned} \tag{3.41}$$

for the error term

$$\mathcal{E}_k^{\mathrm{P}} = a_{k-1} \left\langle \tfrac{1}{p_{P_k}} \textstyle\sum_{i \in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_{k-1} - \boldsymbol{x}_k \right\rangle. \tag{3.42}$$

Note that the telescoping occurs when taking the marginal expectation $\mathbb{E}$ over the entire sequence. The term $\mathcal{E}_k^{\mathrm{D}}$ is defined analogously by substituting $\bar{\boldsymbol{f}}_{k-1/2}$ from (3.40) into the

expression for $\mathcal{I}_k^{\mathrm{D}}$, yielding

$$\mathcal{E}_k^{\mathrm{D}} = a_{k-1} \left\langle \tfrac{1}{q_{Q_k}} \textstyle\sum_{j \in B_{Q_k}} \left(f_j(\boldsymbol{x}_{k-1}) - f_j(\hat{\boldsymbol{x}}_{k-1,Q_k}) \right) \boldsymbol{e}_j, \boldsymbol{y}_k - \boldsymbol{y}_{k-1} \right\rangle . \tag{3.43}$$

We summarize the above using an identity card.

Identity Card 2: Stochastic update method for general objectives

SUBROUTINE$_1$: $\bar{\boldsymbol{g}}_{k-1} = \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1} + \frac{a_{k-1}}{p_{P_k} a_k} \sum_{i \in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i})$

SUBROUTINE$_2$: $\bar{\boldsymbol{f}}_{k-1/2} = \hat{\boldsymbol{f}}_k + \frac{a_{k-1}}{q_{Q_k} a_k} \sum_{j \in B_{Q_k}} (f_j(\boldsymbol{x}_{k-1}) - \hat{f}_{k-1,j}) \boldsymbol{e}_j$

SUBROUTINE$_3$: Update $(\hat{\boldsymbol{x}}_{k,I}, \hat{\boldsymbol{y}}_{k,I})$ for all I using (3.38) and $(\hat{\boldsymbol{g}}_k, \hat{\boldsymbol{f}}_k)$ using (3.37).

Primal error: $\mathcal{E}_k^{\mathrm{P}} = a_{k-1} \left\langle \frac{1}{p_{P_k}} \sum_{i \in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_{k-1} - \boldsymbol{x}_k \right\rangle$

Dual error: $\mathcal{E}_k^{\mathrm{D}} = a_{k-1} \left\langle \frac{1}{q_{Q_k}} \sum_{j \in B_{Q_k}} (f_j(\boldsymbol{x}_{k-1}) - f_j(\hat{\boldsymbol{x}}_{k-1,Q_k})) \, \boldsymbol{e}_j, \boldsymbol{y}_k - \boldsymbol{y}_{k-1} \right\rangle$

Notice that the terms $\mathcal{E}_k^{\mathrm{P}}$ and $\mathcal{E}_k^{\mathrm{D}}$ that appear in (3.42) and (3.43) are measurements of "table bias", or how stale the elements in the tables are compared to the current iterates $\boldsymbol{x}_k$ (for $\mathcal{E}_k^{\mathrm{P}}$) and $\boldsymbol{y}_k$ (for $\mathcal{E}_k^{\mathrm{D}}$). The algorithms below provide two different strategies for achieving convergence while controlling these errors. Because terms of the form $\mathcal{E}_k^{\mathrm{P}}$ will appear in multiple analyses, we collect a repeatedly used bound below.

Lemma 3.4.1. *Consider $\mathcal{F}_{k-1}$-measurable random vectors $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N$ realized in $\mathcal{X}$, and similarly, let $\mathcal{F}_{k-1}$-measurable $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ be realized in $\mathcal{Y}$. For any collection of positive constants $(b_I)_{I=1}^N$ and $(c_I)_{I=1}^N$, we have that*

$$\begin{aligned}
&\mathbb{E} \left\| \tfrac{1}{p_{P_k}} \textstyle\sum_{i \in B_{P_k}} (y_i \nabla f_i(\boldsymbol{x}) - \hat{y}_i \nabla f_i(\hat{\boldsymbol{x}}_{P_k})) \right\|_{\mathcal{X}^*}^2 \\
&\leq 2 \left(\max_J \frac{\boldsymbol{L}_J^2}{p_J b_J} \right) \sum_{I=1}^N b_I \mathbb{E} \left\| \boldsymbol{x} - \hat{\boldsymbol{x}}_I \right\|_{\mathcal{X}}^2 + 2 \left(\max_J \frac{\boldsymbol{G}_J^2}{p_J c_J} \right) \sum_{I=1}^N c_I \mathbb{E} \left\| \boldsymbol{y}_I - \hat{\boldsymbol{y}}_I \right\|_2^2 .
\end{aligned}$$

Proof. First, take the conditional expectation given $\mathcal{F}_{k-1}$ so that

$$\begin{aligned}
&\mathbb{E}_{k-1} \left\| \tfrac{1}{p_{P_k}} \textstyle\sum_{i \in B_{P_k}} (y_i \nabla f_i(\boldsymbol{x}) - \hat{y}_i \nabla f_i(\hat{\boldsymbol{x}}_{P_k})) \right\|_{\mathcal{X}^*}^2 \\
&= \sum_{I=1}^N \frac{1}{p_I} \left\| \textstyle\sum_{i \in B_I} (y_i \nabla f_i(\boldsymbol{x}) - \hat{y}_i \nabla f_i(\hat{\boldsymbol{x}}_I)) \right\|_{\mathcal{X}^*}^2
\end{aligned}$$

Applying Young's inequality, the quantity above is upper bounded via

$$\begin{aligned}
&\sum_{I=1}^{N} \frac{2}{p_I} \left[\left\| \textstyle\sum_{i \in B_I} y_i (\nabla f_i(\boldsymbol{x}) - \nabla f_i(\hat{\boldsymbol{x}}_I)) \right\|_{\mathfrak{X}^*}^2 + \left(\textstyle\sum_{i \in B_I} |y_i - \hat{y}_i| \, \|\nabla f_i(\hat{\boldsymbol{x}}_I)\|_{\mathfrak{X}^*} \right)^2 \right] \\
&\overset{(\circ)}{\leq} \sum_{I=1}^{N} \frac{2}{p_I} \left[\boldsymbol{L}_I^2 \|\boldsymbol{x} - \hat{\boldsymbol{x}}_I\|_{\mathfrak{X}}^2 + \boldsymbol{G}_I^2 \|\boldsymbol{y}_I - \hat{\boldsymbol{y}}_I\|_2^2 \right] \\
&\leq 2 \left(\max_J \frac{\boldsymbol{L}_J^2}{p_J b_J} \right) \sum_{I=1}^{N} b_I \|\boldsymbol{x} - \hat{\boldsymbol{x}}_I\|_{\mathfrak{X}}^2 + 2 \left(\max_J \frac{\boldsymbol{G}_J^2}{p_J c_J} \right) \sum_{I=1}^{N} c_I \|\boldsymbol{y}_I - \hat{\boldsymbol{y}}_I\|_2^2 .
\end{aligned}$$

where we used Assumption 3.2.2 in $(\circ)$. Take the marginal expectation to complete the proof. □

3.4.1 Strategy 1: Non-Uniform Historical Regularization

Here, the goal will be to select the balancing sequences $(w_k^{\mathrm{P}})_{k \geq 1}$ and $(w_k^{\mathrm{D}})_{k \geq 1}$ along with the weights $\boldsymbol{\gamma} = (\gamma_I)_{I=1}^N$ from (3.11) and (3.12) to achieve the desired complexity guarantee. This is the subject of Proposition 3.4.1. The rate conditions will be stated in terms of three quantities that largely depend on the sampling schemes $\boldsymbol{p}$ and $\boldsymbol{q}$ (which are used to define $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$) and the primal regularization weights $\boldsymbol{\gamma}$. Those are

$$\boldsymbol{G}_{\boldsymbol{p}} := \sqrt{\max_I \tfrac{\boldsymbol{G}_I^2}{p_I}}, \boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}} := \sqrt{\max_I \tfrac{\boldsymbol{L}_I^2}{p_I \gamma_I}}, \text{ and } \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{\gamma}} := \sqrt{\max_I \tfrac{\boldsymbol{G}_I^2}{q_I \gamma_I}}. \tag{3.44}$$

Recall that the vectors $\boldsymbol{r} = (r_1, \ldots, r_N)$ and $\boldsymbol{s} = (s_1, \ldots, s_N)$ contain the probabilities by which the primal and dual table blocks are updated at each iteration. We will set these probabilities to the uniform vectors $\boldsymbol{r} = \mathbf{1}/N$ and $\boldsymbol{s} = \mathbf{1}/N$ as they will not affect the convergence rates in this analysis. Moreover, we assume that the objective (3.2) is dual separable (but not necessarily the feasible set). We discuss how this assumption can be avoided by a minor modification of the algorithm after the proof.

Proposition 3.4.1. *Let* $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathrm{ri}(\mathrm{dom}(\phi)) \times \mathrm{ri}(\mathrm{dom}(\psi))$ *and* $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k \geq 1}$ *be generated by the update from Lemma 3.3.1 with non-increasing sequences of weights* $w_k^{\mathrm{P}} \in [0, 1/2]$ *and Lemma 3.3.2 with non-increasing sequence* $w_k^{\mathrm{D}} \in [0, 1/2]$*, with* $\bar{\boldsymbol{g}}_{k-1}$ *and* $\bar{\boldsymbol{f}}_{k-1/2}$ *given*

by (3.39) *and* (3.40)*, respectively. Define* $a_1 = \min\left\{\frac{\sqrt{w_0^{\mathrm{D}}\mu_0\nu_0}}{4\boldsymbol{G_p}}, \frac{\sqrt{w_0^{\mathrm{P}}}\mu_0}{4\sqrt{2}\boldsymbol{L_{p,\gamma}}}, \frac{\sqrt{w_0^{\mathrm{P}}\mu_0\nu_0}}{4\boldsymbol{G_{q,\gamma}}}\right\}$ *and select* $(a_k)_{k\geq 2}$ *such that the conditions*

$$a_k \leq \left\{\frac{\sqrt{w_{k-1}^{\mathrm{D}}(A_k\mu+\mu_0)(A_{k-1}\nu+\nu_0)}}{4\boldsymbol{G_p}}, \frac{\sqrt{w_{k-1}^{\mathrm{P}}(A_k\mu+\mu_0)(A_{k-1}\mu+\mu_0)}}{4\sqrt{2}\boldsymbol{L_{p,\gamma}}}, \frac{\sqrt{w_{k-1}^{\mathrm{P}}(A_k\nu+\nu_0)(A_{k-1}\mu+\mu_0)}}{4\boldsymbol{G_{q,\gamma}}}\right\} \tag{3.45}$$

are satisfied. In addition, impose that for any $\ell = 1, \ldots, t-1$*, it holds that*

$$\frac{\mu}{N}\sum_{k'=0}^{\infty} w_{\ell+k'+1}^{\mathrm{P}} A_{\ell+k'}(1-1/N)^{k'} \leq (w_\ell^{\mathrm{P}} A_\ell + a_\ell)\mu \tag{3.46}$$

$$\frac{\nu}{N}\sum_{k'=0}^{\infty} w_{\ell+k'+1}^{\mathrm{D}} A_{\ell+k'}(1-1/N)^{k'} \leq (w_\ell^{\mathrm{D}} A_\ell + a_\ell)\nu \tag{3.47}$$

and that $A_k \leq \left(1+\frac{1}{2N}\right)^k a_1$ *when* $\mu > 0$ *or* $\nu > 0$*. We have that for any* $\boldsymbol{u} \in \mathcal{X}$ *and* $\boldsymbol{v} \in \mathcal{Y}$*,*

$$\sum_{k=1}^{t} a_k \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k)] + \tfrac{1}{4}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \tfrac{1}{4}\mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] \leq (a_1\mu+\mu_0)\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + (a_1\nu+\nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_0).$$

Proof. Our starting point is Claim 3.3.1, after which we must show that the terms in (3.24), (3.25), and (3.26) are bounded by a quantity that does not grow with t. Recall that for (3.24), we used the three-term decomposition (3.41), which generated two telescoping terms and one error term. The first part of the proof uses this argument and bounds the error term.

1. Controlling (3.24): We follow the arguments at the beginning of this section to produce $\mathcal{E}_k^{\mathrm{P}}$ in (3.42) and $\mathcal{E}_k^{\mathrm{D}}$ in (3.43). We first upper bound the error terms $\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}]$ for $k = 2, \ldots, t-1$ (noting that $\mathcal{E}_1^{\mathrm{P}} = \mathcal{E}_1^{\mathrm{D}} = 0$) and the last term of the telescoping inner products in (3.41). By Young's inequality with parameter $(1-w_k^{\mathrm{P}})(A_{k-1}\mu+\mu_0)/4$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}] &= a_{k-1}\mathbb{E}\left\langle \tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(y_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right\rangle \qquad (3.48)\\ &\leq \frac{1-w_k^{\mathrm{P}}}{4}\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{2a_{k-1}^2}{(1-w_k^{\mathrm{P}})(A_{k-1}\mu+\mu_0)}\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(y_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2\end{aligned}$$

and for the last term of the decomposition (3.41),

$$
\begin{aligned}
&- \mathbb{E}[a_t \left\langle \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{y}_t - \hat{\boldsymbol{g}}_{t-1}^\top \hat{\boldsymbol{y}}_{t-1}, \boldsymbol{u} - \boldsymbol{x}_t \right\rangle] \\
&\leq \frac{1 - w_t^{\mathrm{P}}}{4} \mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \frac{2a_t^2}{(1 - w_t^{\mathrm{P}})(A_t\mu + \mu_0)} \mathbb{E} \left\| \tfrac{1}{p_{P_t}} \textstyle\sum_{i \in B_{P_t}} (y_{t,i} \nabla f_i(\boldsymbol{x}_t) - \hat{y}_{t-1,i} \hat{\boldsymbol{g}}_{t-1,i}) \right\|_{\mathcal{X}^*}^2 \qquad (3.49)
\end{aligned}
$$

To handle the second term in (3.48) and (3.49) for $k \in \{2, \ldots, t\}$, apply Lemma 3.4.1 with $b_I = \gamma_I$ and $c_I = 1$ to achieve

$$
\begin{aligned}
&\mathbb{E} \left\| \tfrac{1}{p_{P_k}} \textstyle\sum_{i \in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i}) \right\|_{\mathcal{X}^*}^2 \\
&\quad \leq 2 \left(\max_I \frac{\boldsymbol{L}_J^2}{p_J \gamma_J} \right) \cdot \sum_{I=1}^N \gamma_I \mathbb{E} \left\| \boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-2,I} \right\|_{\mathcal{X}}^2 + 2 \left(\max_J \frac{\boldsymbol{G}_J^2}{p_J} \right) \mathbb{E} \left\| \boldsymbol{y}_{k-1} - \hat{\boldsymbol{y}}_{k-2} \right\|_2^2 \\
&\quad \leq 4 \underbrace{\left(\max_J \frac{\boldsymbol{L}_J^2}{p_J \gamma_J} \right)}_{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}^2} \cdot \sum_{I=1}^N \gamma_I \mathbb{E}[\Delta_{\mathcal{X}}(\boldsymbol{x}_{k-1}, \hat{\boldsymbol{x}}_{k-2,I})] + 4 \underbrace{\left(\max_J \frac{\boldsymbol{G}_J^2}{p_J} \right)}_{\boldsymbol{G}_{\boldsymbol{p}}^2} \mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{y}_{k-1}, \hat{\boldsymbol{y}}_{k-2,I})],
\end{aligned}
$$

where in the last line we applied $\|\cdot\|_2 \leq \|\cdot\|_{\mathcal{Y}}$ and the strong convexity of Bregman divergences. Recall that $\hat{\mathcal{C}}_{k,I}^{\mathrm{P}} := (A_{k-1}\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}_k, \hat{\boldsymbol{x}}_{k-1,I})$ and $\hat{\mathcal{C}}_k^{\mathrm{D}} := (A_{k-1}\nu + \nu_0)\Delta_{\mathcal{Y}}(\boldsymbol{y}_k, \hat{\boldsymbol{y}}_{k-1})$. Combining the steps above, using that $1/(1 - w_k^{\mathrm{P}}) \leq 2$, and applying the condition (3.45), we have the upper bound

$$
\begin{aligned}
\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}] &\leq \frac{1 - w_k^{\mathrm{P}}}{4} \mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{8a_{k-1}^2 \boldsymbol{G}_{\boldsymbol{p}}^2 \mathbb{E}[\hat{\mathcal{C}}_{k-1}^{\mathrm{D}}]}{(1 - w_k^{\mathrm{P}})(A_{k-1}\mu + \mu_0)(A_{k-2}\nu + \nu_0)} \\
&\quad + \frac{8a_{k-1}^2 \boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}^2 \mathbb{E}\left[\sum_{I=1}^N \gamma_I \hat{\mathcal{C}}_{k-1,I}^{\mathrm{P}}\right]}{(1 - w_k^{\mathrm{P}})(A_{k-1}\mu + \mu_0)(A_{k-2}\mu + \mu_0)} \\
&\leq \frac{1 - w_{k-1}^{\mathrm{P}}}{2} \mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{w_{k-1}^{\mathrm{D}}}{2} \mathbb{E}[\hat{\mathcal{C}}_{k-1}^{\mathrm{D}}] + \frac{w_{k-1}^{\mathrm{P}}}{4} \mathbb{E}\left[\sum_{I=1}^N \gamma_I \hat{\mathcal{C}}_{k-1,I}^{\mathrm{P}}\right]
\end{aligned}
$$

with a similar bound holding for (3.49). These terms will cancel with the corresponding non-positive terms in (3.24).

The upper bounds for $\mathbb{E}[\mathcal{E}_k^{\mathrm{D}}]$ and $-\mathbb{E}[a_k \langle \nabla f(\boldsymbol{x}_k) - \hat{\boldsymbol{f}}_k, \boldsymbol{v} - \boldsymbol{y}_k \rangle]$ follow by very similar arguments as above. Applying Young's inequality with parameter $(1 - w_k^{\mathrm{D}})(A_{k-1}\nu + \nu_0)/4$

we upper bound $\mathbb{E}_{k-1/2}[\mathcal{E}_k^{\mathrm{D}}]$ via

$$a_{k-1}\mathbb{E}_{k-1/2}\left\langle \tfrac{1}{q_{Q_k}}\textstyle\sum_{j\in B_{Q_k}}(f_j(\boldsymbol{x}_{k-1})-\hat{f}_{k-2,j})\boldsymbol{e}_j, \boldsymbol{y}_k-\boldsymbol{y}_{k-1}\right\rangle$$
$$\leq \frac{1-w_k^{\mathrm{D}}}{4}\mathbb{E}_{k-1/2}[\mathcal{C}_k^{\mathrm{D}}] + \frac{2a_{k-1}^2(1-w_k^{\mathrm{D}})^{-1}}{(A_{k-1}\nu+\nu_0)}\sum_{J=1}^{N}\frac{1}{q_J}\left\|\textstyle\sum_{j\in B_J}\left(f_j(\boldsymbol{x}_{k-1})-\hat{f}_{k-1,j}\right)\boldsymbol{e}_j\right\|_{\mathcal{Y}^*}^2 \quad (3.50)$$

Recall that the index R_{k-1} determines which block of the primal table is updated. Thus, it holds in conditional expectation that

$$\begin{aligned}\mathbb{E}_{k-2}\left[\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-1,J}\|_{\mathcal{X}}^2\right] &= \mathbb{E}_{k-2}\left[\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-1,J}\|_{\mathcal{X}}^2\,\mathbb{1}_{J=R_{k-1}}\right] + \mathbb{E}_{k-2}\left[\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-1,J}\|_{\mathcal{X}}^2\,\mathbb{1}_{J\neq R_{k-1}}\right]\\ &= 0 + \mathbb{E}_{k-2}\left[\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-2,J}\|_{\mathcal{X}}^2\,\mathbb{1}_{J\neq R_{k-1}}\right]\\ &\leq \mathbb{E}_{k-2}\left[\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-2,J}\|_{\mathcal{X}}^2\right].\end{aligned}$$

Taking the marginal expectation, the second term of (3.50) can be upper bounded as

$$\begin{aligned}\sum_{J=1}^{N}\frac{1}{q_J}\mathbb{E}\left\|\textstyle\sum_{j\in B_J}\left(f_j(\boldsymbol{x}_{k-1})-\hat{f}_{k-1,j}\right)\boldsymbol{e}_j\right\|_{\mathcal{Y}^*}^2 &\leq \sum_{J=1}^{N}\frac{\boldsymbol{G}_J^2}{q_J}\mathbb{E}\,\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-1,J}\|_{\mathcal{X}}^2\\ &\leq \sum_{J=1}^{N}\frac{\boldsymbol{G}_J^2}{q_J}\mathbb{E}\,\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-2,J}\|_{\mathcal{X}}^2\\ &\leq 2\underbrace{\max_I\tfrac{\boldsymbol{G}_I^2}{q_I\gamma_I}}_{\boldsymbol{G}_{\boldsymbol{q},\gamma}^2}\sum_{J=1}^{N}\gamma_J\Delta_{\mathcal{X}}(\boldsymbol{x}_{k-1},\hat{\boldsymbol{x}}_{k-2,J})_{\mathcal{X}}^2. \quad (3.51)\end{aligned}$$

Invoking condition (3.45) once again and taking the marginal expectation, we have

$$\mathbb{E}[\mathcal{E}_k^{\mathrm{D}}] \leq \frac{1-w_k^{\mathrm{D}}}{4}\mathbb{E}[\mathcal{C}_k^{\mathrm{D}}] + \frac{4a_{k-1}^2\boldsymbol{G}_{\boldsymbol{q},\gamma}^2}{(1-w_k^{\mathrm{D}})(A_{k-1}\nu+\nu_0)(A_{k-2}\mu+\mu_0)}\mathbb{E}\left[\textstyle\sum_{I=1}^{N}\gamma_I\hat{\mathcal{C}}_{k-1,I}^{\mathrm{P}}\right] \quad (3.52)$$
$$\leq \frac{1-w_k^{\mathrm{D}}}{4}\mathbb{E}[\mathcal{C}_k^{\mathrm{D}}] + \frac{w_{k-1}^{\mathrm{P}}}{4}\mathbb{E}\left[\textstyle\sum_{I=1}^{N}\gamma_I\hat{\mathcal{C}}_{k-1,I}^{\mathrm{P}}\right]. \quad (3.53)$$

Similarly, $-\mathbb{E}[a_k\langle\nabla f(\boldsymbol{x}_k)-\hat{\boldsymbol{f}}_k, \boldsymbol{v}-\boldsymbol{y}_k\rangle] \leq \frac{1-w_t^{\mathrm{D}}}{4}\mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] + \frac{w_{t-1}^{\mathrm{P}}}{4}\mathbb{E}\left[\sum_{I=1}^{N}\gamma_I\hat{\mathcal{C}}_{t,I}^{\mathrm{P}}\right]$. All these terms will cancel with corresponding terms in (3.24).

2. Controlling (3.25): For this step, we will express the $\hat{\mathcal{T}}_k^{\mathrm{P}}$ terms as a function of the $\mathcal{T}_k^{\mathrm{P}}$ terms by analyzing the random sampling that governs the block updates. For any $k \geq 1$, write

$$\begin{aligned}\sum_{I=1}^{N} \gamma_I \mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \hat{\boldsymbol{x}}_{k,I})] &= \sum_{I=1}^{N} \gamma_I \left((1/N)\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_k)] + (1-1/N)\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \hat{\boldsymbol{x}}_{k-1,I})]\right) \\ &= (1/N)\sum_{k'=0}^{k} (1-1/N)^{k'} \mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_{k-k'})] \end{aligned} \tag{3.54}$$

Using (3.54), the expression (3.25) can be expanded to

$$\begin{aligned}&\sum_{k=0}^{t-1} w_{k+1}^{\mathrm{P}}(A_k\mu + \mu_0)(1/N)\sum_{k'=0}^{k}(1-1/N)^{k'}\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_{k-k'})] \\ &= \sum_{k'=0}^{t-1}\sum_{k=k'}^{t-1} w_{k+1}^{\mathrm{P}}(A_k\mu + \mu_0)(1/N)(1-1/N))^{k'}\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_{k-k'})] && \text{exchange sums} \\ &= \sum_{k'=0}^{t-1}\sum_{\ell=0}^{t-1-k'} w_{\ell+k'+1}^{\mathrm{P}}(A_{\ell+k'}\mu + \mu_0)\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_\ell)](1/N)(1-1/N)^{k'}. && \text{reparameterize } \ell = k - k'\end{aligned}$$

Using the identity above and $t-1-k' < t-1 < \infty$, we can decompose (3.25) as

$$\begin{aligned}&\frac{1}{2}\sum_{k=1}^{t}\mathbb{E}\left[w_k^{\mathrm{P}}\left(\textstyle\sum_{I=1}^{N}\gamma_I\hat{\mathcal{T}}_{k-1,I}^{\mathrm{P}} - \mathcal{T}_k^{\mathrm{P}}\right) - a_k\mu\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_k)\right] \\ &\leq \frac{1}{2}\sum_{\ell=0}^{t-1}\underbrace{\left(\sum_{k'=0}^{\infty} w_{\ell+k'+1}^{\mathrm{P}}(A_{\ell+k'}\mu + \mu_0)(1/N)(1-1/N)^{k'}\right)}_{\leq w_\ell^{\mathrm{P}}(A_\ell\mu+\mu_0)+a_\ell\mu \text{ for } \ell\geq 1}\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_\ell)] \\ &\quad - \frac{1}{2}\sum_{\ell=0}^{t-1}\left(w_{\ell+1}^{\mathrm{P}}(A_{\ell+1}\mu + \mu_0) + a_{\ell+1}\mu\right)\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_{\ell+1})]\end{aligned}$$

where the inequality under the braces follows from the theorem assumptions. By telescoping the resulting terms, (3.25) is upper bounded by

$$\frac{1}{2}\left(\sum_{k'=0}^{\infty} w_{k'+1}^{\mathrm{P}}(A_{k'}\mu + \mu_0)(1/N)(1-1/N)^{k'}\right)\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_0).$$

Finally, we upper bound the leading constant on $\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0)$ above. We separate the $(A_{k'}\mu + \mu_0)$ constants into two terms each and bound the resulting sums. Because $w^{\mathrm{P}}_{k'+1} \leq 1/2$, we have that $\frac{\mu_0}{2N}\left(\sum_{k'=0}^{\infty} w^{\mathrm{P}}_{k'+1}(1-1/N)^{k'}\right) \leq \frac{\mu_0}{4}$ and because $A_{k'} \leq (1+1/(2N))^{k'} a_1$, we have

$$\frac{\mu}{2N}\left(\sum_{k'=0}^{\infty} w^{\mathrm{P}}_{k'+1} A_{k'}(1-1/N)^{k'}\right) \leq \frac{a_1\mu}{4N}\sum_{k'=0}^{\infty}(1-1/(2N))^{k'} = \frac{a_1\mu}{2}.$$

Summing the two gives the leading constant of $\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0)$ appearing in the statement, as $(a_1\mu/2 + \mu_0/4 + \mu_0/2) \leq a_1\mu + \mu_0$.

3. Controlling (3.26): This will follow from similar steps as those shown above, but will rely upon using the probabilistic arguments on each coordinate block of $\hat{\boldsymbol{y}}_k$. Recall the notation $\Delta_I(\cdot,\cdot)$ from Section 3.2. Using dual-separability of the objective, write

$$\begin{aligned}
\mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{v}, \hat{\boldsymbol{y}}_k)] &= \sum_{I=1}^{N} \mathbb{E}[\Delta_I(\boldsymbol{v}_I, \hat{\boldsymbol{y}}_{k,I})] \\
&= \sum_{I=1}^{N} \left((1/N)\mathbb{E}[\Delta_I(\boldsymbol{v}_I, \boldsymbol{y}_{k,I})] + (1-1/N)\mathbb{E}[\Delta_I(\boldsymbol{v}_I, \hat{\boldsymbol{y}}_{k-1,I})]\right) \\
&= (1/N)\sum_{k'=0}^{k}(1-1/N)^{k'} \underbrace{\sum_{I=1}^{N} \mathbb{E}[\Delta_I(\boldsymbol{v}_I, \boldsymbol{y}_{k-k',I})]}_{\mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_{k-k'})]}.
\end{aligned}$$

The remainder of the argument follows identically to Step 2 and produces the leading constant of $\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_0)$ appearing in the statement. □

Note that the separability of $\Delta_{\mathcal{Y}}(\cdot,\cdot)$ is only used in Step 3, which could be eschewed by updating the entirety of $\hat{\boldsymbol{y}}_{k-1}$ with probability $1/N$ as opposed to the block-wise updates currently being used. Recall the initial distance D_0 from (3.36).

Theorem 3.4.1. *Under Assumption 3.2.2 and Assumption 3.2.1, consider any $\boldsymbol{u} \in \mathcal{X}$, $\boldsymbol{v} \in \mathcal{Y}$ and precision $\varepsilon > 0$. There exists a choice of the sequence $(a_k)_{k=1}^t$, and the parameters w^{P}_k and w^{D}_k such that Algorithm 4 with Identity Card 2 produces an output point $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\mathbb{E}\left[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)\right] \leq \varepsilon$ for t that depends on ε according to*

the following iteration complexities. They depend logarithmically on the constants $C_1 := \sqrt{N}(\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + (\boldsymbol{G}_{\boldsymbol{p}} \vee \boldsymbol{G}_{\boldsymbol{q},\gamma}))D_0 + \mu\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + \nu\Delta_{\mathfrak{Y}}(\boldsymbol{v}, \boldsymbol{y}_0)$, $C_2 := (\sqrt{N}\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + (\boldsymbol{G}_{\boldsymbol{p}} \vee N\boldsymbol{G}_{\boldsymbol{q},\gamma}))D_0 + \mu\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_0)$, *and* $C_3 := (\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + (N\boldsymbol{G}_{\boldsymbol{p}} \vee \boldsymbol{G}_{\boldsymbol{q},\gamma}))D_0 + \nu\Delta_{\mathfrak{Y}}(\boldsymbol{v}, \boldsymbol{y}_0)$.

Case	**Iteration Complexity**
$\mu > 0$ *and* $\nu > 0$	$O\left(\left(N + \frac{\sqrt{N}\boldsymbol{L}_{\boldsymbol{p},\gamma}}{\mu} + \frac{\sqrt{N}(\boldsymbol{G}_{\boldsymbol{p}} \vee \boldsymbol{G}_{\boldsymbol{q},\gamma})}{\sqrt{\mu\nu}}\right)\ln\left(\frac{C_1}{\varepsilon}\right)\right)$
$\mu > 0$ *and* $\nu = 0$	$O\left(\left(N + \frac{\sqrt{N}\boldsymbol{L}_{\boldsymbol{p},\gamma}}{\mu}\right)\ln\left(\frac{C_2}{\varepsilon}\right) + (\boldsymbol{G}_{\boldsymbol{p}} \vee N\boldsymbol{G}_{\boldsymbol{q},\gamma})\sqrt{\frac{\sqrt{\mu_0/\nu_0}D_0 + (a_1\mu/\nu_0)\Delta_{\mathfrak{X}}(\boldsymbol{u},\boldsymbol{x}_0)}{\mu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu > 0$	$O\left(N\ln\left(\frac{C_3}{\varepsilon}\right) + \frac{\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0}D_0}{\varepsilon} + (N\boldsymbol{G}_{\boldsymbol{p}} \vee \boldsymbol{G}_{\boldsymbol{q},\gamma})\sqrt{\frac{\sqrt{\nu_0/\mu_0}D_0 + (a_1\nu/\mu_0)\Delta_{\mathfrak{Y}}(\boldsymbol{v},\boldsymbol{y}_0)}{\nu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu = 0$	$O\left(\frac{\left(\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + (\boldsymbol{G}_{\boldsymbol{p}} \vee \boldsymbol{G}_{\boldsymbol{q},\gamma})\right)D_0}{\varepsilon}\right)$

Proof. We split the proof into the same case-by-case strategy as employed in Theorem 3.3.1. While we may match those arguments exactly for most of the conditions of Proposition 3.4.1, we need to additionally set the correct values of the sequences $(w_k^{\mathrm{P}})_{k\geq 1}$ and $(w^{\mathrm{D}})_{k\geq 1}$ to complete the analysis. In all cases, the requirement that $A_k \leq \left(1 + \frac{1}{2N}\right)^k a_1$ introduces a term of the form $N\ln\left(\frac{G_0}{a_1\varepsilon}\right)$ to the iteration complexity, where $G_0 = (a_1\mu + \mu_0)\Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + (a_1\nu + \nu_0)\Delta_{\mathfrak{Y}}(\boldsymbol{v}, \boldsymbol{y}_0)$. When substituting the particular values of $(w_0^{\mathrm{P}}, w_0^{\mathrm{D}})$ in each case, we derive the constants (C_1, C_2, C_3).

Case 1: $\mu > 0$**,** $\nu > 0$**.** We consider here a constant choice of the sequences and $w_k^{\mathrm{P}} = w_0^{\mathrm{P}}$ and $w_k^{\mathrm{D}} = w_0^{\mathrm{D}}$. Then, all conditions on the growth of $(a_k)_{k\geq 1}$ can be satisfied using $a_k = \alpha A_{k-1}$, where

$$\alpha \lesssim \min\left\{\frac{\sqrt{w_0^{\mathrm{D}}\mu\nu}}{\boldsymbol{G}_{\boldsymbol{p}}}, \frac{\sqrt{w_0^{\mathrm{P}}\mu\nu}}{\boldsymbol{G}_{\boldsymbol{q},\gamma}}, \frac{\sqrt{w_0^{\mathrm{P}}}\mu}{\boldsymbol{L}_{\boldsymbol{p},\gamma}}\right\}.$$

This implies that $A_k = (1+\alpha)A_{k-1}$. Then, considering (3.46), we have that

$$\frac{1}{N}\sum_{k'=0}^{\infty} w_0^{\mathrm{P}} A_{\ell+k'}(1-1/N)^{k'} = \frac{1}{N} w_0^{\mathrm{P}} A_\ell \sum_{k'=0}^{\infty}(1+\alpha)^{k'}(1-1/N)^{k'}.$$

Consider a setting of α and a constant $c > 0$ (which may depend on N) such that

$$\frac{1}{N}\sum_{k'=0}^{\infty}(1+\alpha)^{k'}(1-1/N)^{k'} \leq 1 + \frac{c\alpha}{1+\alpha}, \tag{3.55}$$

where the right-hand side can be tightened to $1 + c\alpha/2$ when $\alpha \leq 1$. Then, by setting $w_0^{\mathrm{P}} \leq 2c^{-1}$, the condition (3.46) is satisfied (with an identical argument holding for (3.47)). When $N = 1$, the term above vanishes, so we may consider $N \geq 2$. To satisfy (3.55), we need that $\alpha \leq \frac{N-c/2-1}{1-N} = \frac{1}{N-1}$ for $c = 2N$, which imposes the condition that $\alpha \leq \frac{1}{N-1}$.

Case 2: $\mu > 0$**,** $\nu = 0$**.** We set $w_k^{\mathrm{D}} = 1/2$ for all $k \geq 0$ and need only set w_k^{P}, which will be piece-wise constant. For all $\ell \leq k^\star$ such that the second condition of (3.45) is the dominant condition, that is, $a_\ell = \alpha_\ell A_{\ell-1}$ for $\alpha_\ell \lesssim \frac{\sqrt{w_{\ell-1}^{\mathrm{P}}}\mu}{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}}$, we may set $w_\ell^{\mathrm{P}} = 1/N$ and $\alpha_\ell \equiv \alpha_{k^\star}$ for all $\ell \in \{0, \ldots, k^\star\}$ and follow the argument of Case 1 to achieve the first term in the complexity. For $\ell \geq k^\star + 1$, we will derive the value of $w_{k^\star+1}^{\mathrm{P}}$. We have that $a_\ell = c\mu\nu_0 \min\left\{\frac{1}{\boldsymbol{G}_{\boldsymbol{p}}^2}, \frac{w_{k^\star+1}^{\mathrm{P}}}{\boldsymbol{G}_{\boldsymbol{q},\boldsymbol{\gamma}}^2}\right\}\ell$ for an absolute constant $c > 0$, and moreover, that

$$A_\ell \leq c\mu\nu_0 \min\left\{\frac{1}{\boldsymbol{G}_{\boldsymbol{p}}^2}, \frac{w_{k^\star+1}^{\mathrm{P}}}{\boldsymbol{G}_{\boldsymbol{q},\boldsymbol{\gamma}}^2}\right\}\frac{\ell(\ell+1)}{2}.$$

Furthermore, the condition (3.46) is satisfied if

$$\frac{1}{N} w_{k^\star+1}^{\mathrm{P}} \sum_{k'=0}^{\infty} \frac{(\ell+k')(\ell+1+k')}{2}(1-1/N)^{k'} \leq w_{k^\star+1}^{\mathrm{P}} \frac{\ell(\ell+1)}{2} + \ell$$

which in turn is satisfied when

$$w_{k^\star+1}^{\mathrm{P}} \underbrace{\frac{1}{N}\sum_{k'=0}^{\infty} \frac{k'(2\ell+k'+1)}{2}(1-1/N)^{k'}}_{\leq N\ell+N^2+1/2} \leq \ell,$$

where the upper bound follows by summing over k' and applying $(1 - 1/N) \leq 1$. Set

$$w^{\mathrm{P}}_{k^\star+1} = \frac{1}{N + N^2/k^\star + 1/(2k^\star)},$$

which introduces the term $\frac{N\boldsymbol{G}_{\boldsymbol{q},\boldsymbol{\gamma}}}{\sqrt{\mu\varepsilon}} \cdot \sqrt{\frac{G_0}{\nu_0}}$ in the given complexity.

Case 3: $\mu = 0$, $\nu > 0$. Here, we may set $w^{\mathrm{P}}_k = 1/2$ for all $k \geq 0$ and derive the required setting for $(w^{\mathrm{D}}_k)_{k\geq 1}$. We may repeat the argument above and for Case 3 of Theorem 3.3.1 to set $w^{\mathrm{D}}_k \sim 1/N^2$, which achieves the given complexity.

Case 4: $\mu = 0$, $\nu = 0$. The conditions (3.46) and (3.47) vanish, so we reuse the sequence $A_t = O(\min\left\{\sqrt{\mu_0\nu_0}/G, \mu_0/L\right\} t)$ as before to complete the proof. □

We instantiate the problem constants by selecting a sampling scheme. Recall from Section 3.2 that $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)$ for $\boldsymbol{\lambda}_I := \sqrt{\boldsymbol{G}_I^2 + \boldsymbol{L}_I^2}$ along with the constants from (3.44). The non-uniform sampling complexity given below follows by letting $p_I \propto \boldsymbol{\lambda}_I$, $\gamma_I \propto \boldsymbol{\lambda}_I$, and $q_I \propto \boldsymbol{G}_I$. The constants appearing in Theorem 3.4.1 are tabulated below.

Constant	**Uniform Sampling**	**Non-Uniform Sampling**
$\boldsymbol{G}_{\boldsymbol{p}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}$	$N \left\|\boldsymbol{G}\right\|_\infty$	$\left\|\boldsymbol{\lambda}\right\|_1^{1/2} \left\|\boldsymbol{G}\right\|_1^{1/2}$
$\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}$	$N \left\|\boldsymbol{L}\right\|_\infty$	$\left\|\boldsymbol{\lambda}\right\|_1$

We discuss memory and per-iteration complexity of the method. If we show that the optimization problem (3.11) can be solved at $\tilde{O}(d)$ cost, then the total arithmetic complexity is given by $\tilde{O}(n(d+N)/N)$, where we recall that we assume uniform block sizes for arithmetic complexity discussions. The relevant terms are the matrix-vector product $\hat{\boldsymbol{g}}_k^\top \hat{\boldsymbol{y}}_k$ and the sum of weighted Bregman divergences.

For the former, let $\hat{\boldsymbol{g}}_{k,I} \in \mathbb{R}^{n/N\times d}$ denote the matrix containing the primal gradients for the elements in block I. Noting that R_k is the block of the primal table that is updated at

each iteration and S_k plays the same role for the dual table, it holds that

$$\begin{aligned}\hat{\boldsymbol{g}}_k^\top \hat{\boldsymbol{y}}_k &= \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_k + (\hat{\boldsymbol{g}}_{k,R_k} - \hat{\boldsymbol{g}}_{k-1,R_k})\hat{\boldsymbol{y}}_{k,R_k} \\ &= \hat{\boldsymbol{g}}_{k-1}^\top \left(\hat{\boldsymbol{y}}_{k-1} + \textstyle\sum_{j\in B_{S_k}} (\hat{y}_{k,j} - \hat{y}_{k-1,j})\boldsymbol{e}_j\right) + (\hat{\boldsymbol{g}}_{k,R_k} - \hat{\boldsymbol{g}}_{k-1,R_k})^\top \hat{\boldsymbol{y}}_{k,R_k} \\ &= \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1} + \hat{\boldsymbol{g}}_{k-1,S_k}^\top (\hat{\boldsymbol{y}}_{k,S_k} - \hat{\boldsymbol{y}}_{k-1,S_k}) + (\hat{\boldsymbol{g}}_{k,R_k} - \hat{\boldsymbol{g}}_{k-1,R_k})^\top \hat{\boldsymbol{y}}_{k,R_k}. \end{aligned} \tag{3.56}$$

Assuming that $\hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1} \in \mathbb{R}^d$ is already stored, everything above can be computed with $\hat{\boldsymbol{g}}_{k,R_k}$, $\hat{\boldsymbol{y}}_{k,S_k}$, and past information at cost $O(nd/N)$. Furthermore, we need not retain the entire table $\hat{\boldsymbol{g}}_k \in \mathbb{R}^{n\times d}$, as $\hat{\boldsymbol{g}}_{k-1,R_k}$ and $\hat{\boldsymbol{g}}_{k-1,S_k}$ above can be recomputed from $\hat{\boldsymbol{x}}_{k-1,R_k}$ and $\hat{\boldsymbol{x}}_{k-1,R_k}$. The entire memory footprint is $O(Nd+n)$, which could be much smaller than $O(nd)$ (for instance, when $N=d$).

For the latter, letting φ be the generator of $\Delta_\chi(\cdot,\cdot)$, we write

$$\sum_{I=1}^N \gamma_I \Delta_\chi(\boldsymbol{x}, \hat{\boldsymbol{x}}_{k,I}) = \varphi(\boldsymbol{x}) + \left\langle \sum_{I=1}^N \gamma_I \nabla\varphi(\hat{\boldsymbol{x}}_{k,I}), \boldsymbol{x} \right\rangle + \text{const}(\boldsymbol{x}),$$

where the term $\text{const}(\boldsymbol{x})$ does not vary with respect to $\boldsymbol{x}$. It then holds that

$$\sum_{I=1}^N \gamma_I \nabla\varphi(\hat{\boldsymbol{x}}_{k,I}) = \sum_{I=1}^N \gamma_I \nabla\varphi(\hat{\boldsymbol{x}}_{k-1,I}) + \gamma_{R_k}\left(\nabla\varphi(\hat{\boldsymbol{x}}_{k,R_k}) - \nabla\varphi(\hat{\boldsymbol{x}}_{k-1,R_k})\right),$$

so we need only compute $\nabla\varphi(\hat{\boldsymbol{x}}_{k,R_k})$ at each iteration. Retaining the $(\nabla\varphi(\hat{\boldsymbol{x}}_{k,I}))_{I=1}^N$ comes at an $O(Nd)$ storage cost, which is the same as the table itself. The total per-iteration complexity is then $\tilde{O}(n(d+N)/N)$.

3.4.2 Strategy 2: Non-Uniform Block Replacement Probabilities

In the previous approach, we relied on the non-uniform weights $(\gamma_I)_{I=1}^N$ in order to achieve complexities that were independent of the number of blocks N. Here, rather than relying on the historical regularization, we will instead tune the sampling probabilities $\boldsymbol{r} = (r_1, \ldots, r_N)$ and $\boldsymbol{s} = (s_1, \ldots, s_N)$, which govern the element R_k of $\hat{\boldsymbol{x}}_{k-1,1}, \ldots, \hat{\boldsymbol{x}}_{k-1,N}$ and which coordinate block S_k of $\hat{\boldsymbol{y}}_{k-1}$ gets updates at each iteration k (see (3.38)).

We bound the same terms as in Proposition 3.4.1, although (3.25) and (3.26) are immediately non-positive as we will set $w_k^{\mathrm{P}} = 0$ and $w_k^{\mathrm{D}} = 0$ for all k. This argument is employed in Proposition 3.4.2. Similarly to (3.44), the resulting complexity will depend on the sampling probabilities $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, and $\boldsymbol{s}$ through the following constants:

$$\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} := \sqrt{\max_{I\in[N]} \frac{\boldsymbol{G}_I^2}{p_I s_I^2}}, \boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}} := \sqrt{\sum_{I=1}^{N} \frac{\boldsymbol{L}_I^2}{p_I r_I^2}}, \text{ and } \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}} := \sqrt{\sum_{I=1}^{N} \frac{\boldsymbol{G}_I^2}{q_I r_I^2}}.$$

Observe the following.

Proposition 3.4.2. *Let $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathrm{ri}(\mathrm{dom}(\phi)) \times \mathrm{ri}(\mathrm{dom}(\psi))$ and $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k\geq 1}$ be generated using $\bar{\boldsymbol{g}}_{k-1}$ and $\bar{\boldsymbol{f}}_{k-1/2}$ given by (3.39) and (3.40), respectively. Define $a_1 = \frac{1}{15}\min\left\{\frac{\sqrt{\mu_0\nu_0}}{\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}}\vee\boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}}, \frac{\mu_0}{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}}\right\}$ and select $(a_k)_{k\geq 2}$ such that both the conditions*

$$\frac{a_k^2}{A_k\mu + \mu_0} \leq \min_I (1 + (s_I \wedge r_I)/5)\frac{a_{k-1}^2}{A_{k-1}\mu + \mu_0} \tag{3.57}$$

$$\frac{a_k^2}{A_k\nu + \nu_0} \leq \min_I (1 + r_I/5)\frac{a_{k-1}^2}{A_{k-1}\nu + \nu_0} \tag{3.58}$$

and

$$a_k \leq \min\left\{\frac{\sqrt{(A_k\mu + \mu_0)(A_{k-1}\nu + \nu_0)}}{10\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}}}, \frac{\sqrt{(A_k\mu + \mu_0)(A_{k-1}\mu + \mu_0)}}{10\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}}, \frac{\sqrt{(A_k\nu + \nu_0)(A_{k-1}\mu + \mu_0)}}{15\boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}}\right\} \tag{3.59}$$

are satisfied. We have that for any $\boldsymbol{u} \in \mathcal{X}$ and $\boldsymbol{v} \in \mathcal{Y}$,

$$\sum_{k=1}^{t} a_k \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k)] + \frac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \frac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] \leq \mathcal{T}_0^{\mathrm{P}} + \mathcal{T}_0^{\mathrm{D}}.$$

Proof. As stated before, we aim to show that the sum of terms in (3.24) is upper bounded by a constant independent of t. This is composed of the terms $\mathbb{E}[\mathcal{I}_k^{\mathrm{P}}]$ and $\mathbb{E}[\mathcal{I}_k^{\mathrm{D}}]$. We divide this task into bounding the primal and dual components separately. As before, by applying the argument leading to the expressions (3.42) and (3.43), bounding the inner product terms $\mathbb{E}[\mathcal{I}_k^{\mathrm{P}}]$ and $\mathbb{E}[\mathcal{I}_k^{\mathrm{D}}]$ reduces to bounding the error terms $\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}]$ and $\mathbb{E}[\mathcal{E}_k^{\mathrm{D}}]$ and the final element of the telescoping inner product terms $a_t \left\langle \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{y}_t - \hat{\boldsymbol{g}}_{t-1}^\top \hat{\boldsymbol{y}}_{t-1}, \boldsymbol{u} - \boldsymbol{x}_t \right\rangle$

and $a_t\langle \nabla f(\boldsymbol{x}_t) - \hat{\boldsymbol{f}}_t, \boldsymbol{v} - \boldsymbol{y}_t\rangle$.

1. Controlling $\mathbb{E}[\mathcal{I}_k^{\mathrm{P}}]$: The first step follows similarly to Step 1 from the proof of Proposition 3.4.1. By Young's inequality with parameter $(A_{k-1}\mu + \mu_0)/2$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}] &= a_{k-1}\mathbb{E}\left\langle \tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(y_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right\rangle \\ &\le \frac{1}{2}\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{a_{k-1}^2}{A_{k-1}\mu+\mu_0}\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(y_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2 \end{aligned} \tag{3.60}$$

and

$$\begin{aligned}&-\mathbb{E}[a_k\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k\rangle] \\ &\le \frac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \frac{a_{t-1}^2}{A_{t-1}\mu+\mu_0}\mathbb{E}\left\|\tfrac{1}{p_{P_t}}\textstyle\sum_{i\in B_{P_t}}(y_{t,i}\nabla f_i(\boldsymbol{x}_t) - \hat{y}_{t-1,i}\hat{\boldsymbol{g}}_{t-1,i})\right\|_{\mathcal{X}^*}^2. \end{aligned} \tag{3.61}$$

Then, we apply Lemma 3.4.1 with $b_I = 1$ and $c_I = 1$ to handle the second term in (3.60) (and (3.61)), and write

$$\begin{aligned}&\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(y_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2 \\ &\le 2\left(\max_J \frac{\boldsymbol{L}_J^2}{p_J}\right)\sum_{I=1}^N \mathbb{E}\left\|\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-2,I}\right\|_{\mathcal{X}}^2 + 2\left(\max_J \frac{\boldsymbol{G}_J^2}{p_J}\right)\sum_{I=1}^N \mathbb{E}\left\|\boldsymbol{y}_{k-1,I} - \hat{\boldsymbol{y}}_{k-2,I}\right\|_2^2. \end{aligned}$$

Individually, the colored table bias terms can be further upper bounded by applying Lemma 3.4.2 (stated after this proof), so that the sums of (3.60) and (3.61) can be further developed to

$$\begin{aligned}\sum_{k=2}^t \mathbb{E}[\mathcal{I}_k^{\mathrm{P}}] &\le \frac{1}{2}\sum_{k=2}^t \mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] \\ &\quad + 10\sum_{k=1}^t \frac{a_k^2}{A_k\mu+\mu_0}\sum_{I=1}^N \frac{\boldsymbol{G}_I^2}{p_I s_I}\sum_{k'=1}^k (1-s_I/2)^{k-k'}\mathbb{E}\left\|\boldsymbol{y}_{k',I} - \boldsymbol{y}_{k'-1,I}\right\|_2^2 \qquad (3.62) \\ &\quad + 10\sum_{k=1}^t \frac{a_k^2}{A_k\mu+\mu_0}\sum_{I=1}^N \frac{\boldsymbol{L}_I^2}{p_I r_I}\sum_{k'=1}^k (1-r_I/2)^{k-k'}\mathbb{E}\left\|\boldsymbol{x}_{k'} - \boldsymbol{x}_{k'-1}\right\|_{\mathcal{X}}^2. \qquad (3.63)\end{aligned}$$

To control the resulting sums (3.62) and (3.63), we exchange the order of summation to

compute them. First, (3.62) can be written as

$$
\begin{aligned}
&10 \sum_{k'=1}^{t} \sum_{I=1}^{N} \frac{\boldsymbol{G}_I^2}{p_I s_I} \mathbb{E} \left\| \boldsymbol{y}_{k',I} - \boldsymbol{y}_{k'-1,I} \right\|_2^2 \cdot \sum_{k=k'}^{t} \frac{a_k^2}{A_k \mu + \mu_0} (1 - s_I/2)^{k-k'} \\
&\leq 50 \sum_{k'=1}^{t} \frac{a_{k'}^2}{A_{k'} \mu + \mu_0} \sum_{I=1}^{N} \frac{\boldsymbol{G}_I^2}{p_I s_I^2} \mathbb{E} \left\| \boldsymbol{y}_{k',I} - \boldsymbol{y}_{k'-1,I} \right\|_2^2 \qquad (3.64) \\
&\leq 50 \underbrace{\left(\max_{I \in [N]} \frac{\boldsymbol{G}_I^2}{p_I s_I^2} \right)}_{\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}}^2} \sum_{k'=1}^{t} \frac{a_{k'}^2}{A_{k'} \mu + \mu_0} \mathbb{E} \left\| \boldsymbol{y}_{k'} - \boldsymbol{y}_{k'-1} \right\|_{\mathcal{Y}}^2, \qquad (3.65)
\end{aligned}
$$

where the inequality (3.64) follows by the given assumption (3.57) that $\frac{a_k^2}{A_k \mu + \mu_0} \leq \min_I (1 + s_I/5) \frac{a_{k-1}^2}{A_{k-1} \mu + \mu_0}$ and the sequence of steps

$$
\begin{aligned}
\frac{a_k^2}{A_k \mu + \mu_0} \sum_{k=k'}^{t} (1 - s_I/2)^{k-k'} &\leq \sum_{k=k'}^{t} [(1 - s_I/2)(1 + s_I/5)]^{k-k'} \frac{a_{k'}^2}{A_{k'} \mu + \mu_0} \\
&\leq \sum_{k=k'}^{t} (1 - s_I/5)^{k-k'} \frac{a_{k'}^2}{A_{k'} \mu + \mu_0} \leq \frac{5}{s_I} \frac{a_{k'}^2}{A_{k'} \mu + \mu_0}.
\end{aligned}
$$

In (3.65), we also used that $\|\cdot\|_2 \leq \|\cdot\|_{\mathcal{Y}}$. This argument is the most technical part of the analysis and is repeated two more times in the remainder of the proof. The first of the two is to upper bound (3.63) by the quantity

$$
50 \underbrace{\left(\sum_{I=1}^{N} \frac{\boldsymbol{L}_I^2}{p_I r_I^2} \right)}_{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}^2} \sum_{k'=1}^{t} \frac{a_{k'}^2}{A_{k'} \mu + \mu_0} \mathbb{E} \left\| \boldsymbol{x}_{k'} - \boldsymbol{x}_{k'-1} \right\|_{\mathcal{X}}^2, \qquad (3.66)
$$

whereas the second appears in the steps used to bound the dual error terms below.

2. Controlling $\mathbb{E}[\mathcal{I}_k^{\mathrm{D}}]$: The following steps are the dual analog of ones shown above for the primal terms. Applying Young's inequality with parameter $(A_{k-1}\nu+\nu_0)/2$ we have that

$$\begin{aligned}
\mathbb{E}_{k-1/2}[\mathcal{E}_k^{\mathrm{D}}] &= a_{k-1}\mathbb{E}_{k-1/2}\left\langle \tfrac{1}{q_{Q_k}}\textstyle\sum_{j\in B_{Q_k}}(f_j(\boldsymbol{x}_{k-1})-\hat{f}_{k-2,j})\boldsymbol{e}_j, \boldsymbol{y}_k-\boldsymbol{y}_{k-1}\right\rangle \\
&\leq \frac{1}{2}\mathbb{E}_{k-1/2}[\mathcal{C}_k^{\mathrm{D}}] + \frac{a_{k-1}^2}{A_{k-1}\nu+\nu_0}\sum_{J=1}^{N}\frac{1}{q_J}\left\|\textstyle\sum_{j\in B_J}\left(f_j(\boldsymbol{x}_{k-1})-\hat{f}_{k-1,j}\right)\boldsymbol{e}_j\right\|_{\mathcal{Y}^*}^2 \\
&\leq \frac{1}{2}\mathbb{E}_{k-1/2}[\mathcal{C}_k^{\mathrm{D}}] + \frac{a_{k-1}^2}{A_{k-1}\nu+\nu_0}\sum_{J=1}^{N}\frac{1}{q_J}\left\|\textstyle\sum_{j\in B_J}\left(f_j(\boldsymbol{x}_{k-1})-\hat{f}_{k-1,j}\right)\boldsymbol{e}_j\right\|_{2}^2 \qquad (3.67)\\
&\leq \frac{1}{2}\mathbb{E}_{k-1/2}[\mathcal{C}_k^{\mathrm{D}}] + \frac{a_{k-1}^2}{A_{k-1}\nu+\nu_0}\sum_{J=1}^{N}\frac{\boldsymbol{G}_J^2}{q_J}\left\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-1,J}\right\|_{\mathcal{X}}^2
\end{aligned}$$

where in (3.67) we used that $\|\cdot\|_{\mathcal{Y}^*}\leq\|\cdot\|_2$. Summing over k and taking the marginal expectation,

$$\sum_{k=2}^{t}\mathbb{E}[\mathcal{I}_k^{\mathrm{D}}] \leq \frac{1}{2}\sum_{k=2}^{t}\mathbb{E}[\mathcal{C}_k^{\mathrm{D}}] + \frac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] + \frac{a_{k-1}^2}{A_{k-1}\nu+\nu_0}\sum_{J=1}^{N}\frac{\boldsymbol{G}_J^2}{q_J}\left\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-1,J}\right\|_{\mathcal{X}}^2, \qquad (3.68)$$

where (3.68) can be upper bounded using the same arguments leading to (3.66) under the given assumption (3.58), yielding

$$25\underbrace{\left(\sum_{J=1}^{N}\frac{\boldsymbol{G}_J^2}{q_J r_J^2}\right)}_{\boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}^2}\sum_{k'=1}^{t}\frac{a_{k'}^2}{A_{k'}\nu+\nu_0}\mathbb{E}\left\|\boldsymbol{x}_{k'}-\boldsymbol{x}_{k'-1}\right\|_{\mathcal{X}}^2. \qquad (3.69)$$

To summarize progress thus far, we must cancel the terms (3.65), (3.66), and (3.69) to complete the proof, which requires setting the appropriate conditions on the sequence $(a_k)_{k\geq 1}$.

3. Deriving the rate conditions: Under the condition (3.59), we may bound (3.65) by $\frac{1}{2}\sum_{k'=1}^{t}\mathbb{E}[\mathcal{C}_{k'}^{\mathrm{D}}]$, (3.66) by $\frac{1}{4}\sum_{k'=1}^{t}\mathbb{E}[\mathcal{C}_{k'}^{\mathrm{P}}]$, and (3.69) by $\frac{1}{4}\sum_{k'=1}^{t}\mathbb{E}[\mathcal{C}_{k'}^{\mathrm{P}}]$. All terms of (3.24) now cancel, completing the proof. □

The following technical lemma was used to express the terms that quantified the table bias terms (distance between the iterates and their counterparts in the respective tables)

with terms that can be canceled by quantities in (3.24).

Lemma 3.4.2. *[Diakonikolas, 2025, Lemma 2] For any $k \geq 1$ and $I \in [N]$, the following hold:*

$$\mathbb{E}\,\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k-1,I}\|_{\mathcal{X}}^2 \leq \frac{5}{r_I}\sum_{k'=1}^{k}(1-r_I/2)^{k-k'}\mathbb{E}\,\|\boldsymbol{x}_{k'} - \boldsymbol{x}_{k'-1}\|_{\mathcal{X}}^2\,,$$
$$\mathbb{E}\,\|\boldsymbol{y}_{k,I} - \hat{\boldsymbol{y}}_{k-1,I}\|_2^2 \leq \frac{5}{s_I}\sum_{k'=1}^{k}(1-s_I/2)^{k-k'}\mathbb{E}\,\|\boldsymbol{y}_{k',I} - \boldsymbol{y}_{k'-1,I}\|_2^2\,.$$

We now convert the Proposition 3.4.2 into a complexity guarantee, using again the constant D_0 from Theorem 3.3.1.

Theorem 3.4.2. *Under Assumption 3.2.2 and Assumption 3.2.1, consider any $\boldsymbol{u} \in \mathcal{X}$, $\boldsymbol{v} \in \mathcal{Y}$ and precision $\varepsilon > 0$. Assume that $\min_I r_I \geq 1/(2N)$ and $\min_I s_I \geq 1/(2N)$. Define There exists a choice of the sequence $(a_k)_{k=1}^t$ such that Algorithm 4 with Identity Card 2 produces an output point $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \in \mathcal{X}\times\mathcal{Y}$ satisfying $\mathbb{E}\left[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)\right] \leq \varepsilon$ for t that depends on ε according to the following iteration complexities. Below, we use the constant $C_0 := \left(\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}\sqrt{\nu_0/\mu_0} + (\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}})\right) D_0$.*

Case	Iteration Complexity
$\mu > 0$ *and* $\nu > 0$	$O\left(\left(N + \frac{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}}{\mu} + \frac{\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{C_0}{\varepsilon}\right)\right)$
$\mu > 0$ *and* $\nu = 0$	$O\left(\left(N + \frac{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}}{\mu}\right)\ln\left(\frac{C_0}{\varepsilon}\right) + (\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}})\sqrt{\frac{\sqrt{\mu_0/\nu_0}D_0}{\mu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu > 0$	$O\left(N\ln\left(\frac{C_0}{\varepsilon}\right) + \frac{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}\sqrt{\nu_0/\mu_0}D_0}{\varepsilon} + (\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}})\sqrt{\frac{\sqrt{\nu_0/\mu_0}D_0}{\nu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu = 0$	$O\left(N\ln\left(\frac{C_0}{\varepsilon}\right) + \frac{\left(\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}\sqrt{\nu_0/\mu_0} + (\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}})\right)D_0}{\varepsilon}\right)$

Proof. The exact case-by-case strategy of Theorem 3.3.1 can be applied (ignoring absolute constant factors) by setting $G \leftarrow \boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}$ and $L \leftarrow \boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}$. The only additional conditions that need to be incorporated are (3.57) and (3.58). Under the given assumption that

$\min_I r_I \geq 1/(2N)$ and $\min_I s_I \geq 1/(2N)$, these conditions will be satisfied when, for all $k \geq 2$, it holds that $\frac{a_k^2}{A_k\mu+\mu_0} \leq (1+1/(10N))\frac{a_{k-1}^2}{A_{k-1}\mu+\mu_0}$ and $\frac{a_k^2}{A_k\nu+\nu_0} \leq (1+1/(10N))\frac{a_{k-1}^2}{A_{k-1}\nu+\nu_0}$. Taking the first condition as an example, it can be rewritten as

$$\frac{a_k^2}{a_{k-1}^2} \leq \left(1+\frac{1}{10N}\right)\frac{A_k\mu+\mu_0}{A_{k-1}\mu+\mu_0}.$$

When $\mu = 0$ and $\nu = 0$, this condition is satisfied automatically as the ratio on the left-hand side is equal to 1 (because a_k is a constant sequence). Otherwise, because A_k is an increasing sequence, we can reduce the condition to $a_k^2 \leq (1+1/(10N))a_{k-1}^2$. The fastest growth condition on $(a_k)_{k\geq 1}$ that is possible under the constraint is $a_k \leq (1+\alpha)a_{k-1}$, where $(1+\alpha) \leq \sqrt{1+1/(10N)}$. Then,

$$\sqrt{1+1/(10N)} \geq \sqrt{1+69/(900N)} \geq 1+1/(30N),$$

so the imposition $\alpha \leq \frac{1}{30N}$ suffices. This adds an $O\left(N\ln\left(C_0/\varepsilon\right)\right)$ term to the resulting complexities and completes the proof. □

We provide similar computations as those used in Theorem 3.4.1 to uncover the dependence on the sampling scheme. We again use $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)$ where $\boldsymbol{\lambda}_I := \sqrt{\boldsymbol{G}_I^2 + \boldsymbol{L}_I^2}$ from Section 3.2. The non-uniform sampling complexity given below follows by letting $p_I \propto \boldsymbol{\lambda}_I^{1/2}$, $r_I \propto \boldsymbol{\lambda}_I^{1/2}$, $s_I \propto \boldsymbol{G}_I^{1/2}$ and $q_I \propto \boldsymbol{G}_I^{1/2}$.

Constant	**Uniform Sampling**	**Non-Uniform Sampling**
$\boldsymbol{G}_{\boldsymbol{p},\boldsymbol{s}} \vee \boldsymbol{G}_{\boldsymbol{q},\boldsymbol{r}}$	$N^{3/2}\left\|\boldsymbol{G}\right\|_2$	$\left\|\boldsymbol{\lambda}\right\|_{1/2}^{1/2}\left\|\boldsymbol{G}\right\|_{1/2}^{1/2}$
$\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}$	$N^{3/2}\left\|\boldsymbol{L}\right\|_2$	$\left\|\boldsymbol{\lambda}\right\|_{1/2}^{3/4}\left\|\boldsymbol{L}\right\|_{1/2}^{1/4}$

Notice that the complexities in Theorem 3.4.1 (as opposed to the ones shown in Theorem 3.4.2) depend on additional factors in N. Thus, taking $\mu > 0$ and $\nu > 0$ as an example, in the case of uniform sampling, the method of Theorem 3.4.1 has an $N^{3/2}(\left\|\boldsymbol{G}\right\|_\infty + \left\|\boldsymbol{L}\right\|_\infty)$ dependence on problem constants, whereas Theorem 3.4.2 has an $N^{3/2}(\left\|\boldsymbol{G}\right\|_2 + \left\|\boldsymbol{L}\right\|_2)$ depen-

dence. These are the same in the case of highly non-uniform Lipschitz constants, but have a $\sqrt{N}$ factor difference for approximately uniform Lipschitz constants.

This method inherits the exact per-iteration complexity as the one analyzed in Theorem 3.4.1. Maintaining the matrix-vector multiplication $\hat{\boldsymbol{g}}_k^\top \hat{\boldsymbol{y}}_k \in \mathbb{R}^d$ operates just as in (3.56). The total arithmetic complexity is given by $\tilde{O}(n(d+N)/N)$. The main difference in the algorithms from the upcoming Section 3.5, from a per-iteration complexity viewpoint, is that the full $\tilde{O}(n)$ cost update of $\boldsymbol{y}_k$ is replaced by a single block update of cost $\tilde{O}(n/N)$. Thus, we aim to improve the per-iteration complexity to $\tilde{O}(nd/N)$ in the separable case.

3.5 An Algorithm for Dual-Separable Problems

Given Definition 3.2.1, we design an algorithm variant that performs stochastic block-wise updates in the dual variable, akin to similar strategies applied to bilinearly coupled objectives [Song et al., 2021]. Precisely, we will only update a single randomly chosen block Q_k on each iteration k, in an effort to achieve an improved complexity guarantee. Updates in this form introduce additional technical challenges because different blocks of the dual iterate $\boldsymbol{y}_k = (\boldsymbol{y}_{k,1}, \ldots, \boldsymbol{y}_{k,N})$ have different dependences on the block index Q_k. As such, we carefully handle the expectations computations in the upcoming Lemma 3.5.1. Furthermore, a key difference in the proof structure of this section is that we will track an auxiliary sequence $(\bar{\boldsymbol{y}}_k)_{k\geq 1}$ of *return values*, such that the algorithm returns $(\boldsymbol{x}_t, \bar{\boldsymbol{y}}_t)$ in the final iteration instead of $(\boldsymbol{x}_t, \boldsymbol{y}_t)$. Conceptually, each block $\bar{\boldsymbol{y}}_{k,J}$ represents the J-th block of $\boldsymbol{y}_k$ if $J = Q_k$, or if block J was the one updated at time k. In other words, it stores all possible block updates that could have occurred from step $k-1$ to step k in one vector. Similar to before, we will define our update sequences in the process of deriving upper and lower bounds on $a_k \mathbb{E}[\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{v})]$ and $a_k \mathbb{E}[\mathcal{L}(\boldsymbol{u}, \bar{\boldsymbol{y}}_k)]$.

Crucially, we do not need to compute the elements of the sequence $(\bar{\boldsymbol{y}}_k)_{k\geq 1}$, as doing so would defeat the purpose of considering coordinate-wise updates. Instead, we may realize (3.10) in expectation by computing only one instance of $\bar{\boldsymbol{y}}_k$ with the following trick [Alacaoglu et al., 2022]: we randomly draw an index $\hat{t}$ (independent of all other randomness

in the algorithm and of $(\boldsymbol{u}, \boldsymbol{v})$) from $\{1, \ldots, t\}$, where $\hat{t} = k$ with probability a_k/A_t. Thus, by computing the conditional expectation over $\hat{t}$ given the sequence of iterates,

$$\mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_{\hat{t}}, \bar{\boldsymbol{y}}_{\hat{t}})|\mathcal{F}_t] = \sum_{k=1}^{t} a_k \,\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \bar{\boldsymbol{y}}_k).$$

In practice, we may simply run the algorithm to iteration $\hat{t}$. We use the following technical lemma to provide expectation formulas regarding $\boldsymbol{y}_k$ and $\bar{\boldsymbol{y}}_k$.

Lemma 3.5.1. *Let $h : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ be block separable in its first argument, i.e., $h(\boldsymbol{y}, \boldsymbol{x}) = \sum_{J=1}^{N} h_J(\boldsymbol{y}_J, \boldsymbol{x})$ for $J \in [N]$ and $\boldsymbol{y} \in \mathcal{Y}$. Assume that $\boldsymbol{y}_{k,J} = \boldsymbol{y}_{k-1,J}$ for all $J \neq Q_k$, that $\bar{\boldsymbol{y}}_{k,Q_k} = \boldsymbol{y}_{k,Q_k}$, and that $\bar{\boldsymbol{y}}_k$ is $\mathcal{F}_{k-1/2}$-measurable. Then, if Q_k is sampled uniformly on $[N]$, it holds that*

$$N\mathbb{E}\left[h_{Q_k}(\boldsymbol{y}_{k,Q_k}, \boldsymbol{x}_k)\right] = N\mathbb{E}\left[h(\boldsymbol{y}_k, \boldsymbol{x}_k)\right] - (N-1)\mathbb{E}\left[h(\boldsymbol{y}_{k-1}, \boldsymbol{x}_k)\right] = \mathbb{E}\left[h(\bar{\boldsymbol{y}}_k, \boldsymbol{x}_k)\right]. \tag{3.70}$$

Proof. Write

$$\begin{aligned}
\mathbb{E}_{k-1/2}\left[h_{Q_k}(\boldsymbol{y}_{k,Q_k}, \boldsymbol{x}_k)\right] &= \mathbb{E}_{k-1/2}\left[h(\boldsymbol{y}_k, \boldsymbol{x}_k)\right] - \mathbb{E}_{k-1/2}\left[\textstyle\sum_{J \neq Q_k} h_J(\boldsymbol{y}_{k,J}, \boldsymbol{x}_k)\right] \\
&= \mathbb{E}_{k-1/2}\left[h(\boldsymbol{y}_k, \boldsymbol{x}_k)\right] - \mathbb{E}_{k-1/2}\left[\textstyle\sum_{J \neq Q_k} h_J(\boldsymbol{y}_{k-1,J}, \boldsymbol{x}_k)\right] \\
&= \mathbb{E}_{k-1/2}\left[h(\boldsymbol{y}_k, \boldsymbol{x}_k)\right] - \frac{N-1}{N} h(\boldsymbol{y}_{k-1}, \boldsymbol{x}_k).
\end{aligned}$$

Take the marginal expectation of both terms to prove the first result. For the second,

$$\begin{aligned}
\mathbb{E}_{k-1/2}\left[h(\boldsymbol{y}_k, \boldsymbol{x}_k)\right] &= \frac{1}{N}\sum_{J=1}^{N} \mathbb{E}_{k-1/2}\left[h_J(\bar{\boldsymbol{y}}_{k,J}, \boldsymbol{x}_k)|J = Q_k\right] \\
&\quad + \frac{N-1}{N}\sum_{J=1}^{N} \mathbb{E}_{k-1/2}\left[h_J(\boldsymbol{y}_{k-1,J}, \boldsymbol{x}_k)|J \neq Q_k\right] \\
&= \frac{1}{N}\sum_{J=1}^{N} h_J(\bar{\boldsymbol{y}}_{k,J}, \boldsymbol{x}_k) + \frac{N-1}{N}\sum_{J=1}^{N} h_J(\boldsymbol{y}_{k-1,J}, \boldsymbol{x}_k) \\
&= \frac{1}{N} h(\bar{\boldsymbol{y}}_k, \boldsymbol{x}_k) + \frac{N-1}{N} h(\boldsymbol{y}_{k-1}, \boldsymbol{x}_k).
\end{aligned}$$

Rearrange terms and apply the marginal expectation to achieve the second result. □

The condition that $\bar{\boldsymbol{y}}_k$ is $\mathcal{F}_{k-1/2}$-measurable reflects the viewpoint that $\bar{\boldsymbol{y}}_k$ can be computed via a deterministic update given $\boldsymbol{x}_k$, after which $\boldsymbol{y}_k$ can be computed exactly from the random triple $(\boldsymbol{y}_{k-1}, \bar{\boldsymbol{y}}_k, Q_k)$.

We proceed to the details of the convergence analysis. The formula for $\bar{\boldsymbol{g}}_{k-1} \in \mathbb{R}^d$ is given after rigorously introducing the sequence $(\bar{\boldsymbol{y}}_k)_{k\geq 0}$, on which $\bar{\boldsymbol{g}}_{k-1}$ depends. We will first specify the upper bound and dual update. We define first the update for $\boldsymbol{y}_k$ (and by extension, the update for $\bar{\boldsymbol{y}}_k$). Because of separability, we may perform this update at $O(b)$ for $b := n/N$ cost on average across blocks. We do so by including only one additive component of ψ in the objective that $\boldsymbol{y}_k$ maximizes. We notice that the strong concavity constant of the objective defining $\boldsymbol{y}_k$ may change if we only use some components of ψ. To account for this, we design a slightly different proximity term from the one in Algorithm 4. Recall the definition of the Bregman divergences $\Delta_J(\cdot,\cdot)$ from Section 3.2. The upper bound is stated in expectation, as the results may not hold for all realizations (as was the case for previous versions of the initial gap bounds).

Lemma 3.5.2. *For all $k \geq 1$, consider the update*

$$\bar{\boldsymbol{y}}_k = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \left\{ a_k \left\langle \boldsymbol{y}, \bar{\boldsymbol{f}}_{k-1/2} \right\rangle - a_k \psi(\boldsymbol{y}) - \tfrac{A_{k-1}\nu+\nu_0}{2} \Delta_y(\boldsymbol{y}, \boldsymbol{y}_{k-1}) \right\}, \tag{3.71}$$

followed by setting $\boldsymbol{y}_{k,Q_k} = \bar{\boldsymbol{y}}_{k,Q_k}$ and $\boldsymbol{y}_{k,J} = \boldsymbol{y}_{k-1,J}$ for $J \neq Q_k$. Then, it holds that

$$\begin{aligned} a_k \mathbb{E}[\mathcal{L}(\boldsymbol{x}_k, \boldsymbol{v})] \leq a_k \mathbb{E}\left[\mathcal{L}(\boldsymbol{x}_k, \bar{\boldsymbol{y}}_k)\right] + \tfrac{N}{2}\mathbb{E}[\mathcal{T}_{k-1}^{\mathrm{D}}] - \tfrac{N}{2}\mathbb{E}[\mathcal{T}_k^{\mathrm{D}}] - \tfrac{N}{2}\mathbb{E}[\mathcal{C}_k^{\mathrm{D}}] \\ + a_k \mathbb{E}\left\langle \boldsymbol{v} - \bar{\boldsymbol{y}}_k, f(\boldsymbol{x}_k) - \bar{\boldsymbol{f}}_{k-1/2} \right\rangle. \end{aligned} \tag{3.72}$$

Proof. As in Lemma 3.3.2, write

$$\begin{aligned}
a_k\mathcal{L}(\boldsymbol{x}_k,\boldsymbol{v}) &= a_k\left\langle \boldsymbol{v},\bar{\boldsymbol{f}}_{k-1/2}\right\rangle - a_k\psi(\boldsymbol{v}) + a_k\phi(\boldsymbol{x}_k) - \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\boldsymbol{y}_{k-1})\\
&\quad + \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\boldsymbol{y}_{k-1}) + a_k\left\langle \boldsymbol{v}, f(\boldsymbol{x}_k) - \bar{\boldsymbol{f}}_{k-1/2}\right\rangle\\
&\le a_k\left\langle \bar{\boldsymbol{y}}_k,\bar{\boldsymbol{f}}_{k-1/2}\right\rangle - a_k\psi(\bar{\boldsymbol{y}}_k) + a_k\phi(\boldsymbol{x}_k) - \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\bar{\boldsymbol{y}}_k,\boldsymbol{y}_{k-1})\\
&\quad + \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\boldsymbol{y}_{k-1}) + a_k\left\langle \boldsymbol{v}, f(\boldsymbol{x}_k) - \bar{\boldsymbol{f}}_{k-1/2}\right\rangle - \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\bar{\boldsymbol{y}}_k)\\
&= a_k\mathcal{L}(\boldsymbol{x}_k,\bar{\boldsymbol{y}}_k) - \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\bar{\boldsymbol{y}}_k,\boldsymbol{y}_{k-1})\\
&\quad + \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\boldsymbol{y}_{k-1}) + a_k\left\langle \boldsymbol{v}-\bar{\boldsymbol{y}}_k, f(\boldsymbol{x}_k) - \bar{\boldsymbol{f}}_{k-1/2}\right\rangle - \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\bar{\boldsymbol{y}}_k),
\end{aligned}$$

where the inequality follows by Lemma 3.2.1 (dropping the non-positive term $\frac{a_k\nu}{2}\Delta_{\mathcal{Y}}(\boldsymbol{v},\bar{\boldsymbol{y}}_k)$). We take the marginal expectation and use Lemma 3.5.1 to achieve

$$\begin{aligned}
\tfrac{A_{k-1}\nu+\nu_0}{2}\mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{v},\boldsymbol{y}_{k-1})] - \tfrac{A_{k-1}\nu+\nu_0}{2}\mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{v},\bar{\boldsymbol{y}}_k)] &= \tfrac{N}{2}\mathbb{E}[\mathcal{T}^{\mathrm{D}}_{k-1}] - \tfrac{N}{2}\mathbb{E}[\mathcal{T}^{\mathrm{D}}_{k}]\\
\tfrac{A_{k-1}\nu+\nu_0}{2}\mathbb{E}[\Delta_{\mathcal{Y}}(\bar{\boldsymbol{y}}_k,\boldsymbol{y}_{k-1})] &= \tfrac{N}{2}\mathbb{E}[\mathcal{C}^{\mathrm{D}}_{k}],
\end{aligned}$$

completing the proof. □

Notice that we need only compute $\bar{\boldsymbol{y}}_{k,Q_k}$ in order to define $\boldsymbol{y}_k$. While the update (3.71) is written for the purpose of the proof, notice that for any $J \in [N]$, under Definition 3.2.1,

$$\bar{\boldsymbol{y}}_{k,J} = \operatorname*{arg\,max}_{\boldsymbol{y}_J\in\mathcal{Y}_J}\left\{a_k\left\langle \boldsymbol{y}_J,\bar{\boldsymbol{f}}_{k-1/2,J}\right\rangle - a_k\psi_J(\boldsymbol{y}_J) - \tfrac{A_{k-1}\nu+\nu_0}{2}\Delta_J(\boldsymbol{y}_J,\boldsymbol{y}_{k-1,J})\right\}.$$

Defining the primal update will reflect two different strategies—namely, those used in Section 3.4. For both cases, recall the table $\hat{\boldsymbol{x}}_{k,1},\dots,\hat{\boldsymbol{x}}_{k,N}$ introduced in (3.37) and (3.38). We will use the primal gradient estimate

$$\bar{\boldsymbol{g}}_{k-1} = \hat{\boldsymbol{g}}_{k-1}^{\top}\boldsymbol{y}_{k-1} + \frac{Na_{k-1}}{a_k}\sum_{i\in B_{P_k}}\left(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i}\right). \tag{3.73}$$

Notice that we do not need to use the table $\hat{\boldsymbol{y}}_k$ introduced in (3.39) from Section 3.4, as the matrix-vector product $\hat{\boldsymbol{g}}_{k-1}^{\top}\boldsymbol{y}_{k-1}$ above can be maintained in $O(bd)$ on average because only $b = n/N$ components of $\boldsymbol{y}_k$ change each iteration. When using these tables of iterates,

we will encounter the familiar terms $\hat{\mathcal{T}}^{\mathrm{P}}_{k,I} = (A_k\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{u}, \hat{\boldsymbol{x}}_{k,I})$ from (3.15) and $\hat{\mathcal{C}}^{\mathrm{P}}_{k,I} = (A_{k-1}\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}_k, \hat{\boldsymbol{x}}_{k-1,I})$ from (3.16) for $I = 1, \dots, N$.

With these elements in hand, we produce the following lower bound, which follows from identical steps to Lemma 3.3.1 and so has its proof omitted. Recall the use of probability weights $\gamma_1, \dots, \gamma_N$ in the update (3.11).

Lemma 3.5.3. *For any $k \geq 1$, let $\bar{\boldsymbol{g}}_{k-1} \in \mathbb{R}^d$ and $w^{\mathrm{P}}_k \in [0, 1)$, consider the update*

$$\boldsymbol{x}_k = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \Big\{ a_k \left\langle \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{x} \right\rangle + a_k \phi(\boldsymbol{x}) + \tfrac{1-w^{\mathrm{P}}_k}{2}(A_{k-1}\mu + \mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}_{k-1}) + \tfrac{w^{\mathrm{P}}_k}{2}(A_{k-1}\mu + \mu_0) \sum_{I=1}^{N} \gamma_I \Delta_{\mathcal{X}}(\boldsymbol{x}, \hat{\boldsymbol{x}}_{k-1,I}) \Big\}. \tag{3.74}$$

For $k \geq 1$, it holds that

$$a_k \mathcal{L}(\boldsymbol{u}, \bar{\boldsymbol{y}}_k) \geq a_k \mathcal{L}(\boldsymbol{x}_k, \bar{\boldsymbol{y}}_k) + a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \bar{\boldsymbol{y}}_k - \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \tag{3.75}$$

$$+ \left(\tfrac{1-w^{\mathrm{P}}_k}{2}\mathcal{T}^{\mathrm{P}}_k - \tfrac{1-w^{\mathrm{P}}_{k-1}}{2}\mathcal{T}^{\mathrm{P}}_{k-1} \right) + \tfrac{w^{\mathrm{P}}_k}{2} \left(\mathcal{T}^{\mathrm{P}}_k - \textstyle\sum_{I=1}^{N} \gamma_I \hat{\mathcal{T}}^{\mathrm{P}}_{k-1,I} \right) \tag{3.76}$$

$$+ \tfrac{1-w^{\mathrm{P}}_k}{2}\mathcal{C}^{\mathrm{P}}_k + \tfrac{w^{\mathrm{P}}_k}{2} \textstyle\sum_{I=1}^{N} \gamma_I \hat{\mathcal{C}}^{\mathrm{P}}_{k,I} + \tfrac{a_k\mu}{2}\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_k). \tag{3.77}$$

The only difference between Lemma 3.5.3 and Lemma 3.3.1 is the replacement of $\boldsymbol{y}_k$ by $\bar{\boldsymbol{y}}_k$, which is also an element of $\mathcal{Y}$. On the dual side, we set $\bar{\boldsymbol{f}}_{k-1/2} = f(\boldsymbol{x}_k)$, so the inner product terms in (3.72) vanish in expectation (in other words, $\mathcal{I}^{\mathrm{D}}_k$ can be thought of as zero). Following the same argument used to produce $\mathcal{E}^{\mathrm{P}}_k$ in (3.42), we may now build the identity card.

Identity Card 3: Stochastic update method for separable objectives

SUBROUTINE$_1$: $\hat{\boldsymbol{g}}_{k-1}^\top \boldsymbol{y}_{k-1} + \frac{Na_{k-1}}{a_k} \sum_{i \in B_{P_k}} (\bar{y}_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i})$

SUBROUTINE$_2$: $\bar{\boldsymbol{f}}_{k-1/2} = f(\boldsymbol{x}_k)$

SUBROUTINE$_3$: Update $\hat{\boldsymbol{x}}_{k,I}$ for all I using (3.38) and $(\hat{\boldsymbol{g}}_k, \hat{\boldsymbol{f}}_k)$ using (3.37).

Primal error: $\mathcal{E}_k^{\mathrm{P}} = a_{k-1} \left\langle \frac{1}{p_{P_k}} \sum_{i \in B_{P_k}} (\bar{y}_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_k - \boldsymbol{x}_{k-1} \right\rangle$

Dual error: $\mathcal{E}_k^{\mathrm{D}} = 0$

Using Lemma 3.5.2 and Lemma 3.5.3, we may produce a version of Claim 3.3.1:

$$\sum_{k=1}^{t} a_k \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k)] \leq \tfrac{1-w_k^{\mathrm{P}}}{2} \left(\mathcal{T}_0^{\mathrm{P}} - \mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] \right) + \tfrac{N}{2} \left(\mathcal{T}_0^{\mathrm{D}} - \mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] \right)$$

$$+ \sum_{k=1}^{t} \mathbb{E}\left[\mathcal{I}_k^{\mathrm{P}}\right] - \underbrace{\mathbb{E}\left[\tfrac{1-w_k^{\mathrm{P}}}{2} \mathcal{C}_k^{\mathrm{P}} + \tfrac{N}{2} \mathcal{C}_k^{\mathrm{D}} + \tfrac{w_k^{\mathrm{P}}}{2} \textstyle\sum_J \gamma_I \hat{\mathcal{C}}_{k,J}^{\mathrm{P}}\right]}_{\text{cancellation terms from Lemma 3.5.2 and Lemma 3.5.3}} \tag{3.78}$$

$$+ \underbrace{\tfrac{1}{2} \sum_{k=1}^{t} \mathbb{E}\left[w_k^{\mathrm{P}} \left(\textstyle\sum_{I=1}^{N} \gamma_I \hat{\mathcal{T}}_{k-1,I}^{\mathrm{P}} - \mathcal{T}_k^{\mathrm{P}} \right) - a_k \mu \Delta_{\mathfrak{X}}(\boldsymbol{u}, \boldsymbol{x}_k) \right]}_{\text{primal table terms from Lemma 3.5.3}}. \tag{3.79}$$

We proceed to the individual analyses to cancel the lines (3.78) and (3.79).

3.5.1 Strategy 1: Non-Uniform Historical Regularization

Here, we do not assume that $w_k^{\mathrm{P}} = 0$, and so will cancel both the lines (3.78) and (3.79). The resulting complexity will depend on the sampling probabilities $\boldsymbol{p}$ and regularization weights $\boldsymbol{\gamma}$ through the following constants:

$$\boldsymbol{G}_{\boldsymbol{p}} := \sqrt{\max_{I \in [N]} \frac{\boldsymbol{G}_I^2}{p_I}} \text{ and } \boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}} := \sqrt{\max_{I \in [N]} \frac{\boldsymbol{L}_I^2}{p_I \gamma_I}}.$$

The proof follows very similar steps to the proof of Proposition 3.4.1.

Proposition 3.5.1. *Let* $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathrm{ri}(\mathrm{dom}(\phi)) \times \mathrm{ri}(\mathrm{dom}(\psi))$ *and* $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k\geq 1}$ *be generated by the update from Lemma 3.5.3 with sequence* $w_k^{\mathrm{P}} \in [0,1)$ *and Lemma 3.5.2, with* $\bar{\boldsymbol{g}}_{k-1}$ *and* $\bar{\boldsymbol{f}}_{k-1/2}$ *given by (3.73), and* $f(\boldsymbol{x}_k)$*, respectively. Define* $a_1 = \min\left\{\frac{\sqrt{(1-w_0^{\mathrm{P}})\mu_0\nu_0}}{2\sqrt{2}\boldsymbol{G}_{\boldsymbol{p}}}, \frac{\sqrt{w_0^{\mathrm{P}}(1-w_0^{\mathrm{P}})}\mu_0}{4\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}}\right\}$ *and select* $(a_k)_{k\geq 2}$ *such that the conditions*

$$a_k \leq \min\left\{\frac{\sqrt{(1-w_k^{\mathrm{P}})(A_k\mu+\mu_0)(A_{k-1}\nu+\nu_0)}}{2\sqrt{2}\boldsymbol{G}_{\boldsymbol{p}}}, \frac{\sqrt{w_k^{\mathrm{P}}(1-w_k^{\mathrm{P}})(A_k\mu+\mu_0)(A_{k-1}\mu+\mu_0)}}{4\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}}\right\}, \tag{3.80}$$

are satisfied. In addition, impose that for any $\ell = 1, \ldots, t-1$*, it holds that*

$$\frac{\mu}{N}\sum_{k'=0}^{\infty} w_{\ell+k'+1}^{\mathrm{P}} A_{\ell+k'}(1-1/N)^{k'} \leq (w_\ell^{\mathrm{P}} A_\ell + a_\ell)\mu \tag{3.81}$$

and that $A_k \leq \left(1+\frac{1}{2N}\right)^k a_1$*. We have that for any* $\boldsymbol{u} \in \mathcal{X}$ *and* $\boldsymbol{v} \in \mathcal{Y}$*,*

$$\sum_{k=1}^{t} a_k \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \bar{\boldsymbol{y}}_k)] + \tfrac{1-w_k^{\mathrm{P}}}{4}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \tfrac{1}{4}\mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] \leq (a_1\mu+\mu_0)\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + \tfrac{1}{2}\mathcal{T}_0^{\mathrm{D}}.$$

Proof. Just as in the proof of Proposition 3.4.1, we aim to show that the sum of terms in (3.78) and (3.79) is a constant independent of t. These lines are exactly analogous to (3.24) and (3.25), and so we control them in a similar fashion.

1. Controlling (3.78): By Young's inequality with parameter $(1-w_k^{\mathrm{P}})(A_{k-1}\mu+\mu_0)/4$,

$$\begin{aligned}\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}] &= a_{k-1}\mathbb{E}\left\langle \tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right\rangle \\ &\leq \frac{1-w_k^{\mathrm{P}}}{4}\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{2a_{k-1}^2}{(1-w_k^{\mathrm{P}})(A_{k-1}\mu+\mu_0)}\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2\end{aligned} \tag{3.82}$$

and for the last term

$$\begin{aligned}&-\mathbb{E}[a_t\langle \nabla f(\boldsymbol{x}_t)^\top \bar{\boldsymbol{y}}_t - \hat{\boldsymbol{g}}_{t-1}^\top \boldsymbol{y}_{t-1}, \boldsymbol{u} - \boldsymbol{x}_t\rangle] \\ &\leq \frac{1-w_t^{\mathrm{P}}}{4}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \frac{2a_t^2}{(1-w_k^{\mathrm{P}})(A_t\mu+\mu_0)}\mathbb{E}\left\|\tfrac{1}{p_{P_t}}\textstyle\sum_{i\in B_{P_t}}(\bar{y}_{t,i}\nabla f_i(\boldsymbol{x}_t) - y_{t-1,i}\hat{\boldsymbol{g}}_{t-1,i})\right\|_{\mathcal{X}^*}^2\end{aligned} \tag{3.83}$$

For the second term in (3.82) and (3.83), apply Lemma 3.4.1 with $b_I = \gamma_I$ and $c_I = 1$ to achieve

$$\begin{aligned}
&\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2 \\
&\quad\le 2\left(\max_I \frac{\boldsymbol{L}_J^2}{p_J\gamma_J}\right)\cdot\sum_{I=1}^N \gamma_I\mathbb{E}\left\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-2,I}\right\|_{\mathcal{X}}^2 + 2\left(\max_J \frac{\boldsymbol{G}_J^2}{p_J}\right)\mathbb{E}\left\|\bar{\boldsymbol{y}}_{k-1}-\boldsymbol{y}_{k-2}\right\|_2^2 \\
&\quad\overset{(*)}{\le} 2\left(\max_I \frac{\boldsymbol{L}_J^2}{p_J\gamma_J}\right)\cdot\sum_{I=1}^N \gamma_I\mathbb{E}\left\|\boldsymbol{x}_{k-1}-\hat{\boldsymbol{x}}_{k-2,I}\right\|_{\mathcal{X}}^2 + 2N\left(\max_J \frac{\boldsymbol{G}_J^2}{p_J}\right)\mathbb{E}\left\|\boldsymbol{y}_{k-1}-\boldsymbol{y}_{k-2}\right\|_{\mathcal{Y}}^2 \\
&\quad\overset{(\circ)}{\le} 4\underbrace{\left(\max_J \frac{\boldsymbol{L}_J^2}{p_J\gamma_J}\right)}_{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}^2}\cdot\sum_{I=1}^N \gamma_I\mathbb{E}[\Delta_{\mathcal{X}}(\boldsymbol{x}_{k-1},\hat{\boldsymbol{x}}_{k-2,I})] + 4N\underbrace{\left(\max_J \frac{\boldsymbol{G}_J^2}{p_J}\right)}_{\boldsymbol{G}_{\boldsymbol{p}}^2}\mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{y}_{k-1},\boldsymbol{y}_{k-2})]
\end{aligned}$$

where in $(*)$ we applied Lemma 3.5.1 and $\|\cdot\|_2 \le \|\cdot\|_{\mathcal{Y}}$ to the second term and in $(\circ)$ we applied $\|\cdot\|_2 \le \|\cdot\|_{\mathcal{Y}}$ and the strong convexity of Bregman divergences.

Combining the steps above, recalling that $\hat{\mathcal{C}}_{k,I}^{\mathrm{P}} := (A_{k-1}\mu+\mu_0)\Delta_{\mathcal{X}}(\boldsymbol{x}_k,\hat{\boldsymbol{x}}_{k-1,I})$, and applying the condition (3.80), we have the upper bound

$$\begin{aligned}
\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}] &\le \frac{1-w_k^{\mathrm{P}}}{4}\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{4Na_{k-1}^2\boldsymbol{G}_{\boldsymbol{p}}^2\mathbb{E}[\mathcal{C}_{k-1}^{\mathrm{D}}]}{(1-w_k^{\mathrm{P}})(A_{k-1}\mu+\mu_0)(A_{k-2}\nu+\nu_0)} \\
&\quad + \frac{4a_{k-1}^2\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{\gamma}}^2\mathbb{E}\left[\sum_{I=1}^N\gamma_I\hat{\mathcal{C}}_{k-1,I}^{\mathrm{P}}\right]}{(1-w_k^{\mathrm{P}})(A_{k-1}\mu+\mu_0)(A_{k-2}\mu+\mu_0)} \\
&\le \frac{1-w_k^{\mathrm{P}}}{2}\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{N}{2}\mathbb{E}[\mathcal{C}_{k-1}^{\mathrm{D}}] + \frac{w_k^{\mathrm{P}}}{4}\mathbb{E}\left[\sum_{I=1}^N\gamma_I\hat{\mathcal{C}}_{k-1,I}^{\mathrm{P}}\right]
\end{aligned}$$

with a similar bound holding for $-\mathbb{E}[a_t\langle\nabla f(\boldsymbol{x}_t)^\top\bar{\boldsymbol{y}}_t - \hat{\boldsymbol{g}}_{t-1}^\top\boldsymbol{y}_{t-1}, \boldsymbol{u}-\boldsymbol{x}_t\rangle]$ except with $\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}]$ replaced by $\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}]$. These terms will cancel with the corresponding non-positive terms in (3.78).

2. Controlling (3.79): This follows from an identical argument to the one used to bound (3.25) in the proof of Proposition 3.4.1, appealing to the condition (3.81). □

The following result mirrors the logic of Theorem 3.4.1; the proof is omitted.

Theorem 3.5.1. *Under Assumption 3.2.2 and Assumption 3.2.1, consider any* $\boldsymbol{u}\in\mathcal{X}$, $\boldsymbol{v}\in\mathcal{Y}$

and precision $\varepsilon > 0$. *Define the initial distance constant*

$$D_{0,N} = \sqrt{\frac{\mu_0}{\nu_0}}\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0) + \sqrt{\frac{\nu_0}{\mu_0}}N\Delta_{\mathcal{Y}}(\boldsymbol{v}, \boldsymbol{y}_0) \tag{3.84}$$

There exists a choice of the sequence $(a_k)_{k=1}^t$ *and parameters* $(w_k^{\mathrm{P}})_{k=0}^t$ *such that Algorithm 4 with Identity Card 3 produces an output point* $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \in \mathcal{X} \times \mathcal{Y}$ *satisfying* $\mathbb{E}\left[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)\right] \leq \varepsilon$ *for* t *that depends on* ε *according to the following iteration complexities. They depend logarithmically on the constants* $C_1 := \sqrt{N}(\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + \boldsymbol{G}_{\boldsymbol{p}})D_{0,N} + \mu\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0)$, *and* $C_2 := (\sqrt{N}\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + \boldsymbol{G}_{\boldsymbol{p}})D_{0,N} + \mu\Delta_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{x}_0)$.

Case	Iteration Complexity
$\mu > 0$ *and* $\nu > 0$	$O\left(\left(N + \frac{\sqrt{N}\boldsymbol{L}_{\boldsymbol{p},\gamma}}{\mu} + \frac{\sqrt{N}\boldsymbol{G}_{\boldsymbol{p}}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{C_1}{\varepsilon}\right)\right)$
$\mu > 0$ *and* $\nu = 0$	$O\left(\left(N + \frac{\sqrt{N}\boldsymbol{L}_{\boldsymbol{p},\gamma}}{\mu}\right)\ln\left(\frac{C_2}{\varepsilon}\right) + \boldsymbol{G}_{\boldsymbol{p}}\sqrt{\frac{\sqrt{\mu_0/\nu_0}D_{0,N} + (a_1\mu/\nu_0)\Delta_{\mathcal{X}}(\boldsymbol{u},\boldsymbol{x}_0)}{\mu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu > 0$	$O\left(\frac{\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0}D_{0,N}}{\varepsilon} + N\boldsymbol{G}_{\boldsymbol{p}}\sqrt{\frac{\sqrt{\nu_0/\mu_0}D_{0,N}}{\nu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu = 0$	$O\left(\frac{\left(\boldsymbol{L}_{\boldsymbol{p},\gamma}\sqrt{\nu_0/\mu_0} + \boldsymbol{G}_{\boldsymbol{p}}\right)D_{0,N}}{\varepsilon}\right)$

In the discussions in Section 3.6, we set $\nu_0 \sim \mu_0/N$, so the $D_{0,N}$ term appearing in Theorem 3.5.1 is interpreted as $\sqrt{N}$ times a constant.

We recall from Section 3.2 that $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N)$ for $\boldsymbol{\lambda}_I := \sqrt{\boldsymbol{G}_I^2 + \boldsymbol{L}_I^2}$. The non-uniform sampling complexity given below follows by letting $p_I \propto \boldsymbol{\lambda}_I$ and $\gamma_I \propto \boldsymbol{L}_I$.

Constant	Uniform Sampling	Non-Uniform Sampling
$\boldsymbol{G}_{\boldsymbol{p}}$	$\sqrt{N}\left\|\boldsymbol{G}\right\|_\infty$	$\left\|\boldsymbol{\lambda}\right\|_1^{1/2}\left\|\boldsymbol{G}\right\|_\infty^{1/2}$
$\boldsymbol{L}_{\boldsymbol{p},\gamma}$	$N\left\|\boldsymbol{L}\right\|_\infty$	$\left\|\boldsymbol{\lambda}\right\|_1^{1/2}\left\|\boldsymbol{L}\right\|_1^{1/2}$

There are strict advantages both in the dependence on $\boldsymbol{G}$ and $\boldsymbol{L}$ in the separable case over, say, the complexities given in Theorem 3.4.1. This primarily results from the freedom to

select the constants $\gamma_1, \ldots, \gamma_N$ based only on the smoothness constants $\boldsymbol{L}$. Thus, in the case of non-uniform sampling, the dependencies from Theorem 3.4.1 (using the same historical regularization strategy) reduce in Theorem 3.5.1 from $\|\boldsymbol{\lambda}\|_1^{1/2} \|\boldsymbol{G}\|_1^{1/2}$ to $\|\boldsymbol{\lambda}\|_1^{1/2} \|\boldsymbol{G}\|_\infty^{1/2}$ and from $\|\boldsymbol{\lambda}\|_1$ to $\|\boldsymbol{\lambda}\|_1^{1/2} \|\boldsymbol{L}\|_1^{1/2}$.

Regarding per-iteration complexity, the exact arguments of Section 3.4.1 apply, except that the update defining $\boldsymbol{y}_k$ occurs at cost $\tilde{O}(n/N)$ instead of $\tilde{O}(n)$. Thus, the total per-iteration complexity is $\tilde{O}(nd/N)$.

3.5.2 Strategy 2: Non-Uniform Block Replacement Probabilities

Like its counterpart Section 3.4.2, this section will rely on choosing the parameter $\boldsymbol{r} = (r_1, \ldots, r_N)$, which governs the update probabilities of the primal table $\hat{\boldsymbol{x}}_{k,1}, \ldots, \hat{\boldsymbol{x}}_{k,N}$. Accordingly, we may set $w_k^{\mathrm{P}} = 0$ to simplify our gap bound to

$$\sum_{k=1}^{t} a_k \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k, \boldsymbol{y}_k)] \leq \tfrac{1}{2}\left(\mathcal{T}_0^{\mathrm{P}} - \mathbb{E}[\mathcal{T}_t^{\mathrm{P}}]\right) + \tfrac{N}{2}\left(\mathcal{T}_0^{\mathrm{D}} - \mathbb{E}[\mathcal{T}_t^{\mathrm{D}}]\right) + \sum_{k=1}^{t-1} \mathbb{E}\left[\mathcal{I}_k^{\mathrm{P}}\right] - \tfrac{1}{2}\sum_{k=1}^{t} \mathbb{E}\left[\mathcal{C}_k^{\mathrm{P}} + N\mathcal{C}_k^{\mathrm{D}}\right] \tag{3.85}$$

By this point, all arguments used in the analysis have been seen before, in that ideas related to separability were employed in Section 3.5.1 and ideas related to block replacement probabilities were employed in Section 3.4.2. Thus, the proofs are relatively short in this section. We cancel (3.85) in Proposition 3.5.2. The resulting complexity will depend on the sampling probabilities $\boldsymbol{p}$ and $\boldsymbol{r}$ through the constants

$$\boldsymbol{G}_{\boldsymbol{p}} := \sqrt{\max_{I \in [N]} \frac{\boldsymbol{G}_I^2}{p_I}} \text{ and } \boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}} := \sqrt{\sum_{I=1}^{N} \frac{\boldsymbol{L}_I^2}{p_I r_I^2}},$$

as described in Proposition 3.5.2.

Proposition 3.5.2. *Let* $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathrm{ri}(\mathrm{dom}(\phi)) \times \mathrm{ri}(\mathrm{dom}(\psi))$ *and* $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k \geq 1}$ *be generated using* $\bar{\boldsymbol{g}}_{k-1}$ *and* $\bar{\boldsymbol{f}}_{k-1/2}$ *given by* (3.73) *and* $f(\boldsymbol{x}_k)$*, respectively. Define* $a_1 = \min\left\{\frac{\sqrt{\mu_0 \nu_0}}{5\sqrt{2}\boldsymbol{G}_{\boldsymbol{p}}}, \frac{\mu_0}{10\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}}\right\}$

and select $(a_k)_{k\geq 2}$ *such that the conditions*

$$a_k \leq \min\left\{\frac{\sqrt{(A_k\mu+\mu_0)(A_{k-1}\nu+\nu_0)}}{5\sqrt{2}\boldsymbol{G_p}}, \frac{\sqrt{(A_k\mu+\mu_0)(A_{k-1}\mu+\mu_0)}}{10\boldsymbol{L_{p,r}}}\right\} \tag{3.86}$$

and $\frac{a_k^2}{A_k\mu+\mu_0} \leq \min_I(1+r_I/5)\frac{a_{k-1}^2}{A_{k-1}\mu+\mu_0}$ *hold. We have that for any* $\boldsymbol{u}\in\mathcal{X}$ *and* $\boldsymbol{v}\in\mathcal{Y}$,

$$\sum_{k=1}^{t} a_k\mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}_k,\boldsymbol{y}_k)] + \tfrac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \tfrac{N}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{D}}] \leq \tfrac{1}{2}\mathcal{T}_0^{\mathrm{P}} + \tfrac{N}{2}\mathcal{T}_0^{\mathrm{D}}.$$

Proof. Mirroring the proof of Proposition 3.4.2, by Young's inequality with parameter $(A_{k-1}\mu+\mu_0)/2$, we have for $k=2,\ldots,t-1$ that

$$\begin{aligned}\mathbb{E}[\mathcal{E}_k^{\mathrm{P}}] &= a_{k-1}\mathbb{E}\left\langle \tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_k - \boldsymbol{x}_{k-1}\right\rangle \\ &\leq \frac{1}{2}\mathbb{E}[\mathcal{C}_k^{\mathrm{P}}] + \frac{a_{k-1}^2}{A_{k-1}\mu+\mu_0}\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2 \qquad (3.87)\end{aligned}$$

$$(3.88)$$

and

$$\begin{aligned}&- a_k\mathbb{E}[\left\langle \nabla f(\boldsymbol{x}_k)^\top\bar{\boldsymbol{y}}_k - \hat{\boldsymbol{g}}_{k-1}^\top\boldsymbol{y}_{k-1}, \boldsymbol{u}-\boldsymbol{x}_k\right\rangle] \\ &\leq \frac{1}{2}\mathbb{E}[\mathcal{T}_t^{\mathrm{P}}] + \frac{a_{t-1}^2}{A_{t-1}\mu+\mu_0}\mathbb{E}\left\|\tfrac{1}{p_{P_t}}\textstyle\sum_{i\in B_{P_t}}(\bar{y}_{t-1,i}\nabla f_i(\boldsymbol{x}_{t-1}) - y_{t-2,i}\hat{\boldsymbol{g}}_{t-2,i})\right\|_{\mathcal{X}^*}^2. \qquad (3.89)\end{aligned}$$

Apply Lemma 3.4.1 to achieve

$$\begin{aligned}&\mathbb{E}\left\|\tfrac{1}{p_{P_k}}\textstyle\sum_{i\in B_{P_k}}(\bar{y}_{k-1,i}\nabla f_i(\boldsymbol{x}_{k-1}) - y_{k-2,i}\hat{\boldsymbol{g}}_{k-2,i})\right\|_{\mathcal{X}^*}^2 \\ &\leq 2\sum_{I=1}^{N}\frac{\boldsymbol{G}_I^2}{p_I}\mathbb{E}\left\|\bar{\boldsymbol{y}}_{k-1,I} - \boldsymbol{y}_{k-2,I}\right\|_2^2 + 2\sum_{I=1}^{N}\frac{\boldsymbol{L}_I^2}{p_I}\mathbb{E}\left\|\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-2,I}\right\|_{\mathcal{X}}^2,\end{aligned}$$

where the last line follows from Young's inequality. For the first term, we compute its

expectation using Lemma 3.5.1,

$$\begin{aligned} 2\sum_{I=1}^{N} \frac{\boldsymbol{G}_I^2}{p_I}\mathbb{E}\left\|\bar{\boldsymbol{y}}_{k-1,I} - \boldsymbol{y}_{k-2,I}\right\|_2^2 &\le 2\max_I \frac{\boldsymbol{G}_I^2}{p_I}\mathbb{E}\left\|\bar{\boldsymbol{y}}_{k-1} - \boldsymbol{y}_{k-2}\right\|_2^2 \\ &\le 2\max_I \frac{\boldsymbol{G}_I^2}{p_I}\mathbb{E}\left\|\bar{\boldsymbol{y}}_{k-1} - \boldsymbol{y}_{k-2}\right\|_{\mathcal{Y}}^2 \\ &= 2N\max_I \frac{\boldsymbol{G}_I^2}{p_I}\mathbb{E}\left\|\boldsymbol{y}_{k-1} - \boldsymbol{y}_{k-2}\right\|_{\mathcal{Y}}^2 \\ &\le 4N\underbrace{\max_I \frac{\boldsymbol{G}_I^2}{p_I}}_{\boldsymbol{G}_{\boldsymbol{p}}^2}\mathbb{E}[\Delta_{\mathcal{Y}}(\boldsymbol{y}_{k-1}, \boldsymbol{y}_{k-2})], \end{aligned}$$

where we used $\|\cdot\|_2 \le \|\cdot\|_{\mathcal{Y}}$ in the second inequality. For the second term, use the same argument leading to (3.66) in the proof of Proposition 3.4.2 to achieve

$$\begin{aligned} 2\sum_{I=1}^{N} \frac{\boldsymbol{L}_I^2}{p_I}\mathbb{E}\left\|\boldsymbol{x}_{k-1} - \hat{\boldsymbol{x}}_{k-2,I}\right\|_{\mathcal{X}}^2 &\le 50\left(\sum_{I=1}^{N}\frac{\boldsymbol{L}_I^2}{p_I r_I^2}\right)\sum_{k'=1}^{t}\frac{a_{k'}^2}{A_{k'}\mu + \mu_0}\mathbb{E}\left\|\boldsymbol{x}_{k'} - \boldsymbol{x}_{k'-1}\right\|_{\mathcal{X}}^2 \\ &\le 100\underbrace{\left(\sum_{I=1}^{N}\frac{\boldsymbol{L}_I^2}{p_I r_I^2}\right)}_{\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}^2}\sum_{k'=1}^{t}\frac{a_{k'}^2}{A_{k'}\mu + \mu_0}\mathbb{E}[\Delta_{\mathcal{X}}(\boldsymbol{x}_{k'}, \boldsymbol{x}_{k'-1})]. \end{aligned}$$

Thus, under the conditions (3.86), it holds that

$$\sum_{k=1}^{t-1}\mathbb{E}\left[\mathcal{I}_k^{\mathrm{P}}\right] \le \tfrac{1}{2}\sum_{k=1}^{t}\mathbb{E}\left[\mathcal{C}_k^{\mathrm{P}} + N\mathcal{C}_k^{\mathrm{D}}\right],$$

which completes the proof. □

As in Section 3.5.1, because the following result follows the same argument as Theorem 3.4.2, the proof is omitted. We use the initial distance quantity $D_{0,N}$ from Theorem 3.5.1.

Theorem 3.5.2. *Under Assumption 3.2.2 and Assumption 3.2.1, consider any $\boldsymbol{u} \in \mathcal{X}$, $\boldsymbol{v} \in \mathcal{Y}$ and precision $\varepsilon > 0$. There exists a choice of the sequences $(a_k)_{k=1}^t$ and $(w_k^{\mathrm{P}})_{k=1}^t$ such that Algorithm 4 with Identity Card 3 produces an output point $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\mathbb{E}\left[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)\right] \le \varepsilon$ for t that depends on ε according to the following iteration complexities. Let $C_{0,N} := (\boldsymbol{L}_{\boldsymbol{p},\boldsymbol{r}}\sqrt{\nu_0/\mu_0} + \boldsymbol{G}_{\boldsymbol{p}})D_{0,N}$, for $D_{0,N}$ defined in (3.84).*

Case	Iteration Complexity
$\mu > 0$ *and* $\nu > 0$	$O\left(\left(N + \frac{\boldsymbol{L_{p,r}}}{\mu} + \frac{\boldsymbol{G_p}}{\sqrt{\mu\nu}}\right) \ln\left(\frac{C_{0,N}}{\varepsilon}\right)\right)$
$\mu > 0$ *and* $\nu = 0$	$O\left(\left(N + \frac{\boldsymbol{L_{p,r}}}{\mu}\right) \ln\left(\frac{C_{0,N}}{\varepsilon}\right) + \boldsymbol{G_p}\sqrt{\frac{\sqrt{\mu_0/\nu_0}D_{0,N}}{\mu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu > 0$	$O\left(N \ln\left(\frac{C_{0,N}}{\varepsilon}\right) + \frac{\boldsymbol{L_{p,r}}\sqrt{\nu_0/\mu_0}D_{0,N}}{\varepsilon} + \boldsymbol{G_p}\sqrt{\frac{\sqrt{\nu_0/\mu_0}D_{0,N}}{\nu\varepsilon}}\right)$
$\mu = 0$ *and* $\nu = 0$	$O\left(N \ln\left(\frac{C_{0,N}}{\varepsilon}\right) + \frac{\left(\boldsymbol{L_{p,r}}\sqrt{\nu_0/\mu_0} + \boldsymbol{G_p}\right)D_{0,N}}{\varepsilon}\right)$

As with Theorem 3.5.1, by setting $\nu_0 \sim \mu_0/N$, we interpret $D_{0,N}$ as $\sqrt{N}$ times a constant. The non-uniform sampling complexity given below follows by letting $p_I \propto \boldsymbol{\lambda}_I^{1/2}$ and $r_I \propto \boldsymbol{L}_I^{1/2}$.

Constant	Uniform Sampling	Non-Uniform Sampling
$\boldsymbol{G_p}$	$\sqrt{N}\,\|\boldsymbol{G}\|_\infty$	$\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\,\|\boldsymbol{G}\|_\infty^{3/4}$
$\boldsymbol{L_{p,r}}$	$N^{3/2}\,\|\boldsymbol{L}\|_2$	$\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\,\|\boldsymbol{L}\|_{1/2}^{3/4}$

We compare the resulting complexity to the non-separable analog in Theorem 3.4.2. There are improvements both in the dependence on $\boldsymbol{G}$ and $\boldsymbol{L}$ when the sampling scheme can be tuned. Theorem 3.5.2 improves the dependence on $\boldsymbol{G}$ from $\|\boldsymbol{\lambda}\|_{1/2}^{1/2}\|\boldsymbol{G}\|_\infty^{1/4}$ to $\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\|\boldsymbol{G}\|_\infty^{3/4}$. As for the dependence on $\boldsymbol{L}$, this improves from $\|\boldsymbol{\lambda}\|_{1/2}^{3/4}\|\boldsymbol{L}\|_{1/2}^{1/4}$ to $\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\|\boldsymbol{L}\|_{1/2}^{3/4}$. The mechanism is analogous to the improvement from Theorem 3.4.1 to Theorem 3.5.1; because $\mathcal{I}_k^{\mathrm{D}} = 0$, the constants do not have to adapt in order to control an error term of the form $\mathcal{E}_k^{\mathrm{D}}$.

Finally, on per-iteration complexity, the exact arguments of Section 3.4.2 apply, except that the update defining $\boldsymbol{y}_k$ may now occur at cost $\tilde{O}(n/N)$ instead of $\tilde{O}(n)$. Thus, the total per-iteration complexity is $\tilde{O}(nd/N)$.

3.6 Discussion & Comparisons

Our discussion covers internal comparisons between full vector methods and stochastic methods, as well as external comparisons to contemporary methods for solving saddle point and

variational inequality problems. We briefly comment on our chosen convergence criterion, as it may differ slightly from those used in comparisons.

3.6.1 Stronger Convergence Criteria

The results discussed in this section are expressed in terms of *global complexity* or *arithmetic complexity*, which is computed by multiplying the number of iterations shown in Theorem 3.3.1 to 3.5.2 by $\tilde{O}(nd)$ for full vector update methods; for block-wise methods with N blocks of size n/N, we use $\tilde{O}(n(d/N+1))$ for full updates of $\boldsymbol{y}_k$ and $\tilde{O}(n(d/N))$ for partial updates of $\boldsymbol{y}_k$. In order to effectively compare to methods using our block coordinate-wise Lipschitz and smoothness constant from Section 3.2, we will apply a particular finite sum decomposition (see (3.93)) that will allow us to directly apply methods from the finite sum variational inequality literature.

Importantly, in the case of randomized algorithms, the complexity in terms of the number of iterations is itself determined by the number t such that the algorithm may output a point $(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)$ satisfying $\mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)] \leq \varepsilon$, where the expectation is taken over all algorithm randomness with $\boldsymbol{u} \in \mathcal{X}$ and $\boldsymbol{v} \in \mathcal{Y}$ fixed (i.e., $(\boldsymbol{u}, \boldsymbol{v})$ is independent of the algorithm randomness). When $\mu > 0$, the criterion is made meaningful by setting $\boldsymbol{u} = \boldsymbol{x}_\star$ as the unique minimizer of the strongly convex objective $\boldsymbol{x} \mapsto \max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. Otherwise, we choose a compact set $\mathcal{U} \subseteq \mathcal{X}$ and consider $\sup_{\boldsymbol{u} \in \mathcal{X}} \mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\tilde{\boldsymbol{x}}_t, \tilde{\boldsymbol{y}}_t)] \leq \varepsilon$, where $\boldsymbol{v}$ is replaced by the unique maximizer $\boldsymbol{y}_\star$ when $\nu > 0$ or another supremum is taken over $\boldsymbol{v} \in \mathcal{V} \subseteq \mathcal{Y}$ for $\mathcal{V}$ compact otherwise. As described in Alacaoglu et al. [2022, Example 1], the "supremum of expected gap" criterion is weaker than the "expected supremum of gap" criterion, as algorithms with divergent behavior can still converge according to the first criterion. We render guarantees for the stronger criterion as a technical detail, as in light of previous work, largely similar steps can be applied to achieve the same complexity guarantee for the expected supremum of gap. To not overcomplicate the proofs, we only highlight the parts of the analysis that change. Consider the argument used to derive (3.41), in which the expectation is applied to the terms $a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle$, after which they telescope. If the supremum

is to be taken, we can no longer apply the expectation to these terms directly. Instead,

$$\begin{aligned} a_k\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k\rangle &= a_k \left\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k \right\rangle \\ &\quad - \langle \underbrace{\tfrac{a_{k-1}}{p_{P_k}} \textstyle\sum_{i\in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i})}_{\boldsymbol{z}_k}, \boldsymbol{u} - \boldsymbol{x}_{k-1}\rangle \\ &\quad - \underbrace{a_{k-1} \left\langle \tfrac{1}{p_{P_k}} \textstyle\sum_{i\in B_{P_k}} (y_{k-1,i} \nabla f_i(\boldsymbol{x}_{k-1}) - \hat{y}_{k-2,i} \hat{\boldsymbol{g}}_{k-2,i}), \boldsymbol{x}_{k-1} - \boldsymbol{x}_k \right\rangle}_{\mathcal{E}_k^{\mathrm{P}}}, \end{aligned}$$

and by taking the expectation over P_k yields the identity

$$\begin{aligned} a_k\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \bar{\boldsymbol{g}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k\rangle &= a_k\langle \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k - \hat{\boldsymbol{g}}_{k-1}^\top \hat{\boldsymbol{y}}_{k-1}, \boldsymbol{u} - \boldsymbol{x}_k\rangle \\ &\quad - a_{k-1}\langle \nabla f(\boldsymbol{x}_{k-1})^\top \boldsymbol{y}_{k-1} - \hat{\boldsymbol{g}}_{k-2}^\top \hat{\boldsymbol{y}}_{k-2}, \boldsymbol{u} - \boldsymbol{x}_{k-1}\rangle \\ &\quad - \langle \boldsymbol{z}_k - \mathbb{E}_{k-1}[\boldsymbol{z}_k], \boldsymbol{u}\rangle - \mathcal{E}_k^{\mathrm{P}}. \end{aligned} \tag{3.90}$$

The familiar term $\mathcal{E}_k^{\mathrm{P}}$ does not depend on $\boldsymbol{u}$ and can be bounded using the same techniques as in the proofs of Theorem 3.4.1 and Theorem 3.4.2, whereas $\boldsymbol{z}_k - \mathbb{E}_{k-1}[\boldsymbol{z}_k]$ is zero-mean conditional on $\mathcal{F}_{k-1}$ (i.e., a martingale difference sequence), but is not necessarily independent of $\boldsymbol{u}$. Next, we may apply Diakonikolas [2025, Lemma 4] (adapted from Alacaoglu and Malitsky [2022, Lemma 3.5]) to achieve

$$\begin{aligned} -\mathbb{E}\left[\sum_{k=1}^{t} \langle \boldsymbol{z}_k - \mathbb{E}_{k-1}[\boldsymbol{z}_k], \boldsymbol{u}_\star\rangle\right] &\le \mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}_\star, \boldsymbol{x}_0)] + \frac{1}{2}\sum_{k=1}^{t} \mathbb{E}\left\|\boldsymbol{z}_k - \mathbb{E}_{k-1}[\boldsymbol{z}_k]\right\|_2^2 \\ &\le \mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}_\star, \boldsymbol{x}_0)] + \frac{1}{2}\sum_{k=1}^{t} \mathbb{E}\left\|\boldsymbol{z}_k\right\|_2^2, \end{aligned}$$

where $\boldsymbol{u}_\star$ is the element of $\boldsymbol{u} \in \mathcal{U}$ that achieves the supremum in the gap criterion (recalling that $\mathcal{U}$ is chosen to be compact). The term $\mathbb{E}[\Delta_{\mathfrak{X}}(\boldsymbol{u}_\star, \boldsymbol{x}_0)]$ is upper bounded by a constant, whereas the $\mathbb{E}\left\|\boldsymbol{z}_k\right\|_2^2$ terms are bounded using the exact same techniques used to bound the $\mathcal{E}_k^{\mathrm{P}}$ terms. We choose to describe the argument in the manner above (as opposed to including it formally in the proofs) as it changes neither the other technical ideas nor the resulting complexities; we comment on this subtlety for the sake of completeness.

Algorithm Type	Global Complexity (big-O)
Full vector (Theorem 3.3.1)	$nd\left(\frac{L}{\mu}+\frac{G}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Stochastic (Theorem 3.4.1)	$\left(\frac{nd}{N}+n\right)\left(N+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1}{\mu}+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{G}\|_1^{1/2}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Stochastic (Theorem 3.4.2)	$\left(\frac{nd}{N}+n\right)\left(N+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{3/4}\|\boldsymbol{L}\|_{1/2}^{1/4}}{\mu}+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{1/2}\|\boldsymbol{G}\|_{1/2}^{1/2}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Block Coordinate-wise (Theorem 3.5.1)	$\frac{nd}{N}\left(N+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{L}\|_1^{1/2}}{\mu}+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{G}\|_\infty^{1/2}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Block Coordinate-wise (Theorem 3.5.2)	$\frac{nd}{N}\left(N+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\|\boldsymbol{L}\|_{1/2}^{3/4}}{\mu}+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\|\boldsymbol{G}\|_\infty^{3/4}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$

Table 3.1: **Complexity Bounds for Full Vector and Stochastic Methods for the case $\mu, \nu > 0$.** Arithmetic or global complexity (i.e., the total number of elementary operations required to compute $(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\mathbb{E}\left[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon\right]$ for fixed $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{X} \times \mathcal{Y}$, with the expectation taken over all algorithmic randomness.

3.6.2 Full Vector Update versus Stochastic Methods

To compare methods both within this chapter and alternatives for solving nonbilinearly coupled min-max problems, we first state some relationships between the constants introduced in Assumption 3.2.2. We then proceed with fine-grained comparisons to alternatives in the existing literature on min-max optimization and monotone variational inequalities. To simplify some comparisons, we assume that $L/\mu \geq 1$ and $G/\sqrt{\mu\nu} \geq 1$, so that they may be interpreted as "primal" and "mixed" condition numbers, respectively. First, observe that by the triangle inequality, in terms of the constants $\boldsymbol{G} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_N)$ and $\boldsymbol{L} = (\boldsymbol{L}_1, \ldots, \boldsymbol{L}_N)$, the constants G and L can be upper bounded as

$$G \leq \|\boldsymbol{G}\|_1 \text{ and } L \leq \|\boldsymbol{L}\|_1 .$$

Comparing the complexities for $\mu, \nu > 0$ in Table 3.1, the findings are summarized as follows.

Remark 3.6.1. *We observe the following in Table 3.1.*

- *The strongly convex case highlights a limitation in the historical regularization method*

(Theorem 3.4.1 and Theorem 3.5.1) as a factor $\sqrt{N}$ is gained in terms of the dependence on $\boldsymbol{G}$ and $\boldsymbol{L}$, which is not observed for the full vector method from Theorem 3.3.1. When comparing Theorem 3.4.1 to Theorem 3.3.1, the dependence on the smoothness constants renders as L versus $\|\boldsymbol{\lambda}\|_1/\sqrt{N}$ and on the Lipschitz constants as G versus $\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{G}\|_1^{1/2}/\sqrt{N}$. In both cases, when the constants are highly non-uniform, we are still afforded an up to $\sqrt{n}$ improvement in complexity from the stochastic method. Note that Theorem 3.5.1 is a strict improvement over Theorem 3.4.1 due to separability.

- *On the other hand, in the highly non-uniform setting, we may gain an up to d factor (resp. n factor) of improvement in terms of complexity using the methods of Theorem 3.4.2 (resp. Theorem 3.5.2) over the full vector method. The results of Theorem 3.4.1 and Theorem 3.4.2 (and by analogy, the results of Theorem 3.5.1 and Theorem 3.5.2) do not have a uniformly dominating method, as the extra factor of $\sqrt{N}$ may be on par with the improvement of the $\|\cdot\|_1$ norm over the $\|\cdot\|_{1/2}$ in terms of the dependence on $(\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\lambda})$.*

3.6.3 Alternative Methods for Min-Max and Variational Inequality Problems

In some comparisons, we must access the smoothness constants of the function $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{y})$ both with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ separately, and additionally, with respect to the pair $(\boldsymbol{x}, \boldsymbol{y})$. For the sake of comparison, assume that ϕ and ψ are differentiable and

$$\|\nabla\phi(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x}')\|_{\mathcal{X}^*} \leq \mu\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}} \text{ and } \|\nabla\psi(\boldsymbol{y}) - \nabla\psi(\boldsymbol{y}')\|_{\mathcal{Y}^*} \leq \nu\|\boldsymbol{y} - \boldsymbol{y}'\|_{\mathcal{Y}}.$$

Then, the following relations hold:

$$\begin{aligned}
\sup_{\boldsymbol{y}\in\mathcal{Y}} \|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{x}'}\mathcal{L}(\boldsymbol{x}',\boldsymbol{y})\|_{\mathcal{X}^*} &\leq (L+\mu)\,\|\boldsymbol{x}-\boldsymbol{x}'\|_{\mathcal{X}}\,, \\
\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla_{\boldsymbol{y}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{y}'}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}')\|_{\mathcal{X}^*} &\leq \nu\,\|\boldsymbol{y}-\boldsymbol{y}'\|_{\mathcal{Y}}\,, \\
\sup_{\boldsymbol{y}\in\mathcal{Y}} \|\nabla_{\boldsymbol{y}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{y}}\mathcal{L}(\boldsymbol{x}',\boldsymbol{y})\|_{\mathcal{X}^*} &\leq G\,\|\boldsymbol{x}-\boldsymbol{x}'\|_{\mathcal{X}}\,, \\
\sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y}')\|_{\mathcal{X}^*} &\leq G\,\|\boldsymbol{y}-\boldsymbol{y}\|_{\mathcal{Y}}\,.
\end{aligned} \tag{3.91}$$

As is done in the case of the ℓ_2-norm, define the norms

$$\|(\boldsymbol{x},\boldsymbol{y})\|^2 := \|\boldsymbol{x}\|_{\mathcal{X}}^2 + \|\boldsymbol{y}\|_{\mathcal{Y}}^2 \;\text{ and }\; \|\nabla\mathcal{L}(\boldsymbol{x},\boldsymbol{y})\|_*^2 := \|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y})\|_{\mathcal{X}^*}^2 + \|\nabla_{\boldsymbol{y}}\mathcal{L}(\boldsymbol{x},\boldsymbol{y})\|_{\mathcal{Y}^*}^2\,.$$

Then, we finally have that

$$\|\nabla\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla\mathcal{L}(\boldsymbol{x}',\boldsymbol{y}')\|_* \leq \sqrt{3\max\left\{L^2+G^2+\mu^2, G^2+\nu^2\right\}}\,\|(\boldsymbol{x},\boldsymbol{y}) - (\boldsymbol{x}',\boldsymbol{y}')\|\,. \tag{3.92}$$

Remark 3.6.2. *We do* not *assume that ϕ and ψ are smooth in our convergence guarantees, as the individual components ϕ and ψ are allowed to be non-differentiable in our framework. The additional assumptions above are made only for the sake of comparison. To this end, we will assume that $\nu \leq G$ (and maintain $\mu \leq L$). Then the constants in (3.91) and (3.92) can be simplified to absolute constants times L and $\sqrt{L^2+G^2}$, respectively. Thus, we define the vector field $(\boldsymbol{x},\boldsymbol{y}) \mapsto \nabla\mathcal{L}(\boldsymbol{x},\boldsymbol{y})$, which is $C\sqrt{L^2+G^2}$-Lipschitz (for an absolute constant $C>0$) and $(\mu\wedge\nu)$-strongly monotone.*

Given the above, we use these constants to compare to algorithms designed for solving variational inequality problems. In the upcoming remarks, we discuss the convex-concave (monotone) and strongly convex-strongly concave (strongly monotone) settings. First, consider the case in which $\mu\wedge\nu > 0$, so that we may observe the dependence on all problem constants in the full vector update settings.

Remark 3.6.3. *We observe the following in Table 3.2.*

- *When using classical methods for monotone variational inequalities, such as dual extrapolation [Nesterov and Scrimali, 2006], we highlight the $(\mu \wedge \nu)$ term, which may suffer when there is a large asymmetry in the strong convexity and strong concavity constants. As emphasized in Section 3.1, ν is often chosen as a small approximation or smoothing parameter, meaning that the theoretical complexity will not necessarily improve for large values of the primal strong convexity constant μ. Observe that all other methods will improve with increasing μ.*

- *In recent works [Jin et al., 2022, Li et al., 2023] and ours, the Lipschitz constants from different components of the objective function are separated in the complexity. In particular, the constants L and ν are decoupled, which is especially helpful in scenarios in which L is much larger than G. Examples include losses that are themselves smooth approximations of non-smooth losses, such as Huber approximations of the mean absolute error function. We improve over Jin et al. [2022], Li et al. [2023], which are designed for general nonbilinearly-coupled objectives, by a logarithmic factor in iteration complexity and achieve the same result in global complexity.*

When viewed as a saddle point or variational inequality problem, notice that (3.2) has a finite sum structure. In order to make direct comparisons, we decompose the objective block-wise, that is, $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{J=1}^{N} \mathcal{L}_J(\boldsymbol{x}, \boldsymbol{y})$, where

$$\mathcal{L}_J(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} \sum_{j \in B_J} y_j f_j(\boldsymbol{x}) - \psi_J(\boldsymbol{y}_J) + \frac{1}{N}\phi(\boldsymbol{x}) & \text{if Definition 3.2.1 (separability) is satisfied} \\ \sum_{j \in B_J} y_j f_j(\boldsymbol{x}) - \frac{1}{N}\psi(\boldsymbol{y}) + \frac{1}{N}\phi(\boldsymbol{x}) & \text{otherwise} \end{cases}. \tag{3.93}$$

Thus, when comparing to methods designed for finite sum objectives, we may consider the overall complexity of querying the oracle $(\mathcal{L}_J, \nabla \mathcal{L}_J)$ to be $O(nd/N)$ if the objective is separable or $O(n(d/N + 1))$ if it is non-separable. Even if $(\mathcal{L}_J, \nabla \mathcal{L}_J)$ is of cost $O(nd/N)$ to

compute, each oracle call may be associated with a single step of the algorithm, which is still $O(n)$ if it updates the primal and dual variables in their entirety. This is the first consideration when computing the global complexities in Table 3.3 and Table 3.4. For the second consideration, we compute the Lipschitzness and smoothness assumption of the individual component functions *on average* with uniform or non-uniform sampling, as is commonly used in analyses of methods for sum-decomposable objectives. Many contemporary results are stated in terms of "on average" smoothness (for min-max problems) and Lipschitzness (for variational inequality problems). Using the same norms defined in (3.92), we say that $\mathcal{L}_1, \dots, \mathcal{L}_N$ are L_{avg}-*smooth on average* according to sampling weights $\boldsymbol{p} = (p_1, \dots, p_N)$ if

$$\mathbb{E}_{J\sim\boldsymbol{p}} \left\| (1/p_J) \left(\nabla \mathcal{L}_J(\boldsymbol{x}, \boldsymbol{y}) - \nabla \mathcal{L}_J(\boldsymbol{x}', \boldsymbol{y}') \right) \right\|_*^2 \leq L_{\text{avg}}^2 \left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}', \boldsymbol{y}') \right\|^2 .$$

Recall the constants $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_N)$, where $\boldsymbol{\lambda}_J = \sqrt{\boldsymbol{G}_J^2 + \boldsymbol{L}_J^2}$ are the Lipschitz constants of each $\mathcal{L}_J$ with respect to the norm $\|\cdot\|$ defined above. The prototypical sampling schemes are the uniform and importance-weighted schemes

$$\mathbb{E}_{J\sim\text{unif}[N]} \left\| n \left(\nabla \mathcal{L}_J(\boldsymbol{x}, \boldsymbol{y}) - \nabla \mathcal{L}_J(\boldsymbol{x}', \boldsymbol{y}') \right) \right\|_*^2 \leq \underbrace{N \left\| \boldsymbol{\lambda} \right\|_2^2}_{\boldsymbol{\lambda}_{\text{unif}}^2} \left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}', \boldsymbol{y}') \right\|^2$$

$$\mathbb{E}_{J\sim\boldsymbol{\lambda}} \left\| (1/\boldsymbol{\lambda}_J) \left(\nabla \mathcal{L}_J(\boldsymbol{x}, \boldsymbol{y}) - \nabla \mathcal{L}_J(\boldsymbol{x}', \boldsymbol{y}') \right) \right\|_*^2 \leq \underbrace{\left\| \boldsymbol{\lambda} \right\|_1^2}_{\boldsymbol{\lambda}_{\text{imp}}^2} \left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}', \boldsymbol{y}') \right\|^2 .$$

Note that under the same sampling scheme, L_{avg}-smoothness on average is also implied by $(1/L_{\text{avg}})$-cocoercivity on average [Cai et al., 2024, Assumption 3]. Finally, for $\varepsilon \searrow 0$, we have that $O\left(N + \sqrt{N}\boldsymbol{\lambda}_{\text{unif}}/\varepsilon\right) = O\left(\sqrt{N}\boldsymbol{\lambda}_{\text{unif}}/\varepsilon\right)$ and $O\left(N + \sqrt{N}\boldsymbol{\lambda}_{\text{imp}}/\varepsilon\right) = O\left(\sqrt{N}\boldsymbol{\lambda}_{\text{imp}}/\varepsilon\right)$ which, combined with the oracle cost for (3.93) , leads to the results in Table 3.3 and Table 3.4.

Remark 3.6.4. *We observe the following in Table 3.3 and Table 3.4.*

- *For the improved dependence on the problem constants* $(\boldsymbol{G}, \boldsymbol{L})$ *and strong convexity constants* (μ, ν) *there are two main themes, which both involve "decoupling" of the two*

constants. Indeed, all results except for ours depend only on the aggregate Lipschitz constants $\boldsymbol{\lambda}$ and the minimum of the strong convexity constants $(\mu \wedge \nu)$, whereas we may (partially) separate these into dependences on "$\boldsymbol{L}$ over μ" terms and "$\boldsymbol{G}$ over $\sqrt{\mu\nu}$" terms.

- *In Table 3.3, fixing (n, d, ε) and without considering the differences between the aggregate constants $\boldsymbol{\lambda}$ and decoupled constants $(\boldsymbol{G}, \boldsymbol{L})$, we notice a dependence on the ℓ_2-norm in the results of Alacaoglu and Malitsky [2022], Cai et al. [2024], and Pichugin et al. [2024], as opposed to the $(1/N)$ times the ℓ_1-norm dependence in Theorem 3.4.1. Because $N^{-1}\|\cdot\|_1 \leq N^{-1/2}\|\cdot\|_2$, we observe a $\sqrt{N}$ improvement in complexity. Comparing the analogous results in Table 3.4, the complexity result of Theorem 3.4.1 scales as $N^{-1/2}\|\cdot\|_1$ improves over, but may still be on par with, the $\|\cdot\|_2$ scaling of Alacaoglu and Malitsky [2022] and Cai et al. [2024].*

- *In comparison to the best result of Diakonikolas [2025] in Table 3.3 for the non-separable case, the improvement of Theorem 3.4.1 comes from the use of the ℓ_1-norm over the $\ell_{1/2}$-norm, which may be up to $\sqrt{N}$ smaller.*

We also mention the work of Boob and Khalafi [2024], which solves a functionally constrained variational inequality formulation akin to Example 3 from Section 3.1. They operate under a completely different set of assumptions, largely to handle the possible unboundedness of the domain $\mathcal{Y}$ of the Lagrange multipliers. Furthermore, the gradient operator and the functional constraints satisfy non-standard deviation control (as opposed to Lipschitzness) inequalities (see Boob and Khalafi [2024, Eq. (1.3) and (1.4)], and thus this work is not directly comparable to ours.

3.6.4 Bilinearly Coupled Problems

While originally motivated by nonbilinearly-coupled min-max problems, the bilinearly-coupled setting constitutes an important special case of (3.2), defined via $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$ for $\mathbf{A} \in \mathbb{R}^{n\times d}$.

For the sake of discussion, we consider the convex-concave case (which includes matrix games for which $\phi \equiv 0$ and $\psi \equiv 0$). Randomized algorithms (such as randomized mirror-prox) for such problems were explored in Juditsky et al. [2011, 2013] to achieve complexity guarantees of the form $O(\sigma^2\varepsilon^{-2} + \|\mathbf{A}\|_{\mathcal{X},\mathcal{Y}^*}\,\varepsilon^{-1})$, where σ^2 is a measurement of noise arising from using stochastic estimates of matrix-vector multiplications and $\|\mathbf{A}\|_{\mathcal{X},\mathcal{Y}^*} = \sup\{\|\mathbf{A}\boldsymbol{x}\|_{\mathcal{Y}^*} : \|\boldsymbol{x}\|_{\mathcal{X}} = 1\}$ is the induced matrix norm. For general problems, the approaches of Chambolle and Pock [2011], Alacaoglu and Malitsky [2022] achieve

$$\min\left\{O\left(nd\,\|\mathbf{A}\|_{2,2}\,\varepsilon^{-1}\right), O\left(nd + \sqrt{nd(n+d)}\,\|\mathbf{A}\|_{\text{Fro}}\,\varepsilon^{-1}\right)\right\}$$

whereas these are reduced by Song et al. [2021] and Alacaoglu et al. [2022][2] to

$$O\left(nd + d\frac{n\max_j \|\mathbf{A}_{j\cdot}\|_2}{\varepsilon}\right), \tag{3.94}$$

when the objective function is *separable*, where $\mathbf{A}_{i\cdot}$ denotes the i-th row of $\mathbf{A}$. In our notation, $G_i = \|\mathbf{A}_{i\cdot}\|_\infty$ when we equip $\mathcal{X}$ with the ℓ_1-norm. Thus, due to separability, we may apply Theorem 3.5.1 with $N = n$ to achieve a global complexity of

$$O\left(nd + d\frac{\sqrt{(\sum_{i=1}^n \|\mathbf{A}_{i\cdot}\|_\infty)\cdot(n\max_j \|\mathbf{A}_{j\cdot}\|_\infty)}}{\varepsilon}\right). \tag{3.95}$$

Because

$$n\max_j \|\mathbf{A}_{j\cdot}\|_2 \overset{(\circ)}{\geq} n\max_j \|\mathbf{A}_{j\cdot}\|_\infty \overset{(*)}{\geq} \sum_{i=1}^n \|\mathbf{A}_{i\cdot}\|_\infty,$$

and $(\circ)$ offers an up to a $\sqrt{d}$-factor improvement and $(*)$ offers an up to n-factor improvement, our result can offer up to an order-$\sqrt{nd}$ improvement overall. This improvement is realized when within-row entries are highly uniform and within-column entries are highly non-uniform, leading to highly non-uniform infinity norms of the rows. For linearly constrained problems, Alacaoglu et al. [2022] achieve $O\left(nd + d\sum_{i=1}^n \|\mathbf{A}_{i\cdot}\|_2/\varepsilon\right)$. It is relevant

[2]Guarantees might be for the expected supremum of gap or supremum of expected gap. We compare them side-by-side due to the argument in the earlier part of the section.

to note that our improvement relies heavily on the non-uniform sampling and non-uniform historical regularization applied in Section 3.4.1 and Section 3.5.1. Non-uniform sampling has been applied in the works above, as well as earlier works such as Alacaoglu et al. [2017] and Chambolle et al. [2018], but our findings show that sampling strategies may not be sufficient to remove the extraneous dimension factors in the iteration complexity.

We also consider a particular application to (non-separably constrained) matrix games, that is, the problem

$$\min_{\boldsymbol{x}\in\Delta^{d-1}} \max_{\boldsymbol{y}\in\Delta^{n-1}} \left[\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) := \boldsymbol{y}^\top \mathbf{A}\boldsymbol{x}\right],$$

for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and probability simplices Δ^{d-1} and Δ^{n-1}, meaning that $\phi \equiv 0$ and $\psi \equiv 0$. Let A_{ij} denote the element (i,j)-th entry of $\mathbf{A}$ and let $\mathrm{nnz}(\mathbf{A})$ be the number of such entries that are non-zero. For this example problem, the results of Carmon et al. [2019, Thm. 1] and Alacaoglu and Malitsky [2022, Coro. 9 & Pg. 37] achieve a global complexity of

$$\tilde{O}\left(\mathrm{nnz}(\mathbf{A}) + \sqrt{\mathrm{nnz}(\mathbf{A})\cdot(n+d)}\frac{\max_{i,j}|A_{ij}|}{\varepsilon}\right) \tag{3.96}$$

whereas the method of Diakonikolas [2025, Thm. 1 & Pg. 19] achieves

$$\tilde{O}\left(\mathrm{nnz}(\mathbf{A}) + (n+d)\frac{\left(\sum_i \|\mathbf{A}_{i\cdot}\|_\infty^{2/3} + \sum_j \|\mathbf{A}_{\cdot j}\|_\infty^{2/3}\right)^{3/2}}{\varepsilon}\right), \tag{3.97}$$

where $\mathbf{A}_{\cdot j}$ is the j-th column of $\mathbf{A}$ and $\mathbf{A}_{i\cdot}$ is defined as in (3.94). Due to the non-separability of Δ^{n-1}, we may apply the result of Theorem 3.4.1 with $N = n$ and $G_i = \|\mathbf{A}_{i\cdot}\|_\infty$ to achieve a global complexity of

$$\tilde{O}\left(\mathrm{nnz}(\mathbf{A}) + \frac{(n+d)\sum_i \|\mathbf{A}_{i\cdot}\|_\infty}{\varepsilon}\right).$$

Notice that this is a direct improvement over (3.97) by replacing the $\ell_{2/3}$-norms with the ℓ_1-norm. On the other hand, for cases when $\mathbf{A}$ is dense but if the infinity norms of the rows are highly non-uniform, the complexity above will improve over (3.96) by a factor of

$\sqrt{\operatorname{nnz}(\mathbf{A})/(n+d)}$ (which is order $\sqrt{n}$ for dense square matrices). If these row norms are not highly non-uniform, but the matrix is still dense, then our complexity will be worse than (3.96) by an $n \cdot \sqrt{(n+d)/\operatorname{nnz}(\mathbf{A})}$-factor (which is also order-$\sqrt{n}$ for dense square matrices).

3.7 Possible Extensions

3.7.1 Certificates of Suboptimality

The goal of this section is to provide an *online accuracy certificate* (or simply *certificate*) for the dual-linear min-max problem (3.2). In other words, we wish to find a continuous function of the primal-dual pair that is zero if and only if it is evaluated at the optimum, and can be computed by the user without necessarily having knowledge of the minimum. This is especially useful in software packages, as it can be use to determine a stopping criterion for the algorithm.

For any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, define the functions

$$\Phi(\boldsymbol{x}) = \max_{\boldsymbol{v} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) \text{ and } \Psi(\boldsymbol{y}) = \min_{\boldsymbol{u} \in \mathcal{X}} \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}).$$

One such certificate is a direct upper bound on the primal-dual gap

$$\operatorname{Gap}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\boldsymbol{v} \in \mathcal{Y}} \{\mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})\} - \min_{\boldsymbol{u} \in \mathcal{X}} \{\mathcal{L}(\boldsymbol{u}, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})\}. \tag{3.98}$$

While the quantity above has the appealing property of agreeing with our notion of suboptimality tracked in the proofs, it is not necessarily the best quantity to track in practice when the strong convexity constants are low (or even zero). Instead, we pursue an upper bound on the *smoothed duality gap* [Walwil and Fercoq, 2025] given by a hyperparameter $\beta = (\beta_{\boldsymbol{x}}, \beta_{\boldsymbol{y}})$ and defined as

$$\begin{aligned} \operatorname{Gap}_{\beta}(\boldsymbol{x}, \boldsymbol{y}) = &\max_{\boldsymbol{v} \in \mathcal{Y}} \left\{ \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \frac{\beta_{\boldsymbol{y}}}{2} \|\boldsymbol{v} - \boldsymbol{y}\|_{\mathcal{Y}}^2 \right\} \\ &- \min_{\boldsymbol{u} \in \mathcal{X}} \left\{ \mathcal{L}(\boldsymbol{u}, \boldsymbol{y}) - \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \boldsymbol{x}\|_{\mathcal{X}}^2 \right\}, \end{aligned} \tag{3.99}$$

for a potential return value $(\boldsymbol{x}, \boldsymbol{y})$ (possibly computed using past iterates of the algorithm), where $\beta_{\boldsymbol{x}} \geq 0$ and $\beta_{\boldsymbol{y}} \geq 0$. This may also be called the *self-centered* smoothed duality gap because the primal regularizer is centered at $\boldsymbol{x}$ and the dual regularizer is centered at $\boldsymbol{y}$. See Fercoq [2023, Proposition 15] for a proof of its status as a convergence certificate, in that $\mathrm{Gap}_\beta(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ and that $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a saddle point of the objective (assuming that one exists) if and only if $\mathrm{Gap}_\beta(\boldsymbol{x}^\star, \boldsymbol{y}^\star) = 0$. Notice in addition that for any choice of β and any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, it holds that

$$\frac{\mu + \beta_{\boldsymbol{x}}}{2} \left\| \boldsymbol{x} - \boldsymbol{u}^\star(\boldsymbol{x}) \right\|_{\mathcal{X}}^2 + \frac{\nu + \beta_{\boldsymbol{y}}}{2} \left\| \boldsymbol{y} - \boldsymbol{v}^\star(\boldsymbol{y}) \right\|_{\mathcal{Y}}^2 \leq \mathrm{Gap}_\beta(\boldsymbol{x}, \boldsymbol{y}) \leq \mathrm{Gap}(\boldsymbol{x}, \boldsymbol{y}),$$

where $\boldsymbol{u}^\star(\boldsymbol{x})$ and $\boldsymbol{v}^\star(\boldsymbol{y})$ denote the minimizer and maximizer of (3.99), respectively. Thus, $\mathrm{Gap}_\beta(\boldsymbol{x}, \boldsymbol{y})$ enjoys the same convergence rate as the primal-dual gap.

We proceed to upper bound $\mathrm{Gap}_\beta(\boldsymbol{x}, \boldsymbol{y})$ in a form that may not depend on unknown quantities such as the saddle-point $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ in the strong convex-strongly concave setting. Instead, we operate with the following toolkit. For any $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{v} \in \mathbb{R}^n$, we assume access to the "proximal gradient" oracles

$$\boldsymbol{x} \mapsto \min_{\boldsymbol{u} \in \mathcal{X}} \langle \boldsymbol{x}, \boldsymbol{u} \rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \left\| \boldsymbol{u} - \boldsymbol{x} \right\|_{\mathcal{X}}^2 \text{ and } \boldsymbol{y} \mapsto \max_{\boldsymbol{v} \in \mathcal{Y}} \langle \boldsymbol{y}, \boldsymbol{v} \rangle - \psi(\boldsymbol{v}) - \frac{\beta_{\boldsymbol{y}}}{2} \left\| \boldsymbol{v} - \boldsymbol{y} \right\|_{\mathcal{Y}}^2, \tag{3.100}$$

for which the output of either map is finite when $\mu + \beta_{\boldsymbol{x}} > 0$ and $\nu + \beta_{\boldsymbol{y}} > 0$. With a slight abuse of notation, consider a sequence of primal-dual iterates $(\boldsymbol{x}_k, \boldsymbol{y}_k)_{k \geq 1}$, where each $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ represents the estimated solution *after* $(n \vee d)k$ *iterations* of an underlying algorithm. Thus, we evaluate the algorithm at every $(n \vee d)$ iterations, so that if we make a full batch gradient computation (at $O(nd)$ cost) the complexity of computing the certificate costs $O(n + d)$ complexity when amortized over the sequence. Finally, our bound can be applied to *any* algorithm, or sequence of primal-dual iterates, including the methods of Chapter 2.

We prove an upper bound on (3.99) that can be computed in practice and supports averaging iterates. We first state a generic upper bound that can aggregate many points and specify practical options after the statement of the result. Below, the notation $i : k$ is used

to index an element of the k-length row of a triangular array.

Lemma 3.7.1. *Consider probability mass weights $(\lambda_{i:k})_{i=1}^{k}$ and evaluation points $\bar{\boldsymbol{x}}_k, \hat{\boldsymbol{x}}_{1:k}, \ldots, \hat{\boldsymbol{x}}_{k:k} \in \mathcal{X}$, and $\hat{\boldsymbol{y}}_{1:k}, \ldots, \hat{\boldsymbol{y}}_{k:k} \in \mathcal{Y}$. Define*

$$\bar{\boldsymbol{y}}_k = \sum_{i=1}^{k} \lambda_{i:k} \hat{\boldsymbol{y}}_{i:k}.$$

Assume that $f_1, \ldots, f_n$ are bounded below by zero. Then, we have that

$$\mathrm{Gap}_\beta(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{y}}_k) \leq \max_{\boldsymbol{v} \in \mathcal{Y}} \left\{ \langle f(\bar{\boldsymbol{x}}_k), \boldsymbol{v} \rangle - \psi(\boldsymbol{v}) - \frac{\beta_{\boldsymbol{y}}}{2} \left\| \boldsymbol{v} - \bar{\boldsymbol{y}}_k \right\|_2^2 \right\} + \phi(\bar{\boldsymbol{x}}_k) + \psi(\bar{\boldsymbol{y}}_k) - \max\left\{ M_k, 0 \right\}, \tag{3.101}$$

where

$$\begin{aligned} M_k = & \min_{\boldsymbol{u} \in \mathcal{X}} \left\{ \sum_{i=1}^{k} \lambda_{i:k} \left\langle \nabla f(\hat{\boldsymbol{x}}_{i:k})^\top \hat{\boldsymbol{y}}_{i:k}, \boldsymbol{u} \right\rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \left\| \boldsymbol{u} - \bar{\boldsymbol{x}}_k \right\|_{\mathcal{X}}^2 \right\} \\ & + \sum_{i=1}^{k} \lambda_{i:k} \left\langle \hat{\boldsymbol{y}}_{i:k}, f(\hat{\boldsymbol{x}}_{i:k}) - \nabla f(\hat{\boldsymbol{x}}_{i:k}) \hat{\boldsymbol{x}}_{i:k} \right\rangle. \end{aligned}$$

Proof. First, we expand the two terms in (3.99) and observe

$$\begin{aligned} & \min_{\boldsymbol{u} \in \mathcal{X}} \left\{ \mathcal{L}(\boldsymbol{u}, \bar{\boldsymbol{y}}_k) - \mathcal{L}(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{y}}_k) + \frac{\beta_{\boldsymbol{x}}}{2} \left\| \boldsymbol{u} - \bar{\boldsymbol{x}}_k \right\|_{\mathcal{X}}^2 \right\} \\ & \quad = \langle f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{y}}_k \rangle + \phi(\bar{\boldsymbol{x}}_k) - \min_{\boldsymbol{u} \in \mathcal{X}} \left\{ \langle f(\boldsymbol{u}), \bar{\boldsymbol{y}}_k \rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \left\| \boldsymbol{u} - \bar{\boldsymbol{x}}_k \right\|_{\mathcal{X}}^2 \right\} \\ & \max_{\boldsymbol{v} \in \mathcal{Y}} \left\{ \mathcal{L}(\bar{\boldsymbol{x}}_k, \boldsymbol{v}) - \mathcal{L}(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{y}}_k) - \frac{\beta_{\boldsymbol{y}}}{2} \left\| \boldsymbol{v} - \bar{\boldsymbol{y}}_k \right\|_{\mathcal{Y}}^2 \right\} \\ & \quad = \max_{\boldsymbol{v} \in \mathcal{Y}} \left\{ \langle f(\bar{\boldsymbol{x}}_k), \boldsymbol{v} \rangle - \psi(\boldsymbol{v}) - \frac{\beta_{\boldsymbol{y}}}{2} \left\| \boldsymbol{v} - \bar{\boldsymbol{y}}_k \right\|_{\mathcal{Y}}^2 \right\} - \langle f(\bar{\boldsymbol{x}}_k), \bar{\boldsymbol{y}}_k \rangle + \psi(\bar{\boldsymbol{y}}_k), \end{aligned}$$

which sum to

$$\begin{aligned} \mathrm{Gap}(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{y}}_k) = & \max_{\boldsymbol{v} \in \mathcal{Y}} \left\{ \langle f(\bar{\boldsymbol{x}}_k), \boldsymbol{v} \rangle - \psi(\boldsymbol{v}) - \frac{\beta_{\boldsymbol{y}}}{2} \left\| \boldsymbol{v} - \boldsymbol{y} \right\|_{\mathcal{Y}}^2 \right\} + \phi(\bar{\boldsymbol{x}}_k) + \psi(\bar{\boldsymbol{y}}_k) \\ & - \min_{\boldsymbol{u} \in \mathcal{X}} \left\{ \langle \bar{\boldsymbol{y}}_k, f(\boldsymbol{u}) \rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \left\| \boldsymbol{u} - \bar{\boldsymbol{x}}_k \right\|_{\mathcal{X}}^2 \right\}. \end{aligned} \tag{3.102}$$

Note that all terms are computable besides the minimization over $u \in \mathcal{X}$. Thus, the rest of the

proof relies on lower bounding the minimization term $\min_{\boldsymbol{u}\in\mathcal{X}} \left\{ \langle \bar{\boldsymbol{y}}, f(\boldsymbol{u})\rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 \right\}$, which is non-negative by assumption. Write

$$\begin{aligned}
&\langle \bar{\boldsymbol{y}}_k, f(\boldsymbol{u})\rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 \\
&= \sum_{i=1}^{k} \lambda_{i:k} \langle \hat{\boldsymbol{y}}_{i:k}, f(\boldsymbol{u})\rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 \\
&\geq \sum_{i=1}^{k} \lambda_{i:k} \left(\langle \hat{\boldsymbol{y}}_{i:k}, f(\hat{\boldsymbol{x}}_{i:k})\rangle + \left\langle \nabla f(\hat{\boldsymbol{x}}_{i:k})^\top \hat{\boldsymbol{y}}_{i:k}, \boldsymbol{u} - \hat{\boldsymbol{x}}_{i:k} \right\rangle \right) + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 ,
\end{aligned}$$

by convexity so that

$$\begin{aligned}
&\min_{\boldsymbol{u}\in\mathcal{X}} \left\{ \langle \bar{\boldsymbol{y}}, f(\boldsymbol{u})\rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 \right\} \\
&\geq \min_{\boldsymbol{u}\in\mathcal{X}} \left\{ \sum_{i=1}^{k} \lambda_{i:k} \left\langle \nabla f(\hat{\boldsymbol{x}}_{i:k})^\top \hat{\boldsymbol{y}}_{i:k}, \boldsymbol{u} \right\rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 \right\} \\
&\quad + \sum_{i=1}^{k} \lambda_{i:k} \langle \hat{\boldsymbol{y}}_{i:k}, f(\hat{\boldsymbol{x}}_{i:k})\rangle - \sum_{i=1}^{k} \lambda_{i:k} \left\langle \nabla f(\hat{\boldsymbol{x}}_{i:k})^\top \hat{\boldsymbol{y}}_{i:k}, \hat{\boldsymbol{x}}_{i:k} \right\rangle .
\end{aligned}$$

Substituting the definition of M_k completes the proof. □

One practical note is that the term that usually blows up when $\mu \approx 0$ is M_k—the minimization over $\boldsymbol{u} \in \mathcal{X}$—whenever the primal gradient is not nearly zero.

To convert this upper bound into an algorithm, we make the choices

$$\hat{\boldsymbol{x}}_{i:k} = \boldsymbol{x}_i, \quad \hat{\boldsymbol{x}}_{i:k} = \boldsymbol{x}_i, \text{ and } \bar{\boldsymbol{x}}_k = \sum_{i=1}^{k} \lambda_{i:k} \boldsymbol{x}_i.$$

For the weights $(\lambda_{i:k})_{i=1}^k$, we use an exponential moving average. Notice that all terms in (3.101) can be computed easily if every term that is indexed by k is maintained at every iteration of the optimization problem. See Algorithm 5 for an implementation-friendly description.

Figure 3.5 compares the certificate from Algorithm 5 to the standard (primal) gradient norm criterion on the distributionally robust classification and regression benchmarks from

Algorithm 5 Dual-Linear Online Accuracy Certificate

Inputs: Initial points $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, total epochs K, averaging constant $\lambda \in [0, 1)$, hyperparameters $\beta = (\beta_{\boldsymbol{x}}, \beta_{\boldsymbol{y}})$.

$\boldsymbol{z}_0 = \nabla f(\boldsymbol{x}_0)^\top \boldsymbol{y}_0 \in \mathbb{R}^d$, $r_0 = \langle \boldsymbol{y}_0, f(\boldsymbol{x}_0) - \nabla f(\boldsymbol{x}_0)\boldsymbol{x}_0 \rangle \in \mathbb{R}$, $(\bar{\boldsymbol{x}}_0, \bar{\boldsymbol{y}}_0) = (\boldsymbol{x}_0, \boldsymbol{y}_0)$.

for $k = 1, \ldots, K$ **do**

  $(\boldsymbol{x}_k, \boldsymbol{y}_k) = \text{RunEpoch}(\boldsymbol{x}_{k-1}, \boldsymbol{y}_{k-1})$.

  $(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{y}}_k) = (1 - \lambda) \cdot (\bar{\boldsymbol{x}}_{k-1}, \bar{\boldsymbol{y}}_{k-1}) + \lambda \cdot (\boldsymbol{x}_k, \boldsymbol{y}_k)$.

  $\boldsymbol{z}_k = (1 - \lambda) \cdot \boldsymbol{z}_{k-1} + \lambda \cdot \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{y}_k$.

  $r_k = (1 - \lambda) \cdot r_{k-1} + \lambda \cdot \langle \boldsymbol{y}_k, f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\boldsymbol{x}_k \rangle$.

  $M_k = \min_{\boldsymbol{u} \in \mathcal{X}} \left\{ \langle \boldsymbol{z}_k, \boldsymbol{u} \rangle + \phi(\boldsymbol{u}) + \frac{\beta_{\boldsymbol{x}}}{2} \|\boldsymbol{u} - \bar{\boldsymbol{x}}_k\|_{\mathcal{X}}^2 \right\} + r_k$.

  $\varepsilon_k = \max_{\boldsymbol{v} \in \mathcal{Y}} \left\{ \langle f(\bar{\boldsymbol{x}}_k), \boldsymbol{v} \rangle - \psi(\boldsymbol{v}) - \frac{\beta_{\boldsymbol{y}}}{2} \|\boldsymbol{v} - \bar{\boldsymbol{y}}_k\|_{\mathcal{Y}}^2 \right\} - \max\{M_k, 0\} + \phi(\bar{\boldsymbol{x}}_k) + \psi(\bar{\boldsymbol{y}}_k)$.

Output: Sequence $(\varepsilon_k)_{k=1}^K$.

Section 2.9 under the same experimental setup. Precisely, the gradient norm criterion is computed as

$$\|\nabla \Phi(\boldsymbol{x}_k)\|_{\mathcal{X}^*}^2 \text{ for } \Phi(\boldsymbol{x}) := \max_{\boldsymbol{v} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{v}) \tag{3.103}$$

at epoch k. For the certificate parameters, we set $\beta_{\boldsymbol{x}} = 1$, $\beta_{\boldsymbol{y}} = 0$, and $\nu = 0.05$ in all experiments. In contrast, μ is tuned via training an empirical risk minimization model and choosing the value that yields the best generalization performance. Note that for some datasets, this may be as small as $\mu \approx 10^{-5}$, making a positive value of $\beta_{\boldsymbol{x}}$ essential. We find that the certificate, while not guaranteed to be an upper bound to the primal-dual gap, remains close in experimental settings using a flat hyperparameter choice. Thus, the tolerance can be set at approximately the same magnitude as would be required of the primal-dual gap/suboptimality.

3.7.2 Lower Bounds

In this section, we show that under particular parameter regimes, the full vector update method analyzed in Theorem 3.3.1 achieves a matching lower bound on the number of iterations k required to achieve an ε-suboptimal primal-dual pair $(\boldsymbol{x}_k, \boldsymbol{y}_k)$. This helps calibrate

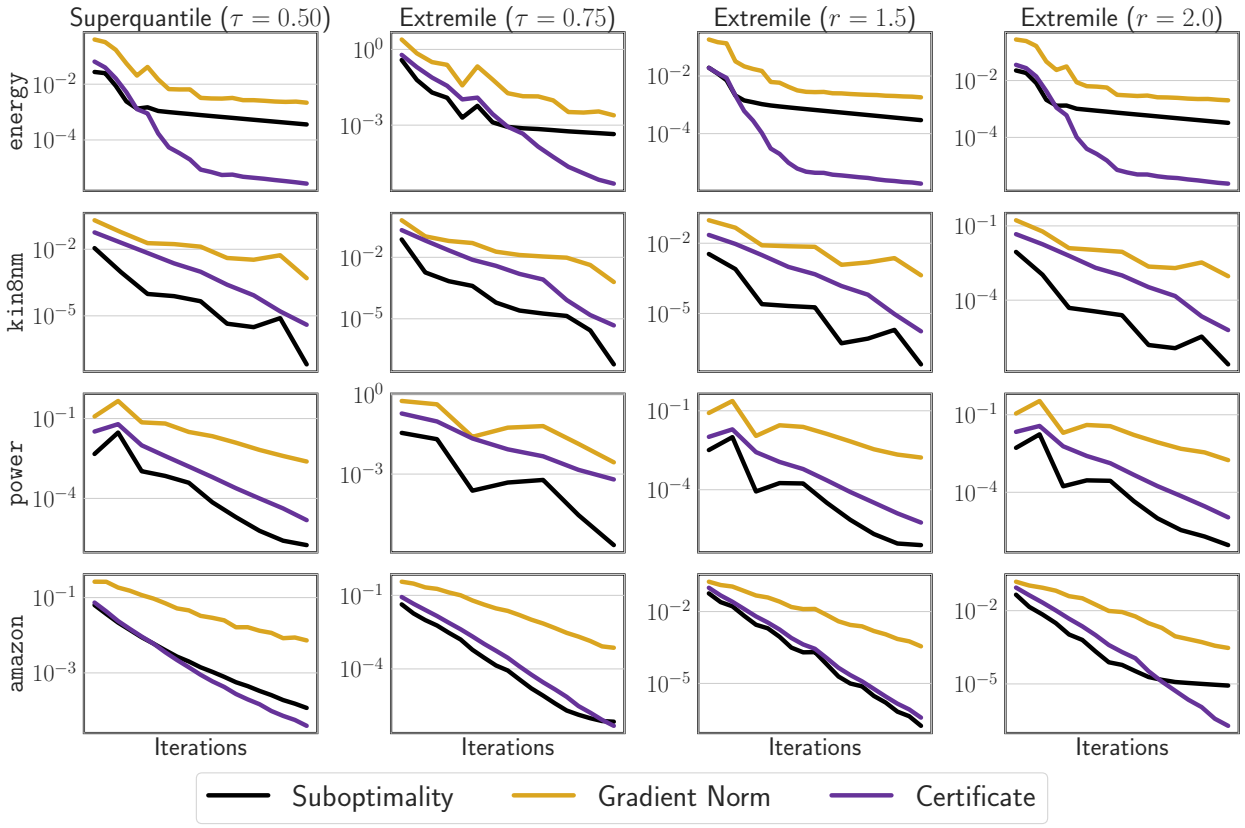

Figure 3.5: **Experimental Evaluation of Online Accuracy Certificates.** In all plots, the x-axis refers to the iteration count, which may differ between datasets. Each line represents the gradient norm (3.103), certificate (3.101), and the primal-dual gap (3.98).

the upper bounds stated in previous sections to the true problem hardness of semilinear min-max programs. We assume that $\mu > 0$ and $\nu > 0$, guaranteeing the existence of a unique solution $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ satisfying

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}) \leq \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \leq \min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^\star),$$

i.e. a saddle point of the objective $\mathcal{L}$. We use the gap criterion (3.4) by setting $(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, and recognize by strong convexity that

$$\mathrm{Gap}^{\boldsymbol{x}^\star, \boldsymbol{y}^\star}(\boldsymbol{x}, \boldsymbol{y}) := \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^\star) - \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{y}) \geq \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{X}}^2 + \frac{\nu}{2} \|\boldsymbol{y} - \boldsymbol{y}^\star\|_{\mathcal{Y}}^2 \geq 0.$$

The lower bound will be applied to the term $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mathcal{X}}^2$, and as a result apply to the gap criterion. The proof follows similar techniques as those used in Zhang et al. [2022] for bilinearly coupled objectives:

1. We describe precisely the class of algorithms considered, and verify that the method of Theorem 3.3.1 is such an algorithm.

2. We then introduce a "hard" instance which satisfies Assumption 3.2.2 and Assumption 3.2.1, and then characterize particular subspaces for which the primal-dual iterates are guaranteed to be members.

3. Finally, we compute by hand an arbitrarily close approximation $\hat{\boldsymbol{x}}^\star$ of the optimal primal solution $\boldsymbol{x}^\star$, and show that for d sufficiently large and $\sqrt{\mu\nu}/G \leq 1/2$,

$$\|\boldsymbol{x}_{2k} - \hat{\boldsymbol{x}}^\star\|_{\mathcal{X}}^2 \geq C(1 - 2\sqrt{\mu\nu}/G)^{2k} \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}^\star\|_{\mathcal{X}}^2 .$$

This describes the parameter regime

$$\mu G^2 \geq \nu L^2, \tag{3.104}$$

indicating that the mixed condition number in Theorem 3.3.1 is dominant. This reflects practice in the case of the examples from Section 3.2. In distributionally robust optimization, ν is often chosen as a smoothing parameter for a problem that is originally non-smooth in the primal. In fully composite optimization, we have that $(1/\nu)$ is the smoothness constant of F, which could be large for F "close to non-smooth". Finally, in the case of optimization with functional constraints, ν refers to a penalty applied to the Lagrange multipliers, which would originally be unpenalized in order to enforce the constraint exactly. Remarkably, in the parameter regime (3.104), we may use a *bilinear* example as our hard instance. We fix the dual-linear convention in this section as well, and handle the primal-linear case at the end of the section.

Algorithm Class and Hard Instance This section contains the portions of the proof that generalize and improve on the result of Zhang et al. [2022]. We define our class of methods to satisfy a linear span assumption involving proximal oracles. This definition will reflect [Zhang et al., 2022, Definition 2.1], but account for a number of changes: 1) possible nonbilinearity, 2) differing problem dimensions $n \neq d$, 3) non-Euclidean proximal oracles, 4) extrapolated gradient estimates, and 5) the use of $\boldsymbol{x}_k$ in the update for $\boldsymbol{y}_k$. Define the non-Euclidean proximal operators

$$p_{\eta_k,\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{g}}) := \arg\min_{\boldsymbol{x}\in\mathcal{X}} \langle \tilde{\boldsymbol{g}}, \boldsymbol{x}\rangle + \phi(\boldsymbol{x}) + \eta_k \Delta_{\mathcal{X}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \qquad (\eta_k \geq 0, \tilde{\boldsymbol{x}} \in \mathcal{X}, \tilde{\boldsymbol{g}} \in \mathcal{X}^*)$$
$$p_{\delta_k,\psi}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{f}}) := \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \left\langle \tilde{\boldsymbol{f}}, \boldsymbol{y}\right\rangle - \psi(\boldsymbol{y}) - \delta_k \Delta_{\mathcal{Y}}(\boldsymbol{y}, \tilde{\boldsymbol{y}}) \qquad (\delta_k \geq 0, \tilde{\boldsymbol{y}} \in \mathcal{Y}, \tilde{\boldsymbol{f}} \in \mathcal{Y}^*)$$

and consider the following definition.

Definition 3.7.1 (Dual-Linear Proximal Algorithm Class)**.** We define a (dual-linear) *deterministic proximal gradient algorithm* to be a sequence of primal-dual iterates $(\boldsymbol{x}_k, \boldsymbol{y}_k)_{k\geq 0}$ satisfying $(\boldsymbol{x}_k, \boldsymbol{y}_k) \in \mathcal{H}_k^{\mathcal{X}} \times \mathcal{H}_k^{\mathcal{Y}}$, where $\mathcal{H}_k^{\mathcal{X}} \subseteq \mathcal{X}$ and $\mathcal{H}_k^{\mathcal{Y}} \subseteq \mathcal{Y}$ are constructed from the following steps. First, for $k \geq 0$, define the subspaces

$$\mathcal{G}_k^{\mathcal{X}} = \operatorname{span}\left\{\nabla f(\boldsymbol{x}_i)^\top \boldsymbol{y}_i : 0 \leq i \leq k\right\} \subseteq \mathcal{X}^*$$
$$\mathcal{G}_k^{\mathcal{Y}} = \operatorname{span}\left\{f(\boldsymbol{x}_i) : 0 \leq i \leq k+1\right\} \subseteq \mathcal{Y}^*.$$

Define $\mathcal{H}_0^{\mathcal{X}} = \operatorname{span}\{\boldsymbol{x}_0\}$ and $\mathcal{H}_0^{\mathcal{Y}} = \operatorname{span}\{\boldsymbol{y}_0\}$, and for $k \geq 1$, we recursively define

$$\mathcal{H}_k^{\mathcal{X}} = \operatorname{span}\left(\left\{p_{\eta_k\phi}(\tilde{\boldsymbol{x}}_{k-1}, \tilde{\boldsymbol{g}}_{k-1}) : \tilde{\boldsymbol{x}}_{k-1} \in \mathcal{H}_{k-1}^{\mathcal{X}}, \tilde{\boldsymbol{g}}_{k-1} \in \mathcal{G}_{k-1}^{\mathcal{X}}\right\} \cup \mathcal{H}_{k-1}^{\mathcal{X}}\right),$$
$$\mathcal{H}_k^{\mathcal{Y}} = \operatorname{span}\left(\left\{p_{\delta_k\psi}(\tilde{\boldsymbol{y}}_{k-1}, \tilde{\boldsymbol{f}}_k) : \tilde{\boldsymbol{y}}_{k-1} \in \mathcal{H}_{k-1}^{\mathcal{Y}}, \tilde{\boldsymbol{f}}_{k-1} \in \mathcal{G}_{k-1}^{\mathcal{Y}}\right\} \cup \mathcal{H}_{k-1}^{\mathcal{Y}}\right).$$

We must confirm that the algorithm analyzed in Theorem 3.3.1 adheres to this definition.

Lemma 3.7.2. *The primal-dual updates* (3.11) *and* (3.12) *define a deterministic proximal gradient algorithm in the sense of Definition 3.7.1.*

Proof. First for any $k \geq 1$, notice the primal update (3.11) and (3.12) can be written as

$$\boldsymbol{x}_k = p_{\eta_k \phi}(\boldsymbol{x}_{k-1}, \bar{\boldsymbol{g}}_{k-1}) \text{ for } \eta_k = \frac{A_{k-1}\mu + \mu_0}{a_k} \geq 0, \tag{3.105}$$

$$\boldsymbol{y}_k = p_{\delta_k \psi}(\boldsymbol{y}_{k-1}, f(\boldsymbol{x}_k)) \text{ for } \delta_k = \frac{A_{k-1}\nu + \nu_0}{a_k} \geq 0. \tag{3.106}$$

By the definition of the primal gradient estimate $\bar{\boldsymbol{g}}_{k-1}$ from (3.27), we also have that

$$(\bar{\boldsymbol{g}}_{k-1}, f(\boldsymbol{x}_k)) \in \mathcal{G}_{k-1}^{\mathcal{X}} \times \mathcal{G}_{k-1}^{\mathcal{Y}}.$$

By construction, we have that initial iterates $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{H}_0^{\mathcal{X}} \times \mathcal{H}_0^{\mathcal{Y}}$. Combine this base case with the recursions (3.105), and (3.106) to prove by induction that $\boldsymbol{x}_k \in \mathcal{H}_k^{\mathcal{X}}$ and $\boldsymbol{y}_k \in \mathcal{H}_k^{\mathcal{Y}}$. □

Next, we proceed to define a special case of the objective $\mathcal{L}$ and establish a *zero-chain* property for the sequence of iterates. Consider a setting in which $n \geq d$, $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_2$ and $\|\cdot\|_{\mathcal{Y}} = \|\cdot\|_2$ (the ℓ_2-norms on $\mathbb{R}^d$ and $\mathbb{R}^n$), respectively. The objective is written

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \underbrace{\left\langle \boldsymbol{y}, \tfrac{G}{2}\boldsymbol{A}\boldsymbol{x} \right\rangle}_{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle} - \underbrace{\frac{\nu}{2}\|\boldsymbol{y}\|_2^2}_{\psi(\boldsymbol{y})} + \underbrace{\boldsymbol{c}^\top \boldsymbol{x} + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2}_{\phi(\boldsymbol{x})}, \tag{3.107}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{c} \in \mathbb{R}^d$ are to-be-specified. As a result, we have that

$$\Delta_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 \text{ and } \Delta_{\mathcal{Y}}(\boldsymbol{y}, \boldsymbol{y}') = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{y}'\|_2^2.$$

We characterize the subspaces $\mathcal{H}_k^{\mathcal{X}}$ and $\mathcal{H}_k^{\mathcal{Y}}$ from Definition 3.7.1 on this particular example, which will reveal the structure of $\boldsymbol{x}_k$ that is used in the next step. First, notice that the proximal oracles can be computed directly for (3.107). Recalling that we defined $(\boldsymbol{x}_k, \boldsymbol{y}_k) = (\boldsymbol{x}_0, \boldsymbol{y}_0)$ for $k < 0$, we have for all $k \geq 1$,

$$\boldsymbol{x}_k = \frac{\eta_k, \boldsymbol{x}_{k-1} - \bar{\boldsymbol{g}}_{k-1}}{\mu + \eta_k} \subseteq \operatorname{span}\left\{\boldsymbol{x}_{k-1}, \boldsymbol{c}, \boldsymbol{A}^\top \boldsymbol{y}_{k-1}, \boldsymbol{A}^\top \boldsymbol{y}_{k-2}, \ldots, \boldsymbol{A}^\top \boldsymbol{y}_0\right\} \tag{3.108}$$

$$\boldsymbol{y}_k = \frac{\delta_k, \boldsymbol{y}_{k-1} - \bar{\boldsymbol{f}}_{k-1}}{\nu + \delta_k} \subseteq \operatorname{span}\left\{\boldsymbol{y}_{k-1}, \boldsymbol{A}\boldsymbol{x}_k, \boldsymbol{A}\boldsymbol{x}_{k-1}, \ldots, \boldsymbol{A}\boldsymbol{x}_0\right\}. \tag{3.109}$$

We can then present the key result of this step.

Lemma 3.7.3. *For any determistic proximal gradient algorithm in the sense of Definition 3.7.1 with $\boldsymbol{x}_0 = \mathbf{0}_d$ and $\boldsymbol{y}_0 = \mathbf{0}_n$ applied to the problem (3.107), we have that $\mathcal{H}_1^{\mathcal{X}} = \{\boldsymbol{c}\}$, $\mathcal{H}_1^{\mathcal{Y}} = \operatorname{span}\{\boldsymbol{A}\boldsymbol{c}\}$, and for all $k \geq 2$,*

$$\mathcal{H}_k^{\mathcal{X}} \subseteq \operatorname{span}\left\{(\boldsymbol{A}^\top \boldsymbol{A})^i \boldsymbol{c} : 0 \leq i \leq k-1\right\}, \tag{3.110}$$

$$\mathcal{H}_k^{\mathcal{Y}} \subseteq \operatorname{span}\left\{(\boldsymbol{A}\boldsymbol{A}^\top)^i (\boldsymbol{A}\boldsymbol{c}) : 0 \leq i \leq k-1\right\}. \tag{3.111}$$

Proof. Using that $\boldsymbol{x}_0 = \mathbf{0}_d$ and $\boldsymbol{y}_0 = \mathbf{0}_n$, and directly applying the formulas (3.108) and (3.109) we achieve the base cases

$$\mathcal{H}_1^{\mathcal{X}} = \{\boldsymbol{c}\} \text{ and } \mathcal{H}_1^{\mathcal{Y}} \subseteq \operatorname{span}\{\boldsymbol{A}\boldsymbol{c}\}.$$

We argue the general case by induction. Fix $k \geq 2$ and assume that for $\{\kappa : 0 \leq \kappa \leq k-1\}$, we have that (3.110) and (3.111) hold. Then, we may rewrite the conclusion of (3.108) and (3.109) as

$$\boldsymbol{x}_k \subseteq \operatorname{span}\left\{\boldsymbol{x}_{k-1}, \boldsymbol{c}, \boldsymbol{A}^\top \boldsymbol{y}_{k-1}\right\} \text{ and } \boldsymbol{y}_k \subseteq \operatorname{span}\left\{\boldsymbol{y}_{k-1}, \boldsymbol{A}\boldsymbol{x}_k\right\},$$

and we may apply these inclusions in an alternating manner to claim

$$\begin{aligned}
\boldsymbol{x}_k \subseteq \operatorname{span}\left\{\boldsymbol{x}_{k-1}, \boldsymbol{c}, \boldsymbol{A}^\top \boldsymbol{y}_{k-1}\right\} &\subseteq \operatorname{span}\left\{\boldsymbol{x}_{k-1}, \boldsymbol{c}, \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x}_{k-1}, \boldsymbol{A}^\top \boldsymbol{y}_{k-2}\right\} \\
&\subseteq \operatorname{span}\left\{\boldsymbol{x}_{k-1}, \boldsymbol{c}, (\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}_{k-1}, \ldots, (\boldsymbol{A}^\top \boldsymbol{A})\boldsymbol{x}_1\right\} \\
&= (\boldsymbol{A}^\top \boldsymbol{A})\left(\operatorname{span}\left\{(\boldsymbol{A}^\top \boldsymbol{A})^i \boldsymbol{c} : 0 \leq i \leq k-2\right\}\right) \\
&= \operatorname{span}\left\{(\boldsymbol{A}^\top \boldsymbol{A})^i \boldsymbol{c} : 0 \leq i \leq k-1\right\}.
\end{aligned}$$

Arguing similarly for the sequence $(\boldsymbol{y}_k)_{k \geq 1}$, we have that

$$\begin{aligned}
\boldsymbol{y}_k &\subseteq \operatorname{span}\left\{\boldsymbol{y}_{k-1}, \boldsymbol{A}\boldsymbol{c}, (\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{y}_{k-1}, \ldots, (\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{y}_1\right\} \\
&= \operatorname{span}\left\{(\boldsymbol{A}\boldsymbol{A}^\top)^i (\boldsymbol{A}\boldsymbol{c}) : 0 \leq i \leq k-1\right\}.
\end{aligned}$$

This completes the proof. □

Note that $\boldsymbol{x}_0 = \mathbf{0}_d$ and $\boldsymbol{y}_0 = \mathbf{0}_n$ can hold without loss of generality, as we may shift the input space to satisfy this condition. Now, we are ready to partially specify $\boldsymbol{A}$ and $\boldsymbol{c}$. We will set $\boldsymbol{A}$ to recover a variant of Nesterov's tridiagonal matrix (see [Nesterov, 2018, Section 2.3]), so that for

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1 \end{bmatrix}, \tag{3.112}$$

we have that

$$\boldsymbol{A}\boldsymbol{A}^\top = \begin{bmatrix} \mathbf{B} & \mathbf{0}_{d\times n-d} \\ \mathbf{0}_{n-d\times d} & \mathbf{0}_{n-d\times n-d} \end{bmatrix}, \quad \boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} 1 & -1 & & \\ -1 & \ddots & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{bmatrix}. \tag{3.113}$$

The operator norm of $\boldsymbol{A}$ will be the spectral norm $\|\boldsymbol{A}\|_{2,2} \leq 2$, so the multiplier $\frac{G}{2}$ makes f adhere to Assumption 3.2.2 for any $L \geq 0$. Using these specifications and Lemma 3.7.3, we establish the zero-chain property for the primal-dual sequence.

Corollary 3.7.1. *Let $n \geq d$ and use the notation $\boldsymbol{e}_{j:p}$ to denote the j-th standard basis vector in $\mathbb{R}^p$. Using $\boldsymbol{A}$ as defined in (3.113) and $\boldsymbol{c} = \beta\boldsymbol{e}_{1:d}$ for a constant $\beta \in \mathbb{R}$, we have that*

$$\mathcal{H}_k^{\mathcal{X}} \subseteq \operatorname{span}\left\{\boldsymbol{e}_{1:d}, \ldots, \boldsymbol{e}_{\min\{k,d\}:d}\right\}$$
$$\mathcal{H}_k^{\mathcal{Y}} \subseteq \operatorname{span}\left\{\boldsymbol{e}_{1:n}, \ldots, \boldsymbol{e}_{\min\{k,d\}:n}\right\}.$$

Proof. First, note that

$$\boldsymbol{A}\boldsymbol{c} = \beta \boldsymbol{A}\boldsymbol{e}_{1:d} = \beta \boldsymbol{e}_{1:d}.$$

Additionally, because $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ are tridiagonal we have that

$$\begin{aligned}
\boldsymbol{u} \in \operatorname{span}\left\{\boldsymbol{e}_{j:p} : j \le p \le d\right\} &\implies \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{u} \in \operatorname{span}\left\{\boldsymbol{e}_{j:p+1} : j \le p \le d-1\right\} \\
\boldsymbol{v} \in \operatorname{span}\left\{\boldsymbol{e}_{j:p} : j \le p \le n\right\} &\implies \boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{v} \in \operatorname{span}\left\{\boldsymbol{e}_{j:p+1} : j \le p \le d-1\right\}.
\end{aligned}$$

Combining the displays above with Lemma 3.7.3, we have that for $k \ge 2$,

$$\begin{aligned}
\mathcal{H}_k^{\mathcal{X}} &\subseteq \operatorname{span}\left\{\boldsymbol{e}_{1:d}, \ldots, \boldsymbol{e}_{\min\{k,d\}:d}\right\} \\
\mathcal{H}_k^{\mathcal{Y}} &\subseteq \operatorname{span}\left\{\boldsymbol{e}_{1:n}, \ldots, \boldsymbol{e}_{\min\{k,d\}:n}\right\}
\end{aligned}$$

the result as desired. □

Notice in the result above that d is a limiting factor, in that the zero components of $\boldsymbol{A}\boldsymbol{A}^\top$ prevent any further non-zero entries from appearing beyond element d. Thus, in the upcoming proofs, we will increase d to achieve the desired properties.

Dual-Linear Lower Bounds After having established Corollary 3.7.1, the rest of the argument follows very similarly to Zhang et al. [2022, Theorem 3.5]. Noting that $n \ge d$ we first partially maximize the objective over $\boldsymbol{y}$ (which can be solved by hand), to construct the primal gap $\Phi(\boldsymbol{x})$. We first recall that $\mathcal{Y}$ may include vectors with negative components without harming convexity, as all functions $(f_j)_{j=1}^n$ are linear, giving

$$\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{2}\boldsymbol{x}^\top \left(\frac{G^2}{4\nu}\boldsymbol{A}^\top\boldsymbol{A} + \mu I\right)\boldsymbol{x} + \boldsymbol{c}^\top\boldsymbol{x}.$$

We now fully specify the objective by setting $\beta = \frac{G^2}{4\nu}$ (recalling that $\boldsymbol{c} = \beta\boldsymbol{e}_{1:d}$), yielding

$$\Phi(\boldsymbol{x}) = \frac{G^2}{4\nu}\left(-\frac{1}{2}\boldsymbol{x}^\top\left(\boldsymbol{A}^\top\boldsymbol{A} + \alpha I\right)\boldsymbol{x} + \boldsymbol{e}_{1:d}^\top\boldsymbol{x}\right)$$

for $\alpha = \frac{4\mu\nu}{G^2}$. Consider the following result, which defines an approximation to the maximizer of the expression above.

Lemma 3.7.4. *[Zhang et al., 2022, Lemma 3.3] Consider the system of linear equations*

$$\left(\boldsymbol{A}^\top \boldsymbol{A} + \alpha I\right) \boldsymbol{x} = \boldsymbol{e}_{1:d},$$

for $\alpha > 0$ and $\boldsymbol{A}^\top \boldsymbol{A}$ defined in (3.113). Denote the unique solution of this system as $\boldsymbol{x}^\star$, which is guaranteed to exist by positive definiteness of $\boldsymbol{A}^\top \boldsymbol{A} + \alpha I$. Denote by

$$q \equiv q(\alpha) = \frac{1}{2}\left((2+\alpha) - \sqrt{(2+\alpha)^2 - 4}\right) \in (0,1)$$

the smallest root of the quadratic equation $1 - (2+\alpha)q + q^2 = 0$. Then, by defining

$$\hat{x}_i^\star = \frac{q^i}{1-q}, \forall i \in \{1, \ldots, d\} \tag{3.114}$$

and $\hat{\boldsymbol{x}}^\star = (\hat{x}_1^\star, \ldots, \hat{x}_d^\star)$ we have that

$$\|\hat{\boldsymbol{x}}^\star - \boldsymbol{x}^\star\|_2 \leq \frac{q^{d+1}}{\alpha(1-q)}.$$

We may now make the full claim.

Proposition 3.7.1. *Consider any algorithm satisfying Definition 3.7.1 with $\boldsymbol{x}_0 = \boldsymbol{0}_d$ and $\boldsymbol{y}_0 = \boldsymbol{0}_n$ applied to the problem (3.107), with $n \geq d \geq k$. We have that when $\sqrt{\alpha} = 2\sqrt{\mu\nu}/G \leq 1$,*

$$\mathrm{Gap}^{\boldsymbol{x}^\star, \boldsymbol{y}^\star}(\boldsymbol{x}_k, \boldsymbol{y}_k) \geq \frac{\mu}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \geq \frac{\mu}{8\sqrt{2}}\left(1 - \sqrt{\alpha}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - O((1-\alpha)^{2d+2})/\alpha)$$

Before stating the proof, note that d is a free parameter, so the second term in the lower bound can be made arbitrarily small.

Proof. By Corollary 3.7.1, we have that

$$\boldsymbol{x}_k \in \mathcal{H}_k^{\mathcal{X}} \subseteq \mathrm{span}\left\{\boldsymbol{e}_{1:d}, \ldots, \boldsymbol{e}_{k:d}\right\},$$

indicating that the last $d-k$ elements of $\boldsymbol{x}_k$ are necessarily zero. On the other hand, by the definition of $\hat{\boldsymbol{x}}^\star$, it must hold that elements $k+1,\ldots,d$ of $\hat{\boldsymbol{x}}^\star$ are non-zero with formula (3.114). Thus, we have that

$$\begin{aligned}\|\boldsymbol{x}_{2k}-\hat{\boldsymbol{x}}^\star\|_2^2 &\geq \sum_{i=k+1}^{d}(\hat{x}_i^\star)^2 = \left(\frac{q^k}{1-q}\right)^2\left(q^2+\ldots+q^{2(d-k)}\right)\\ &\geq \frac{q^{2k}}{2}\|\hat{\boldsymbol{x}}^\star\|_2^2 = \frac{q^{2k}}{2}\|\hat{\boldsymbol{x}}^\star-\boldsymbol{x}_0\|_2^2.\end{aligned}$$

Next, by using the triangle inequality $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and by the assumption that $\alpha=4\mu\nu/G\leq 1$, we have that

$$q(\alpha)=\frac{1}{2}\left(2+\alpha-\sqrt{\alpha^2+4\alpha}\right)\geq 1-\sqrt{\alpha}\geq 0,$$

so that

$$\|\boldsymbol{x}_{2k}-\hat{\boldsymbol{x}}^\star\|_2^2\geq\frac{q^{2k}}{2}\|\boldsymbol{x}_0-\hat{\boldsymbol{x}}^\star\|_2^2\geq\frac{1}{2}\left(1-\sqrt{\alpha}\right)^{2k}\|\boldsymbol{x}_0-\hat{\boldsymbol{x}}^\star\|_2^2.$$

By using that $\alpha/4\leq 1$, we also have the upper bound

$$q(\alpha)=1+\alpha/2-\sqrt{\alpha^2/4+\alpha}\leq 1+\alpha/2-\sqrt{\alpha^2/2}=1-\frac{\sqrt{2}-1}{2}\alpha.$$

Then, applying the upper bound and second claim of Lemma 3.7.4, we have that

$$\begin{aligned}\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|_2^2 &\geq \frac{1}{4\sqrt{2}}\left(1-\sqrt{\alpha}\right)^k\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2^2-\frac{3q^{2d+2}}{2\alpha(1-q)}\\ &= \frac{1}{4\sqrt{2}}\left(1-\sqrt{\alpha}\right)^k\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2^2-O((1-\alpha)^{2d+2})/\alpha).\end{aligned}$$

□

Primal-Linear Lower Bounds While all upper bounds can be proved with perfect symmetry in the primal-linear setting, this is not obvious in the case of the lower bounds due to the condition $n\geq d$. We modify the argument of Section 3.7 to establish hardness guarantees

for the problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \left[\mu(\boldsymbol{x},\boldsymbol{y}) := \langle g(\boldsymbol{y}),\boldsymbol{x}\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})\right], \tag{3.115}$$

where $\mathcal{X} \subseteq \left\{\boldsymbol{x}\in\mathbb{R}^d : x_l \geq l \text{ if } g_l \text{ is not affine}\right\}$ and the components of $g:\mathcal{Y}\to\mathcal{X}^*$ satisfy an analogous form of Assumption 3.2.2.

Assumption 3.7.1. Assume that each component of $g=(g_1,\ldots,g_d)$ is convex, and there exist $G>0$ and $L\geq 0$ such that for all $\boldsymbol{y}\in\mathcal{Y}$,

$$\|\nabla g(\boldsymbol{y})\|_{\mathcal{Y}\to\mathcal{X}^*} := \sup\left\{\|\nabla g(\boldsymbol{y})\bar{\boldsymbol{y}}\|_{\mathcal{X}^*} : \bar{\boldsymbol{y}}\in\mathcal{Y}, \|\bar{\boldsymbol{y}}\|_{\mathcal{Y}}=1\right\} \leq G.$$

and

$$\left\|\sum_{l=1}^d x_j(\nabla g_l(\boldsymbol{y}) - \nabla g_l(\boldsymbol{y}'))\right\|_{\mathcal{Y}^*} \leq L\|\boldsymbol{y}-\boldsymbol{y}'\|_{\mathcal{Y}} \text{ for all } \boldsymbol{x}\in\mathcal{X}. \tag{3.116}$$

Here, we still assume $n\geq d$, so that for this primal-linear setting, the linear variable is of smaller dimension than the nonlinear dimension. We also maintain Assumption 3.2.1. We still pursue a bound of the form

$$\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}^\star\|_{\mathcal{X}}^2 \geq C(1-\alpha)\|\boldsymbol{x}_0-\hat{\boldsymbol{x}}^\star\|_{\mathcal{X}}^2,$$

where C is an absolute constant and $\alpha = \min\left\{\sqrt{\mu\nu}/G, \nu/L\right\}$, where $\hat{\boldsymbol{x}}^\star$ arbitrarily close approximation of the optimal primal solution $\boldsymbol{x}^\star$. Consequently, the parameter regime for this lower bound to be tight is $\nu G^2 \geq \mu L^2$.

To lay the groundwork, we use an analogous algorithm class to Definition 3.7.1.

Definition 3.7.2 (Primal-Linear Proximal Algorithm Class)**.** We define a (primal-linear) *determistic proximal gradient algorithm* to be a sequence of primal-dual iterates $(\boldsymbol{x}_k,\boldsymbol{y}_k)_{k\geq 0}$ satisfying $(\boldsymbol{x}_k,\boldsymbol{y}_k)\in\mathcal{H}_k^{\mathcal{X}}\times\mathcal{H}_k^{\mathcal{Y}}$, where $\mathcal{H}_k^{\mathcal{X}}\subseteq\mathcal{X}$ and $\mathcal{H}_k^{\mathcal{Y}}\subseteq\mathcal{Y}$ are constructed from the following

steps. First, for $k \geq 0$, define the subspaces

$$\mathcal{G}_k^{\mathcal{X}} = \operatorname{span}\left\{g(\boldsymbol{y}_i) : 0 \leq i \leq k\right\} \subseteq \mathcal{X}^*, \tag{3.117}$$

$$\mathcal{G}_k^{\mathcal{Y}} = \operatorname{span}\left\{\nabla g(\boldsymbol{y}_i)^\top \boldsymbol{x}_i : 0 \leq i \leq k+1\right\} \subseteq \mathcal{Y}^*. \tag{3.118}$$

Define $\mathcal{H}_0^{\mathcal{X}} = \operatorname{span}\{\boldsymbol{x}_0\}$ and $\mathcal{H}_0^{\mathcal{Y}} = \operatorname{span}\{\boldsymbol{y}_0\}$, and for $k \geq 1$, we recursively define

$$\mathcal{H}_k^{\mathcal{X}} = \operatorname{span}\left(\left\{p_{\eta_k \phi}(\tilde{\boldsymbol{x}}_{k-1}, \tilde{\boldsymbol{g}}_{k-1}) : \tilde{\boldsymbol{x}}_{k-1} \in \mathcal{H}_{k-1}^{\mathcal{X}}, \tilde{\boldsymbol{g}}_{k-1} \in \mathcal{G}_{k-1}^{\mathcal{X}}\right\} \cup \mathcal{H}_{k-1}^{\mathcal{X}}\right)$$
$$\mathcal{H}_k^{\mathcal{Y}} = \operatorname{span}\left(\left\{p_{\gamma_k \psi}(\tilde{\boldsymbol{y}}_{k-1}, \tilde{\boldsymbol{f}}_k) : \tilde{\boldsymbol{y}}_{k-1} \in \mathcal{H}_{k-1}^{\mathcal{Y}}, \tilde{\boldsymbol{f}}_{k-1} \in \mathcal{G}_{k-1}^{\mathcal{Y}}\right\} \cup \mathcal{H}_{k-1}^{\mathcal{Y}}\right).$$

We also use the same exact hard instance as before, namely

$$\mu(\boldsymbol{x}, \boldsymbol{y}) = \underbrace{\left\langle \tfrac{G}{2} \boldsymbol{A}^\top \boldsymbol{y}, \boldsymbol{x} \right\rangle}_{\langle g(\boldsymbol{y}), \boldsymbol{x} \rangle} - \underbrace{\frac{\nu}{2} \|\boldsymbol{y}\|_2^2}_{\psi(\boldsymbol{y})} + \underbrace{\boldsymbol{c}^\top \boldsymbol{x} + \frac{\mu}{2} \|\boldsymbol{x}\|_2^2}_{\phi(\boldsymbol{x})}, \tag{3.119}$$

with Euclidean proximal oracles. In this case, just as before, we have for all $k \geq 1$,

$$\boldsymbol{x}_k = \frac{\eta_k, \boldsymbol{x}_{k-1} - \bar{\boldsymbol{g}}_{k-1}}{\mu + \eta_k} \subseteq \operatorname{span}\left\{\boldsymbol{x}_{k-1}, \boldsymbol{c}, \boldsymbol{A}^\top \boldsymbol{y}_{k-1}, \boldsymbol{A}^\top \boldsymbol{y}_{k-2}, \ldots, \boldsymbol{A}^\top \boldsymbol{y}_0\right\} \tag{3.120}$$

$$\boldsymbol{y}_k = \frac{\gamma_k, \boldsymbol{y}_{k-1} - \bar{\boldsymbol{f}}_{k-1}}{\nu + \gamma_k} \subseteq \operatorname{span}\left\{\boldsymbol{y}_{k-1}, \boldsymbol{A}\boldsymbol{x}_k, \boldsymbol{A}\boldsymbol{x}_{k-1}, \ldots, \boldsymbol{A}\boldsymbol{x}_0\right\}. \tag{3.121}$$

As a result, we have that Lemma 3.7.3 and Corollary 3.7.1 follow in this setting as well, as they are fully determined by the equations displayed above. After this, the argument follows *exactly* to Section 3.7.2. This results in the following claim.

Proposition 3.7.2. *Consider any algorithm satisfying Definition 3.7.2 with* $\boldsymbol{x}_0 = \boldsymbol{0}_d$ *and* $\boldsymbol{y}_0 = \boldsymbol{0}_n$ *applied to the problem* (3.119)*, with* $n \geq d \geq k$*. Consider* $\alpha = \frac{4\mu\nu}{G^2} \leq 1$*. We have that*

$$\operatorname{Gap}^{\boldsymbol{x}^\star, \boldsymbol{y}^\star}(\boldsymbol{x}_k, \boldsymbol{y}_k) \geq \frac{\mu}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 \geq \frac{\mu}{8\sqrt{2}} \left(1 - \sqrt{\alpha}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - O((1-\alpha)^{2d+2})/\alpha)$$

Proof. Follow an identical proof to Proposition 3.7.1. □

Having established preliminary lower bounds for deterministic algorithms, extending

these constructions to the setting in which L/μ comprises the dominant term in the objective would be a natural direction for future research. We discuss this and other directions in Section 3.8.

3.8 Perspectives & Future Work

This chapter presented a deep dive into the semilinear min-max problem, which interpolates between two problem classes that have been fertile ground for optimization research for decades: the bilinearly and nonbilinearly-coupled min-max problems. We employed a constructive convergence analysis and derived algorithms for the convex-concave, strongly convex-strongly concave, strongly convex-concave, and convex-strongly concave in a unified manner. Key ideas were adaptive sampling, which made use of possibly non-uniform Lipschitz and smoothness constants among the component functions $f_1, \ldots, f_n$, and a historical regularization term on the primal update that can achieve complexity improvements in even special cases such as bilinear problems.

Because the focus of this work was on deriving upper bounds on complexity, there are opportunities to derive tight lower bounds for both deterministic and stochastic settings. The constructions generally differ for stochastic methods (e.g., that of Woodworth and Srebro [2016]) from the Nesterov tri-diagonal approach taken in Section 3.7. Furthermore, our lower bound only applies to particular parameter regimes (when $G/\sqrt{\mu\nu} \gtrsim L/\mu$). To extend this analysis to arbitrary parameter regimes, we hypothesize that our high-dimensional quadratic objective would be replaced by a Lipschitz continuous and smooth function, such as a multivariate analog of the Huber loss, such as

$$\ell_\delta(\boldsymbol{x}) = \begin{cases} \frac{1}{2}\|\boldsymbol{x}\|_2^2 & \|\boldsymbol{x}\|_2 \leq \delta \\ \delta\left(\|\boldsymbol{x}\|_2 - \frac{1}{2}\delta\right) & \|\boldsymbol{x}\|_2 > \delta \end{cases}.$$

An interesting extension from both theoretical and algorithmic viewpoints is the possibility to solve a generic nonbilinearly-coupled optimization problem by alternating between 1) linearizing the objective in the dual variables, and 2) solving the resulting dual linear

min-max problem. Fixing the strongly convex-strongly concave setting for concreteness, and recalling the formulation (3.1), one may consider $(\bar{\boldsymbol{x}}_k, \bar{\boldsymbol{y}}_k)$ a solution of

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \ \langle \boldsymbol{y}, \nabla_{\boldsymbol{y}} c(\boldsymbol{x}, \bar{\boldsymbol{y}}_{k-1})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x}), \tag{3.122}$$

and inspect the limiting value as $k \to \infty$.

Returning to the outline of Chapter 1, we devoted significant attention to the topics introduced in Section 1.2.2. While the relationship between distributional robustness and convex optimization (e.g., statistical learning with linear models) allows for elegant mathematical analyses, it is also critical to consider highly expressive models such as neural networks (as noted in Section 2.12). Furthermore, core statistical questions regarding generalization remain, such as the optimal selection of training data or the resulting sample complexity of modern learning methods. Both points are addressed in the upcoming chapter using the backdrop of Section 1.2.3).

Reference	Global Complexity
Nesterov and Scrimali [2006, Theorem 3]	$O\left(nd\frac{\sqrt{L^2+G^2}}{\mu\wedge\nu}\ln(1/\varepsilon)\right)$
Wang and Li [2020, Theorem 3]	$\tilde{O}\left(nd\sqrt{\frac{L}{\mu}+\frac{G\cdot\max\{L,G\}}{\mu\nu}}\ln(1/\varepsilon)\right)$
Lin et al. [2020, Theorem 9] Borodich et al. [2024, Theorem 3]	$\tilde{O}\left(nd\sqrt{\frac{L^2+G^2}{\mu\nu}}\ln(1/\varepsilon)\right)$
Carmon et al. [2022, Theorem 3] Lan and Li [2023, Theorem 4.2]	$O\left(nd\sqrt{\frac{L^2+G^2}{\mu\nu}}\ln(1/\varepsilon)\right)$
Kovalev and Gasnikov [2022, Theorem 3]	$O\left(nd\sqrt{\frac{L^2+G^2}{\mu(\mu\wedge\nu)}}\ln(1/\varepsilon)\right)$
Jin et al. [2022, Theorem 1] Li et al. [2023, Corollary 3.4]	$\tilde{O}\left(nd\left(\frac{L}{\mu}+\frac{G}{\sqrt{\mu\nu}}\right)\ln(1/\varepsilon)\right)$
This work (Theorem 3.3.1)	$O\left(nd\left(\frac{L}{\mu}+\frac{G}{\sqrt{\mu\nu}}\right)\ln(1/\varepsilon)\right)$

Table 3.2: **Complexity Bounds for General Nonbilinarly-Coupled Objectives for $\mu,\nu>0$.** Runtime or global complexity (i.e. the total number of elementary operations required to compute $(\boldsymbol{x},\boldsymbol{y})$ satisfying $\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{y})\leq\varepsilon$ for fixed $(\boldsymbol{u},\boldsymbol{v})\in\mathcal{X}\times\mathcal{Y}$. The methods considered call the entire list of primal first-order oracles $(f_j,\nabla f_j)$ for $j=1,\ldots,n$ on each iteration. The method of Kovalev and Gasnikov [2022, Corollary 1] achieves its claim by swapping the role of $\boldsymbol{x}$ and $\boldsymbol{y}$, which is not possible for (3.2). Carmon et al. [2022, Theorem 3] requires a bounded diameter assumption on $\mathcal{Y}$, which is generally required for $\boldsymbol{L}_1,\ldots,\boldsymbol{L}_N$ to be finite if each of $f_1,\ldots,f_n$ is nonlinear.

Reference	Additional Structure	Global Complexity (big-$\tilde{O}$)
Alacaoglu and Malitsky [2022, Corollary 6] Cai et al. [2024, Theorem 4.2]	Constants known	$\frac{n(d+N)}{\sqrt{N}\varepsilon}\left\|\boldsymbol{\lambda}\right\|_1$
Alacaoglu and Malitsky [2022, Corollary 6] Cai et al. [2024, Theorem 4.2] Pichugin et al. [2024, Corollary 1]		$\frac{n(d+N)}{\varepsilon}\left\|\boldsymbol{\lambda}\right\|_2$
Diakonikolas [2025, Thm. 1 & Eq. (38)]	Constants known + Separable	$\frac{n(d+N)}{N\varepsilon}\left\|\boldsymbol{\lambda}\right\|_{1/2}$ $\frac{nd}{N\varepsilon}\left\|\boldsymbol{\lambda}\right\|_{1/2}$
Diakonikolas [2025, Thm. 1 & Eq. (36)]	 + Separable	$\frac{n(d+N)}{\varepsilon}\sqrt{N}\left\|\boldsymbol{\lambda}\right\|_2$ $\frac{nd}{\varepsilon}\sqrt{N}\left\|\boldsymbol{\lambda}\right\|_2$
This work(Theorem 3.4.1) **This work (Theorem 3.5.1)**	Constants known + Separable	$\frac{n(d+N)}{N\varepsilon}\left(\left\|\boldsymbol{\lambda}\right\|_1^{1/2}(\left\|\boldsymbol{\lambda}\right\|_1^{1/2}+\left\|\boldsymbol{G}\right\|_1^{1/2})\right)$ $\frac{nd}{\sqrt{N}\varepsilon}\left(\left\|\boldsymbol{\lambda}\right\|_1^{1/2}(\left\|\boldsymbol{L}\right\|_1^{1/2}+\left\|\boldsymbol{G}\right\|_\infty^{1/2})\right)$
This work (Theorem 3.4.2) **This work (Theorem 3.5.2)**	Constants known + Separable	$\frac{n(d+N)}{N\varepsilon}\left(\left\|\boldsymbol{\lambda}\right\|_{1/2}^{1/2}(\left\|\boldsymbol{\lambda}\right\|_{1/2}^{1/4}\left\|\boldsymbol{L}\right\|_{1/2}^{1/4}+\left\|\boldsymbol{G}\right\|_{1/2}^{1/2})\right)$ $\frac{nd}{\sqrt{N}\varepsilon}\left(\left\|\boldsymbol{\lambda}\right\|_{1/2}^{1/4}(\left\|\boldsymbol{L}\right\|_{1/2}^{3/4}+\left\|\boldsymbol{G}\right\|_\infty^{3/4})\right)$

Table 3.3: **Complexity Bounds for Convex-Concave Finite-Sum Objectives.** Arithmetic or global complexity (i.e., the total number of elementary operations required to compute $(\boldsymbol{x},\boldsymbol{y})$ satisfying $\mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{y})]\leq\varepsilon$ for fixed $(\boldsymbol{u},\boldsymbol{v})\in\mathcal{X}\times\mathcal{Y}$. We use $\boldsymbol{\lambda}$ as defined in Section 3.2. The objective is assumed to be **convex-concave** and have a finite sum structure $\mathcal{L}(\boldsymbol{x},\boldsymbol{y})=\sum_{J=1}^{N}\mathcal{L}_J(\boldsymbol{x},\boldsymbol{y})$. The expectation is taken over any randomness incurred by the algorithm.

Reference	Additional Structure	Global Complexity (big-$\tilde{O}$)
Palaniappan and Bach [2016, Theorem 2]	Constants known	$\frac{n(n+d)}{N}\left(N+\frac{L^2+G^2}{(\mu\wedge\nu)^2}\right)\ln(1/\varepsilon)$
Alacaoglu and Malitsky [2022, Corollary 27] Cai et al. [2024, Theorem 4.6]	Constants known	$\frac{n(d+N)}{N}\left(N+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1}{\mu\wedge\nu}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Alacaoglu and Malitsky [2022, Corollary 27] Cai et al. [2024, Theorem 4.6]		$\frac{n(d+N)}{N}\left(N+\frac{N\|\boldsymbol{\lambda}\|_2}{\mu\wedge\nu}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Diakonikolas [2025, Thm. 1 & Eq. (38)]	Constants known + Separable	$\frac{n(d+N)}{N}\left(N+\frac{\|\boldsymbol{\lambda}\|_{1/2}}{\mu\wedge\nu}\right)\ln\left(\frac{1}{\varepsilon}\right)$ $\frac{nd}{N}\left(N+\frac{\|\boldsymbol{\lambda}\|_{1/2}}{\mu\wedge\nu}\right)\ln\left(\frac{1}{\varepsilon}\right)$
Diakonikolas [2025, Thm. 1 & Eq. (36)]	+ Separable	$\frac{n(d+N)}{N}\left(N+\frac{N^{3/2}\|\boldsymbol{\lambda}\|_2}{\mu\wedge\nu}\right)\ln\left(\frac{1}{\varepsilon}\right)$ $\frac{nd}{N}\left(N+\frac{N^{3/2}\|\boldsymbol{\lambda}\|_2}{\mu\wedge\nu}\right)\ln\left(\frac{1}{\varepsilon}\right)$
This work(Theorem 3.4.1) **This work (Theorem 3.5.1)**	Constants known + Separable	$\frac{n(d+N)}{N}\left(N+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1}{\mu}+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{G}\|_1^{1/2}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$ $\frac{nd}{N}\left(N+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{L}\|_1^{1/2}}{\mu}+\frac{\sqrt{N}\|\boldsymbol{\lambda}\|_1^{1/2}\|\boldsymbol{G}\|_\infty^{1/2}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$
This work (Theorem 3.4.2) **This work (Theorem 3.5.2)**	Constants known + Separable	$\frac{n(d+N)}{N}\left(N+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{3/4}\|\boldsymbol{L}\|_{1/2}^{1/4}}{\mu}+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{1/2}\|\boldsymbol{G}\|_{1/2}^{1/2}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$ $\frac{nd}{N}\left(N+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\|\boldsymbol{L}\|_{1/2}^{3/4}}{\mu}+\frac{\|\boldsymbol{\lambda}\|_{1/2}^{1/4}\|\boldsymbol{G}\|_\infty^{3/4}}{\sqrt{\mu\nu}}\right)\ln\left(\frac{1}{\varepsilon}\right)$

Table 3.4: **Complexity Bounds for Strongly Convex-Strongly Concave Finite-Sum Objectives**: Arithmetic or global complexity (i.e., the total number of elementary operations required to compute $(\boldsymbol{x},\boldsymbol{y})$ satisfying $\mathbb{E}[\mathrm{Gap}^{\boldsymbol{u},\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{y})]\leq\varepsilon$ for fixed $(\boldsymbol{u},\boldsymbol{v})\in\mathcal{X}\times\mathcal{Y}$. We use $\boldsymbol{\lambda}$ as defined in Section 3.2. The objective is assumed to be (μ,ν)-**strongly convex-strongly concave** and have a finite sum structure $\mathcal{L}(\boldsymbol{x},\boldsymbol{y})=\sum_{J=1}^{N}\mathcal{L}_J(\boldsymbol{x},\boldsymbol{y})$. The expectation is taken over any randomness in the algorithm.

Chapter 4

GENERALIZATION CAPABILITIES OF ZERO-SHOT PREDICTION

4.1 Introduction

In Chapter 2 and Chapter 3, we studied statistical and algorithmic aspects of out-of-distribution generalization from the perspective of linear models and convex optimization. While this setting allows for acute statements of algorithm performance, an undisputed driver of modern generalization capabilities—such as zero-shot prediction (ZSP)—is the expressivity of overparametrized neural network models (see Section 1.2.3). Nonetheless, we note that neural networks (in particular, transformers [Vaswani et al., 2017]) merely provide a canvas; the primary driver of performance in both ZSP and language modeling is the data used to pre-train the model [Fang et al., 2023, Gadre et al., 2023, Xu et al., 2024, Li et al., 2024a]. Furthermore, the prompting mechanism (as used in (1.7)), though relatively less-studied, has been seen to be an essential determinant of downstream performance [Pratt et al., 2023]. We now frame our approach to addressing them.

Let us first describe the pre-training phase. Consider random variables X and Z governed by a joint probability measure $P \equiv P_{X,Z}$ over $\mathcal{X} \times \mathcal{Z}$. We may interpret (X, Z) as an image-caption pair. Given this data-generating distribution, we assume access to $(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)$ drawn i.i.d. realization from P with empirical measure $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(\boldsymbol{x}_i, \boldsymbol{z}_i)}$. As mentioned in Section 1.2.1, the pre-training phase involves learning a parameter $\boldsymbol{\theta}$ by minimizing a self-supervised learning (SSL) loss of the form (1.5). Specifically, losses such as CLIP/InfoNCE [van den Oord et al., 2019], VICReg [Bardes et al., 2022], BarlowTwins [Zbontar et al., 2021], and spectral embedding methods [HaoChen et al., 2021, Balestriero

and LeCun, 2022, Tan et al., 2024] can be written in the form

$$\mathcal{R}(\boldsymbol{\theta}, P_n) = \mathbb{E}_{P_n^{\otimes b}} \left[\ell_b(\boldsymbol{\theta}, (X_1, Z_1), \ldots, (X_b, Z_b)) \right], \tag{4.1}$$

where $P_n^{\otimes b}$ denotes sampling b points uniformly randomly with replacement from $\{(\boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^n$ and ℓ_b is a loss function that, at a high level, involves comparison between points in the mini-batch. The specific forms of the loss are not relevant for our main investigation, in which we analyze a prominent method used to design the pre-training set of the original CLIP model [Radford et al., 2021] as well as open-source reproductions [Xu et al., 2024]. In the creation of both models, the scientists first resampled the pre-training set in order to achieve particular marginal distributions over metadata (such as word counts in the text modality). We prove that when these marginal distributions coincide with the true data-generating distribution, linear functionals (such as expected loss) can be estimated with a reduced variance via a non-asymptotic mean squared error bound. The bound decomposes into an $O(1/n)$ variance term, an $\tilde{O}(1/n^2)$ bias term, and an $\tilde{O}(1/n^{3/2})$ cross term. The first-order variance term recovers the asymptotic variance previously discovered by Bickel et al. [1991] by the tools of asymptotic semiparametric efficiency theory. Our bound not only explicitly exposes the higher-order terms non-asymptotically, but also displays a new closed-form formula for the variance term, which was previously described in terms of projections. We hypothesize, and confirm with experimentation, that this procedure may stabilize the loss estimate (4.1) and lead to improved downstream performance of the resulting models.

For the downstream task, the practitioner aims to use the parameter $\boldsymbol{\theta}$ and an auxiliary prompting procedure (described in Section 1.2.3). We prove a basic identity that relates (1.7) to a two-stage regression function (from X to Z to an unseen label Y). The same identity allows us to describe the fundamental limit of ZSP, i.e., how close its performance may approximate the Bayes optimal predictor for the downstream task, in terms of a conditional dependence relationship on (X, Z, Y).

The analyses above are formalized and elaborated upon in Section 4.2. Results regarding (upstream) data curation through marginal matching and the improved estimation of

linear functionals are contained in Section 4.3. The connection between ZSP and two-stage regression is proved, and its implications are given in Section 4.4. Experimental results that validate our viewpoints on both improving downstream performance with marginal-corrected pre-training datasets and optimized prompting are contained in Section 4.5. Extensions and future work are discussed in Section 4.6 and Section 4.7.

4.2 Preliminaries

Recall that we used $P_{X,Z}$ to denote the joint probability measure of the image-caption pairs in the pre-training set (e.g. arbitrary content from the Internet that is not necessarily associated with any task). For a downstream task, we consider an additional random variable Y realized in a (possibly non-discrete) set $\mathcal{Y}$, indicating a category or class label. We define a joint distribution $P^{\mathcal{T}} \equiv P^{\mathcal{T}}_{X,Z,Y}$ governing the data from the downstream task (e.g., image classification on CIFAR-10). Crucially, notice that we consider a random variable Z governed by $P^{\mathcal{T}}$, representing a hypothetical (but unobserved) caption for the image being classified. Instead, this latent variable will be useful for defining a theoretical counterpart of the ZSP procedure (1.7). The first part of this chapter will only concern the pre-training phase and $P_{X,Z}$, whereas the second part will consider the downstream task and $P^{\mathcal{T}}_{X,Z,Y}$. We outline the main approaches below.

4.2.1 Curating Pre-Training Data

Firstly, to understand the effects of data curation on pre-training, we formally analyze the marginal matching method of Radford et al. [2021], Xu et al. [2024] introduced in Section 4.1. For this example, first assume that $\mathcal{X}$ and $\mathcal{Z}$ are finite sets with $|\mathcal{X}| = m$ and $|\mathcal{Z}| = l$, as the method is applied to discretized versions of the original data in practice. The general approach is to select two user-defined marginal distributions P_X on $\mathcal{X}$ and P_Z on $\mathcal{Z}$ and "adjust" the empirical measure P_n so that it marginalizes to (P_X, P_Z) in either variable. In our statistical analysis, we assume that P_X and P_Z are in fact the true marginal distributions of the data-generating distribution P. Under a frequentist viewpoint, this captures a

setting in which high-quality, paired examples are relatively scarce or expensive (i.e., sampling or integration under P is "hard"), but unpaired examples are so abundant that one can estimate the marginals with negligible error. This was the historical justification for the setting, studied as early as 1940 in the context of cross-tabulated data for census applications [Deming and Stephan, 1940, Ireland and Kullback, 1968]. From a Bayesian viewpoint, the marginals P_X and P_Z can encode prior information, inductive bias, or side information that a distribution estimator may incorporate for improved statistical accuracy [Miller and han Liu, 2002]. The marginal matching done in the works above is equivalent to estimating P using some number of steps of the following recursion: we define $P_n^{(0)} = P_n$ as the empirical measure, and for $k \geq 1$ construct

$$P_n^{(k)}(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} \cdot P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ odd} \\ \frac{P_Z(\boldsymbol{z})}{P_{n,Z}^{(k-1)}(\boldsymbol{z})} \cdot P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ even} \end{cases}. \tag{4.2}$$

The conditions under which the iterations of (4.2) are well-defined are given in Section 4.3. These operations reduce to rescaling the rows of an $(m \times l)$-matrix by $P_X / P_{n,X}^{(k-1)}$ or its columns by $P_Z / P_{n,Z}^{(k-1)}$. This very algorithm has a decades-old history and is known in various contexts as Sinkhorn-Knopp matrix scaling [Sinkhorn, 1967], iterative proportional or bi-proportional fitting [Johnston and Pattie, 1993], and raking ratio estimation [Thompson, 2000]. We hypothesize that the improvements of the data "balancing" procedure (4.2) in large-scale training are derived from the improvement of $P_n^{(k)}$ over P_n as an estimator for P (at a linear functional representing the expected loss). In Section 4.3, we prove non-asymptotic bounds on the mean squared error of the estimator $\mathbb{E}_{P_n^{(k)}}[h(X, Z)]$ for $\mathbb{E}_P[h(X, Z)]$ for an integrable test function $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$. The bound decomposes into an $O(n^{-1})$ first-order variance term and an $\tilde{O}(k^6 n^{-3/2})$ higher-order term. The first-order term is shown to have a strict improvement over the empirical measure baseline with a fine-grained dependence on the spectra of two conditional mean operators associated to P. The higher-order term can be used to compute the asymptotic variance, resolving some efficiency questions originally

raised in Bickel et al. [1991]. Our proof technique relies on a recursion decomposition for balancing-based estimators, which may be of independent interest.

While we focus on the applications of balancing to data curation, we also comment that for specific foundation models, such as CLIP [Radford et al., 2021], SwaV [Caron et al., 2020], and variations on this theme [Jones et al., 2022, Asano et al., 2020] also contain a balancing operation in the computation of their objectives (see Liu et al. [2024, Section 2]). While not originally derived in this way, in the particular case of CLIP, the objective is computed by applying the iterations of (4.2) to an unnormalized measure defined on the elements of the *mini-batch*. In fact, only $k = 1$ iteration is applied to each marginal in the original CLIP objective, whereas Liu et al. [2024] observe performance improvements by increasing the number of iterations and further committing to the viewpoint that the objective implicitly computes (4.2). We identify this intersection of balancing and learning objectives as a valuable line of future work.

4.2.2 Prompting and Downstream Classification

Secondly, we consider the downstream task of predicting Y from X, and specifically, using a predictor of the form (1.7) which relies on pre-trained encoders. Note that we do not use the finiteness assumption on $\mathcal{X}$ or $\mathcal{Z}$ in these results. We identify a population parameter and describe how each component of the pipeline (including prompting) is composed to serve as an estimator of the parameter. Let $f : \mathcal{Y} \to \mathbb{R}$ be a function and consider the problem of predicting $f(Y)$ given $X = \boldsymbol{x}$. The use of f serves only to handle multiple tasks such as classification (e.g., $f(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = 1\}$) and regression ($f(\boldsymbol{y}) = y$) in a unified manner. Now, define the minimum $P^{\mathcal{T}}_{X,Y}$-expected risk (Bayes optimal) predictor as

$$\eta_\star(\boldsymbol{x}) := \mathbb{E}_{P^{\mathcal{T}}_{X,Y}}\left[f(Y)|X\right](\boldsymbol{x}). \tag{4.3}$$

Because ZSP operates by "translating" the image classification problem to a text classification problem via prompting, it is natural to consider the entire procedure (1.7) to be a

finite-sample estimator of the two-stage regression function

$$\eta_\rho(\boldsymbol{x}) := \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[f(Y)|Z\right]|X\right](\boldsymbol{x}), \tag{4.4}$$

where $\rho_{Y,Z}$ is the *prompting* distribution. This probability measure represents the pseudo-captions used to define the predictor (1.7), which include template-based prompts as well as class-conditional prompts (those for which the caption text differs for each class label) drawn from large language models. To support the claim of (4.4) as the underlying estimand mathematically, we prove a key relationship in Section 4.4. Under the assumption that $P_{X,Z} \ll P_X \otimes P_Z$, we define the likelihood ratio (or Radon-Nikodym derivative) $\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) = \left(\frac{\mathrm{d}P_{X,Z}}{\mathrm{d}(P_X \otimes P_Z)}\right)(\boldsymbol{x}, \boldsymbol{z})$. Then, it holds that

$$\eta_\rho(\boldsymbol{x}) = \mathbb{E}_{\rho_{Y,Z}}\left[f(Y) \cdot \mathsf{R}(\boldsymbol{x}, Z)\right] + \mathrm{err}(P_Z, \rho_Z), \qquad P_X\text{-almost surely}, \tag{4.5}$$

where $\mathrm{err}(P_Z, \rho_Z)$ measures the disagreement between the distribution of the pre-training captions and the captions drawn from the prompting distribution determined by the user. To understand the significance of this result, recall the encoders $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ introduced in (1.7). When parameterized via $\boldsymbol{\theta}$ as $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv (\boldsymbol{\alpha}_{\boldsymbol{\theta}}, \boldsymbol{\beta}_{\boldsymbol{\theta}})$, they are produced by optimizing the self-supervised learning (SSL) objective (4.1) over a large pre-training set. Several works (see Gutmann and Hyvärinen [2012], Oko et al. [2025] and references therein) have established that popular SSL objectives such as the CLIP and Noise Contrastive Estimation achieve the exact population minimum when $\langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{z}) \rangle = \log \mathsf{R}(\boldsymbol{x}, \boldsymbol{z})$, assuming it exists. This form (4.5) can be related to (1.7), as is done in Section 4.4, connecting an otherwise mysterious empirical procedure to a well-defined estimand with clear limits. In other words, we are able to use the two-stage prediction form (4.4) to comment on the mathematical limits of prompting as a replacement for downstream training data in Section 4.4.

4.3 Non-Asymptotic Analysis of Variance Reduction

We expand on the problem setup introduced in Section 4.2.1, in which we consider sample spaces $(\mathcal{X}, \mathcal{Z})$, along with true and unknown joint distribution P on $\mathcal{X} \times \mathcal{Z}$ with known marginals (P_X, P_Z). Recall the ERM notation of Section 1.2.1. For an integrable test function $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, we introduce the empirical process notation $P(h) := \mathbb{E}_P[h(X, Z)]$. We then let $\Theta = \mathbb{R}$ and define the squared error risk functional

$$\mathcal{R}(\boldsymbol{\theta}, P) = (\boldsymbol{\theta} - P(h))^2.$$

Thus, considering both the empirical measure P_n and the marginal-rebalanced empirical measure $P_n^{(k)}$ from (4.2), we define

$$\boldsymbol{\theta}_n := P_n(h) = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}(\boldsymbol{\theta}, P_n) \text{ and } \boldsymbol{\theta}_n^{(k)} := P_n^{(k)}(h) = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}(\boldsymbol{\theta}, P_n^{(k)}) \tag{4.6}$$

We present theoretical guarantees on the mean squared error (MSE) of the data-balanced estimator $\boldsymbol{\theta}_n^{(k)}$ (taken over the randomness in the sample) and highlight relevant points in the proofs. We first give context on the main innovations of the analysis and then outline its high-level steps. These innovations include relating the nonlinear iterations of balancing over probability measures to linear operators on a vector space and using a singular value decomposition of these operators to quantify their effect after a finite number of iterations. Furthermore, by scaling the number of iterations appropriately, we can characterize the estimator using the limit of balancing iterations, which is an object of interest in applications including optimal transport.

We make the following assumption throughout, which is usually satisfied by the desired marginals P_X and P_Z in practice: the target marginals $P_X(\boldsymbol{x}) > 0$ and $P_Z(\boldsymbol{z}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$. The iterative procedure given in (4.2) is visualized in Figure 4.1 (left). The iterations are well-defined for all k under the event that $\text{Supp}(P_{n,X}) = \text{Supp}(P_X)$ and $\text{Supp}(P_{n,Z}) = \text{Supp}(P_Z)$, i.e., all observed row counts and column counts are non-empty.[1]

[1]Due to this technical consideration, we define $P_n^{(k)}$ to be the empirical measure P_n when this condition

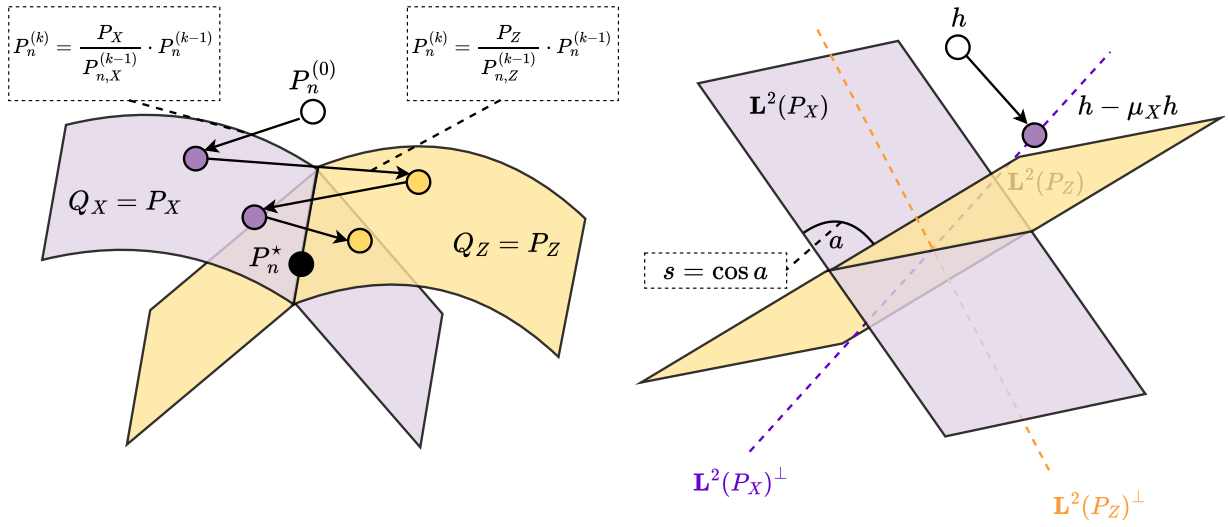

Figure 4.1: **Data Balancing.** Nonlinear and linear operators associated with each iteration of (4.2). **Left:** Visualization of the exact iterations of (4.2) in the space of probability measures. The purple set contains joint distributions with $\mathcal{X}$-marginal equal to P_X, whereas the golden set contains joint distributions with $\mathcal{Z}$-marginal equal to P_Z. **Right:** Visualization of $\mathbf{L}^2(P)$, the operators defining (4.9), and the singular values given in (4.11).

To provide background, the scheme of alternating the operators (4.2) is often seen as an iterative algorithm to solve the problem

$$\min_{Q \in \Pi(P_X, P_Z)} \mathrm{KL}(Q \| P_n), \tag{4.7}$$

where $\Pi(P_X, P_Z)$ denotes the set of probability measures on $\mathcal{X} \times \mathcal{Z}$ that marginalize to P_X and P_Z in each variable and $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. The iterations (4.2) are based on the alternating minimization approach of solving

$$P_n^{(k)}(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} \arg\min_{\{Q: Q_X = P_X\}} \mathrm{KL}(Q \| P_n^{(k-1)}) & k \text{ odd} \\ \arg\min_{\{Q: Q_Z = P_Z\}} \mathrm{KL}(Q \| P_n^{(k-1)}) & k \text{ even} \end{cases},$$

which inspires the viewpoint of balancing as alternating *information projections*. As we

is not satisfied, which we show occurs with low probability. See Appendix B.3.4 for details.

show in Appendix B.2, the iterations of (4.2) can equivalently be defined using the KL, reverse KL, or χ^2-divergences. This viewpoint is relevant as previously, efforts have been made (e.g., in Bickel et al. [1991]) to analyze the variance reduction afforded by the solution to (4.7) directly. However, quantifying the variance reduction (in terms of properties of P) using this approach is challenging, as there is no closed-form expression for the solution of (4.7). A key mathematical outcome of our analysis is that the closed-form expressions of the projections (4.2) can be used to compute the reduction in mean squared error at each iteration. Thus, by letting $k \equiv k(n) \to \infty$ (scaled appropriately against n), we can determine the reduction for the solution of (4.7) for large n. This is the subject of Theorem 4.3.1.

From Information Projections to Orthogonal Projections First, we will show that the variance reduction resulting from each nonlinear iteration of (4.2) is associated with a linear operator applied to h. Thus, instead of analyzing the alternating information projections over probability measures, we may use familiar tools to understand alternating orthogonal projections in a vector space. To define them, we first let $\mathbf{L}^2(P)$ to be the set of functions $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ satisfying $\mathbb{E}_P\left[h^2(X, Z)\right] < \infty$. Even though $\mathcal{X} \times \mathcal{Z}$ is finite, working within $\mathbf{L}^2(P)$ will be analytically convenient. Let $\mathbf{L}^2(P_X)$ be the subspace of $\mathbf{L}^2(P)$ containing functions that only depend on the first argument $\boldsymbol{x} \in \mathcal{X}$ and define $\mathbf{L}^2(P_Z)$ analogously. These are the solid-colored subspaces in Figure 4.1 (right). Next, let $\mu_X : \mathbf{L}^2(P) \to \mathbf{L}^2(P_X)$ and $\mu_Z : \mathbf{L}^2(P) \to \mathbf{L}^2(P_Z)$ be defined as, for any $h \in \mathbf{L}^2(P)$,

$$\mu_X h = \operatorname*{arg\,min}_{f \in \mathbf{L}^2(P_X)} \mathbb{E}_P\left[(h(X, Z) - f(X))^2\right] \implies [\mu_X h](\boldsymbol{x}, \boldsymbol{z}) := \mathbb{E}_P\left[h(X, Z)|X\right](\boldsymbol{x})$$

The operator μ_X is an orthogonal projection onto $\mathbf{L}^2(P_X)$. The orthogonal projection operator μ_Z onto $\mathbf{L}^2(P_Z)$ is defined analogously. We may also define the conditional centering/debiasing operators $\mathcal{C}_X = I - \mu_X$ and $\mathcal{C}_Z = I - \mu_Z$, which each project onto the orthogonal complements of $\mathbf{L}^2(P_X)$ and $\mathbf{L}^2(P_Z)$, visualized as subspaces with dotted lines in Figure 4.1 (right). To understand the importance of the conditional mean and debiasing operators, we give a recursive formula that forms the backbone of our analysis. Define $\mu_k = \mu_X$ for k odd

and $\mu_k = \mu_Z$ for k even, and define $\mathcal{C}_k$ similarly. Thus, we have by linearity of expectation that

$$\begin{aligned}[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mathcal{C}_k h) + \overbrace{[P_n^{(k)} - P](\mu_k h)}^{=0} \\ &= [P_n^{(k-1)} - P](\mathcal{C}_k h) + [P_n^{(k)} - P_n^{(k-1)}](\mathcal{C}_k h) \\ &= \underbrace{[P_n^{(0)} - P](\mathcal{C}_1 \ldots \mathcal{C}_k h)}_{\text{first-order term}} + \underbrace{\textstyle\sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\mathcal{C}_\ell \ldots \mathcal{C}_k h)}_{\text{higher-order terms}}. \end{aligned} \tag{4.8}$$

To justify the first line, we discuss the case when k is odd. Notice that $\mu_X h$ is only a function of X, so its expectation only depends on P_X that is equal to $P_{n,X}^{(k)}$ (the $\mathcal{X}$-marginal of $P_n^{(k)}$) by (4.2). The last line follows by unrolling the previous step $k-1$ times. This recursive expansion is proven formally in Proposition B.3.1 in Appendix B.3. Given the expansion, the mean squared error can be computed by taking the expectation of the square of (4.8). We show that the second moment of the first-order term in (4.8) is equal to σ_k^2/n where

$$\sigma_0^2 := \mathbb{V}\mathrm{ar}(h) \text{ and } \sigma_k^2 := \mathbb{V}\mathrm{ar}(\mathcal{C}_1 \ldots \mathcal{C}_k h) \text{ for } k \geq 1, \tag{4.9}$$

and all other terms are $\tilde{O}(k^6 n^{-3/2})$. Thus, by exactly computing the constant in the dominating term, we may quantify the asymptotic variance reduction. Our first main result concerns the higher-order terms and shows that it is indeed dominated by the first-order term. Note that the empirical mean $\boldsymbol{\theta}_n^{(0)} = \frac{1}{n}\sum_{i=1}^n h(X_i, Z_i)$ is unbiased, and so its MSE is equal to σ_0^2/n. Define in addition

$$p_\star := \min\{\min_{\boldsymbol{x}} P_X(\boldsymbol{x}), \min_{\boldsymbol{z}} P_Z(\boldsymbol{z})\}$$

which measures the non-uniformity of the target marginals. We have that $p_\star$ is positive because both P_X and P_Z are positive. We now state the first main result.

Theorem 4.3.1. *For a sequence of data balancing estimators $(\boldsymbol{\theta}_n^{(k)})_{k\geq 1}$ as defined in (4.6), there exists an absolute constant $C > 0$ and distribution dependent constants $s \in [0, 1)$ and $\sigma_{gap}^2 \geq 0$, such that the following holds: For $n \geq C[\log_2(2n/p_\star) + m\log(n+1)]/p_\star^2$ and $k \geq 1$,*

we have

$$\mathbb{E}_P\left[(\boldsymbol{\theta}_n^{(k)} - \boldsymbol{\theta})^2\right] \leq \frac{\sigma_0^2 - \sigma_{gap}^2}{n} + O\left(\frac{s^k}{n}\right) + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right). \tag{4.10}$$

Furthermore, $\sigma_{gap}^2 > 0$ *except for the pathological case of* $\mu_X h$ *being a constant function.*

The quantities σ_{gap}^2 and s are quantified toward the end of this section and are dependent on singular decays of the conditional mean operators for each variable under P. Showing Theorem 4.3.1 boils down to showing that the higher-order term in (4.8) is $O(n^{-1})$ with high probability. Using the expression (4.2) and assuming that $\ell \geq 1$ is odd, we see that

$$[P_n^{(\ell)} - P_n^{(\ell-1)}](\mathcal{C}_\ell \ldots \mathcal{C}_k h) = \sum_{\boldsymbol{x},\boldsymbol{z}} \left[\frac{P_X(\boldsymbol{x})}{P_{n,X}^{(\ell-1)}(\boldsymbol{x})} - 1\right] \cdot [\mathcal{C}_\ell \ldots \mathcal{C}_k h](\boldsymbol{x},\boldsymbol{z}) P_n^{(\ell-1)}(\boldsymbol{x},\boldsymbol{z}).$$

The first (blue) term in the product quantifies the disagreement between the $\mathcal{X}$-marginal of $P_n^{(\ell-1)}$ and the true marginal, which can be bounded in terms of $\mathrm{KL}(P_{n,X}^{(0)} \| P_X)$ and is shown to be $O(n^{-1/2})$ with high probability via techniques from information theory. The second term can be unrolled recursively in a similar fashion to (4.8) itself, which will consequently be $\tilde{O}(n^{-1/2})$ as well; this is the most technical part of the analysis (see Appendix B.3.3). We also discuss various extensions such as balancing with misspecified marginals and handling continuous data; see Section 4.6.1 and Section 4.6.2.

Given Theorem 4.3.1, a natural next step is to quantify the gap between σ_0^2 and σ_k^2, which requires finer-grained properties of $\mathcal{C}_X$ and $\mathcal{C}_Z$. Notably, we show that as $k \to \infty$, σ_k^2 approaches the limiting value $\sigma^2 - \sigma_{\text{gap}}^2$. Thus, via (4.10), by using $k = o(n^{1/12})$ (excluding logarithmic terms for simplicity), one obtains the asymptotic variance of the solution to (4.7). This contrasts with Albertus and Berthet [2019], in which the dependence of a quantity similar to (4.10) is exponential in k, meaning that $k = o(\log(n))$ is required for convergence under this argument.

From Orthogonal Projections to Variance Reduction We now clarify what is precisely meant by the "spectrum" of the conditional mean operators μ_X and μ_Z. As proven

using a *singular value decomposition* (Proposition B.1.1) in Appendix B.1.1, there exists a basis $\{\alpha_j\}_{j=1}^m$ of $\mathbf{L}^2(P_X)$, a basis $\{\beta_j\}_{j=1}^m$ of $\mathbf{L}^2(P_Z)$, and real values $\{s_j\}_{j=1}^m$, that satisfy

$$\mu_Z \alpha_j = s_j \beta_j \text{ and } \mu_X \beta_j = s_j \alpha_j \text{ for } j \in \{1, \ldots, m\}. \tag{4.11}$$

Furthermore, $\alpha_1 = \mathbf{1}_{\mathcal{X}}$ and $\beta_1 = \mathbf{1}_{\mathcal{Z}}$ leading to the equality $\langle f, \alpha_1 \rangle_{\mathbf{L}^2(P_X)} = \mathbb{E}_{P_X}[f(X)]$. Finally, $s_1 = 1$ and s_j is non-negative and non-increasing in j. For a concrete example, consider $m = 2$, in which case P can be written as a matrix in $\mathbb{R}^{2\times 2}$ and elements of $\mathbf{L}^2(P_X)$ and $\mathbf{L}^2(P_X)$ are vectors in $\mathbb{R}^2$. Then, in the case of uniform marginals, we can verify directly that (4.11) can be satisfied by setting

$$\alpha_1 = \beta_1 = \begin{bmatrix} +1 \\ \\ +1 \end{bmatrix}, \alpha_2 = \beta_2 = \begin{bmatrix} +1 \\ \\ -1 \end{bmatrix}, \text{ and } P = \frac{1}{4}\begin{bmatrix} 1+s & 1-s \\ \\ 1-s & 1+s \end{bmatrix} \tag{4.12}$$

for $s = s_2$ (the second largest singular value). Thus, as $s \to 1$, the distribution becomes "fully dependent" as Z and X are completely determined by one another. As $s \to 0$, P approaches the product measure. Geometrically, because $\alpha_1 = \beta_1$, we know that the angle a between the subspaces $\mathbf{L}^2(P_X)$ and $\mathbf{L}^2(P_Z)$ is given by the angle between α_2 and β_2. By computing their inner product in $\mathbf{L}^2(P)$, we have that $\langle \alpha_2, \beta_2 \rangle_{\mathbf{L}^2(P)} = \langle P, \alpha_2 \beta_2^\top \rangle = s = \cos a$. Thus, $s = 0$ indicates orthogonality of these subspaces, alluding to the independence of X and Z (see the right panel of Figure 4.1).

Returning to $m \geq 2$, we consider the following as a sufficient condition for variance reduction: the operators μ_X and μ_Z have a positive spectral gap, i.e., $s_2 < s_1$. Note that this assumption is satisfied when $P(\boldsymbol{x}, \boldsymbol{z}) > 0$ for all $(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}$ by the Perron–Frobenius Theorem [Horn and Johnson, 2013, Chapter 8]. Using the intuition from Figure 4.1, this rules out pathological cases such as Z being a deterministic function of X. Under the spectral gap condition, the singular values $\{s_j\}_{j=2}^m$ that are strictly less than 1 will determine a geometric rate of decay in variance given in Corollary 4.3.1. The left and right singular functions $\alpha_j : \mathcal{X} \to \mathbb{R}$ and $\beta_j : \mathcal{Z} \to \mathbb{R}$ will define a useful coordinate system to represent projections

of h when analyzing $\boldsymbol{\theta}_n^{(k)}$.

Indeed, let $\bar{h} = P(h)$ be the centered test function. Because $\mu_X \bar{h} \in \mathbf{L}^2(P_X)$ and $\mu_Z \bar{h} \in \mathbf{L}^2(P_Z)$, we may decompose this function on the two bases to write

$$\mu_X \bar{h} = \sum_{j=1}^{m} u_j \alpha_j \quad \text{and} \quad \mu_Z \bar{h} = \sum_{j=1}^{m} v_j \beta_j. \tag{4.13}$$

Corollary 4.3.1 below relates the (normalized) variance σ_k^2 of the first-order term to the variance of the sample mean $\boldsymbol{\theta}_n^{(0)}$. In fact, it shows that the variance reduction $\sigma_0^2 - \sigma_k^2$ decays geometrically to the quantity

$$\sigma_{\text{gap}}^2 := \sum_{j=2}^{m} \left[\frac{u_j^2 + v_j^2 - 2 s_j u_j v_j}{1 - s_j^2} \right].$$

For simplicity, we only present the result for k even, i.e., σ_{2t}^2.

Corollary 4.3.1. *The variance reduction achieved by $t+1$ iterations of the $\mathcal{C}_Z \mathcal{C}_X$ operator can be quantified as*

$$\sigma_0^2 - \sigma_{2(t+1)}^2 = \sigma_{gap}^2 - \sum_{j=2}^{m} \frac{s_j^2 (v_j - s_j u_j)^2}{1 - s_j^2} s_j^{4t} = \sum_{j=2}^{m} \left[u_j^2 + (1 - s_j^{4t+2}) \frac{(v_j - s_j u_j)^2}{1 - s_j^2} \right].$$

Intuitively, the operators $\mathcal{C}_X$ and $\mathcal{C}_Z$ are the main sources of the variance reduction via orthogonality. Since $\alpha_1 = \mathbf{1}_{\mathcal{X}}$, we can see that the reduction will always be strictly positive as long as $\mu_X \bar{h}$ is not a constant function. Finally, using $s := s_2 \geq s_j$ for $j \geq 2$ gives the second term in Theorem 4.3.1.

4.4 Theoretical Limits of Zero-Shot Prediction

We now consider the questions raised in Section 4.2.2. Rather than introducing particular estimators and considering their sample complexity, we focus on deriving a form that relates ZSP to the Bayes optimal predictor on the downstream task, giving an outline to analyze many possible estimation procedures. Before entering into the details, we briefly comment on how our analysis compares to similar theory for few-shot learning (FSL).

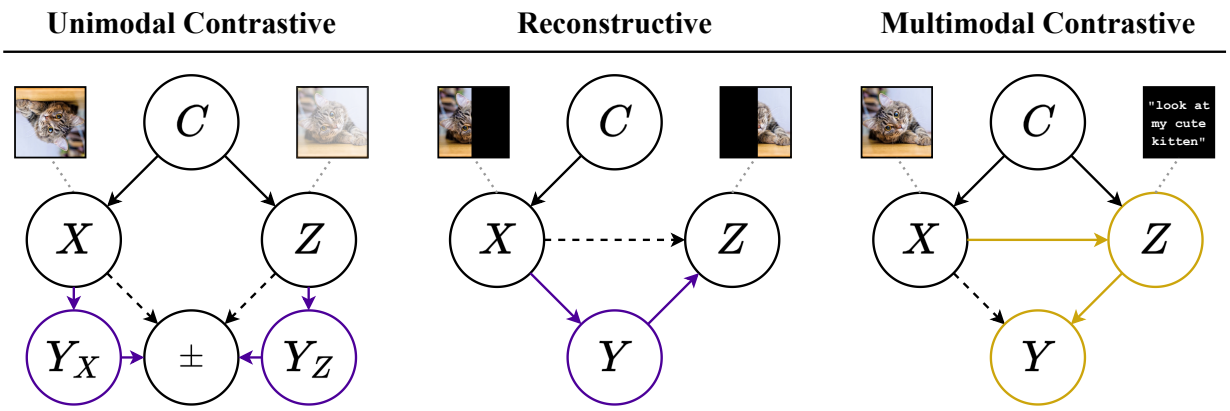

Figure 4.2: **Graphical Models of Self-Supervised Prediction Paths.** Each directed graphical model corresponds to the data types and dependence structures for various SSL pre-training approaches. The variable C represents an unobserved context that determines the observed data-generating distribution. Dotted lines indicate the possibility of presence or absence of the arrow. Methods that are compatible with FSL learn the label Y as a latent variable in the process of solving the pretext task. Methods compatible with ZSP may learn the relationship between X and Z directly, whereas the relationship between Z and Y is estimated via prompting.

The prevailing techniques of recent FSL theory are visualized in Figure 4.2. In unimodal contrastive learning (left), X and Z are augmented/corrupted images, and the pretext task is to identify examples derived from the same ("+") or different ("−") underlying image [Chen et al., 2020]. In reconstructive SSL (center), the encoder is pre-trained to predict a hidden portion of the raw/embedded image [Assran et al., 2023]. If the dotted arrows were absent, the only path to solve the pretext task *is* through the label, from which generalization guarantees follow. This motivates another prevalent assumption of exact/approximate conditional independence of X and Z given Y (e.g., as in Lee et al. [2021]). We also avoid this assumption, which is unrealistic in the multimodal context as the dependence between an image and its caption is unlikely to be fully explained by a coarse label such as "cat from CIFAR-10". In summary, FSL methods learn *through* the label during pre-training. For multimodal contrastive learning (Figure 4.2, right), the ideal dependence structure is fundamentally different.

For illustrative purposes, let us first consider the "compatible" case, that is, when $P_{X,Z} = P^{\mathcal{T}}_{X,Z}$ and $\rho_{Y,Z} = P^{\mathcal{T}}_{Y,Z}$, recalling that $P^{\mathcal{T}}_{X,Y,Z}$ is the evaluation distribution augmented with its latent caption. If X and Y are conditionally independent given Z, then η_ρ (from (4.4)) is in fact identical to $\eta_\star$ (from (4.3)) because

$$\begin{aligned}\eta_\rho &= \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[f(Y)|Z\right]|X\right] \\ &= \mathbb{E}_{P^{\mathcal{T}}_{X,Z}}\left[\mathbb{E}_{P^{\mathcal{T}}_{Y,Z}}\left[f(Y)|Z\right]|X\right] \\ &= \mathbb{E}_{P^{\mathcal{T}}_{X,Z}}\left[\mathbb{E}_{P^{\mathcal{T}}_{X,Y,Z}}\left[f(Y)|Z,X\right]|X\right] = \eta_\star\end{aligned}$$

by the tower property of conditional expectation under $P^{\mathcal{T}}_{X,Y,Z}$. Thus, one naturally interprets ZSP as learning *toward* the label using prompting as the essential connection between Z and Y. For this reason, we will quantify the difference between (4.3) and (4.4) based on a conditional dependence measure (see (4.18) below), which formalizes the information-theoretic cost of using natural language as a proxy for image classification. We dub this the *residual dependence* between X and Y that is unexplained by Z. In addition to this, we quantify the incompatibility of the evaluation, pre-training, and prompting distributions, as $P_{X,Z} \neq P^{\mathcal{T}}_{X,Z}$ and $\rho_{Y,Z} \neq P^{\mathcal{T}}_{Y,Z}$ in general. We argue that this taxonomy of SSL methods based on their ideal graphical model is useful not only for understanding them conceptually but also for analyzing them mathematically.

We first present the precise description of the identity (4.5) (reproduced below in (4.15)). We then demonstrate the connection between the right-hand side of (4.15) with the empirical ZSP (1.7). Having validated the viewpoint that ZSP is a form of two-stage prediction (i.e., the left-hand side of (4.15)), we use this viewpoint to bound the $\mathbf{L}^2(P^{\mathcal{T}}_X)$ distance between η_ρ and $\eta_\star$. Observe the following, which relies on basic measure-theoretic manipulations (see Appendix B.4).

Lemma 4.4.1. *Assume that the pre-training distribution $P_{X,Z}$ satisfies: (i) $P_{X,Z} \ll P_X \otimes P_Z$ with Radon-Nikodym derivative $\mathsf{R} = \left(\frac{dP_{X,Z}}{d(P_X \otimes P_Z)}\right)$, and (ii) that f is integrable under $P^{\mathcal{T}}_Y$. We*

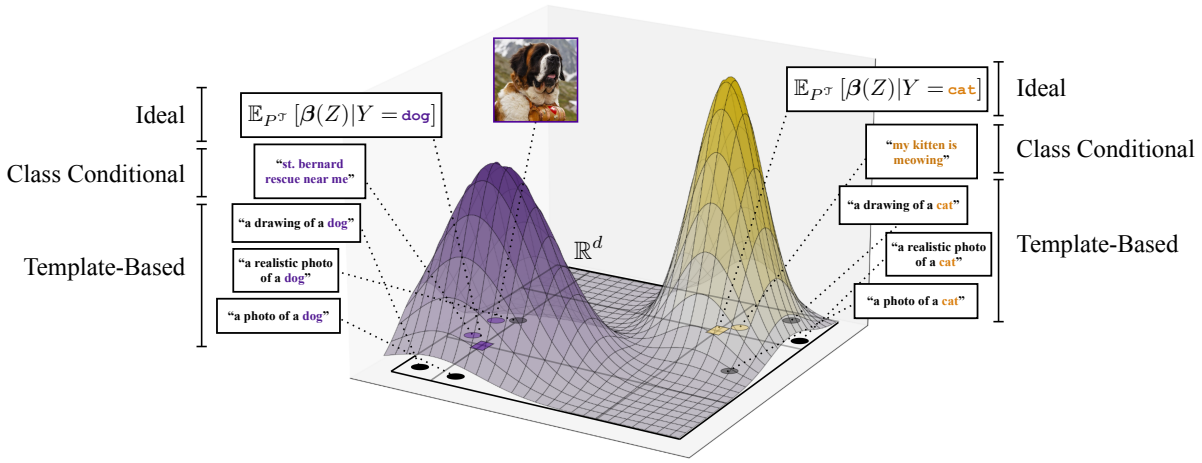

Figure 4.3: **Illustration of Prompting Approaches.** A hypothetical distribution of embeddings $\boldsymbol{\beta}(Z)$ parametrized by two classes ("cat" and "dog"). Three prompting strategies (template-based, class-conditional, and idealized) are shown with example text and resulting embeddings in $\mathbb{R}^d$. Prompt bias is the distance of the average of the circular points to the square target points.

may then define the function

$$g_\rho(\boldsymbol{z}) := \mathbb{E}_{\rho_{Y,Z}}[f(Y)|Z](\boldsymbol{z}). \tag{4.14}$$

Then, P_X-almost surely, we have the equality

$$\eta_\rho(\boldsymbol{x}) = \mathbb{E}_{\rho_{Y,Z}}[f(Y) \cdot \mathsf{R}(\boldsymbol{x}, Z)] + \underbrace{\int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\,(\,dP_Z(\boldsymbol{z}) - \,d\rho_Z(\boldsymbol{z}))}_{\mathrm{err}(P_Z, \rho_Z)}. \tag{4.15}$$

Proof. By Lemma B.4.1, we already have that for P_X-almost all $\boldsymbol{x} \in \mathcal{X}$, the identity

$$\begin{aligned}
\eta_\rho(\boldsymbol{x}) &= \mathbb{E}_{P_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] \\
&= \mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] + \mathbb{E}_{P_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] - \mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] \\
&= \mathbb{E}_{\rho_Z}[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)] + \int_{\mathcal{Z}} g_\rho(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\,(\,\mathrm{d}P_Z(\boldsymbol{z}) - \,\mathrm{d}\rho_Z(\boldsymbol{z}))\,.
\end{aligned}$$

Now, unpacking the first term on the right-hand side above, we recognize that for fixed

$\boldsymbol{x} \in \mathcal{X}$, the random variable $\mathsf{R}(\boldsymbol{x}, Z)$ is $\sigma(Z)$-measurable, so via the properties of conditional expectation [Schilling, 2017, Theorem 27.11 (vii)] in $\mathbf{L}^1(\rho_Z)$, we may write

$$\mathbb{E}_{\rho_Z}\left[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)\right] = \mathbb{E}_{\rho_Z}\left[\mathbb{E}_{\rho_{Y,Z}}\left[f(Y)|Z\right]\mathsf{R}(\boldsymbol{x}, Z)\right] = \mathbb{E}_{\rho_Z}\left[\mathbb{E}_{\rho_{Y,Z}}\left[f(Y)\mathsf{R}(\boldsymbol{x}, Z)|Z\right]\right].$$

Using the expression above and the tower property of the conditional expectation, we write

$$\mathbb{E}_{\rho_Z}\left[g_\rho(Z)\mathsf{R}(\boldsymbol{x}, Z)\right] = \mathbb{E}_{\rho_Z}\left[\mathbb{E}_{\rho_{Y,Z}}\left[f(Y)\mathsf{R}(\boldsymbol{x}, Z)|Z\right]\right] = \mathbb{E}_{\rho_{Y,Z}}\left[f(Y)\mathsf{R}(\boldsymbol{x}, Z)\right],$$

completing the proof. □

The Radon-Nikodym derivative R appearing in Lemma 4.4.1 is a fundamental quantity known as the *information density* in the information theory literature[2]. It is equal to unity almost surely if and only if X and Z are independent under $P_{X,Z}$. For this reason, its deviation from unity constitutes a dependence measure which we employ later in this section.

Proceeding, we relate the first term of (4.15) to the procedure (1.7). Recall from Section 4.2.2 that the encoders $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ optimize popular loss functions such as CLIP if $\langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{z})\rangle = \log \mathsf{R}(\boldsymbol{x}, \boldsymbol{z})$, or equivalently, $\exp\langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{z})\rangle = \mathsf{R}(\boldsymbol{x}, \boldsymbol{z})$. To now make the connection between the right-hand side of (4.15) and (1.7), consider the setting of binary classification ($\mathcal{Y} = \{0, 1\}$) in which ρ_Y places equal weight on each class. Define $\eta_\rho^{(0)}$ and $\eta_\rho^{(1)}$ as instances of (4.4) when setting $f(\boldsymbol{y}) \equiv f^{(0)}(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = 0\}$ and $f(\boldsymbol{y}) \equiv f^{(1)}(\boldsymbol{y}) = \mathbb{1}\{\boldsymbol{y} = 1\}$, respectively. First, assume that $P_Z = \rho_Z$ (there is no change between the pre-training and prompting distribution, marginally over Z). Then we may then classify $\boldsymbol{x}$ as the value of $\boldsymbol{y}$ that maximizes $\eta_\rho^{(\boldsymbol{y})}(\boldsymbol{x})$. This classifier is exactly equivalent to

$$\arg\max_{\boldsymbol{y} \in \{0,1\}} \mathbb{E}_{\rho_{Y,Z}}\left[\exp\langle \boldsymbol{\alpha}(\boldsymbol{x}), \boldsymbol{\beta}(Z)\rangle \,|Y = \boldsymbol{y}\right], \tag{4.16}$$

[2]This term actually refers to $(\boldsymbol{x}, \boldsymbol{z}) \mapsto \log \mathsf{R}(\boldsymbol{x}, \boldsymbol{z})$, but for simplicity, we use it for R—see Dytso et al. [2023, Eq. (11)].

which can be seen by directly calculating

$$\begin{aligned}
&\mathbb{E}_{\rho_{Y,Z}}\left[f^{(1)}(Y)\cdot \mathsf{R}(\boldsymbol{x},Z)\right]\\
&=\mathbb{E}_{P_Y^{\mathcal{T}}}\left[\mathbb{E}_{\rho_{Y,Z}}\left[f^{(1)}(Y)\cdot \mathsf{R}(\boldsymbol{x},Z)|Y\right]\right]\\
&=\frac{1}{2}\mathbb{E}_{\rho_{Y,Z}}\left[f^{(1)}(Y)\cdot \mathsf{R}(\boldsymbol{x},Z)|Y=1\right]+\frac{1}{2}\mathbb{E}_{\rho_{Y,Z}}\left[f^{(1)}(Y)\cdot \mathsf{R}(\boldsymbol{x},Z)|Y=0\right]\\
&=\frac{1}{2}\mathbb{E}_{\rho_{Y,Z}}\left[\mathsf{R}(\boldsymbol{x},Z)|Y=1\right]
\end{aligned}$$

and using $\mathsf{R}(\boldsymbol{x},\boldsymbol{z})=\exp\langle\boldsymbol{\alpha}(\boldsymbol{x}),\boldsymbol{\beta}(\boldsymbol{z})\rangle$. The expression (4.16) not only resembles (1.7) down to an exponentiation of the inner product but also, by relaxing the assumption that $P_Z=\rho_Z$, we may observe exactly where the effect of distribution shift arises, that is, the introduction of error in the coverage of the prompts.

We now formalize the manner in which we measure the conditional dependence and "incompatibility" of the evaluation distribution $P^{\mathcal{T}}_{X,Y}$, pre-training distribution $P_{X,Z}$, and prompting distribution $\rho_{Y,Z}$. To do so, we will need to make several mild regularity conditions on $P_{X,Y,Z}$. We use the notion of regular conditional distribution, or r.c.d. (Definition B.4.1), introduced in Appendix B.4.

Assumption 4.4.1. The joint probability $P^{\mathcal{T}}_{X,Y,Z}$ on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$ satisfies the following constraints.

- **Agrees jointly with the evaluation distribution:** For all measurable sets $A\subseteq\mathcal{X}\times\mathcal{Y}$, we have that $P^{\mathcal{T}}_{X,Y,Z}(A\times\mathcal{Z})=P^{\mathcal{T}}_{X,Y}(A)$ (i.e., $P^{\mathcal{T}}_{X,Y,Z}$ agrees with the given marginal $P^{\mathcal{T}}_{X,Y}$).

- **Agrees conditionally with the pre-training distribution:** There exists a measurable set $\mathcal{X}_1\subseteq\mathcal{X}$ with $P^{\mathcal{T}}_X(\mathcal{X}_1)=1$ such that the regular conditional distributions $P_{Z|X=\boldsymbol{x}}$ and $P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$ on $\mathcal{Z}$ exist. Furthermore, these satisfy $P_{Z|X=\boldsymbol{x}}=P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$ for all $\boldsymbol{x}\in\mathcal{X}_1$.

- **Regularity of conditional distributions:** There exists a measurable set $\mathcal{Z}_1\subseteq\mathcal{Z}$

with $P_Z^{\mathcal{T}}(\mathcal{Z}_1) = 1$ such that the regular conditional distributions $P_{X,Y|Z=\boldsymbol{z}}^{\mathcal{T}}$ on $\mathcal{X} \times \mathcal{Y}$ exists for all $\boldsymbol{z} \in \mathcal{Z}_1$. Furthermore, we have the absolute continuity relation $P_{X,Y|Z=\boldsymbol{z}}^{\mathcal{T}} \ll P_{X|Z=\boldsymbol{z}}^{\mathcal{T}} \otimes P_{Y|Z=\boldsymbol{z}}^{\mathcal{T}}$ with Radon-Nikodym derivative

$$\mathsf{S}_{\boldsymbol{z}} := \frac{\mathrm{d}P_{X,Y|Z=\boldsymbol{z}}^{\mathcal{T}}}{\mathrm{d}(P_{X|Z=\boldsymbol{z}}^{\mathcal{T}} \otimes P_{Y|Z=\boldsymbol{z}}^{\mathcal{T}})}, \tag{4.17}$$

that satisfies $\mathbb{E}_{P_{X,Y|Z=\boldsymbol{z}}^{\mathcal{T}}}[\mathsf{S}_{\boldsymbol{z}}(X,Y)] < \infty$ for each $\boldsymbol{z} \in \mathcal{Z}_1$ and $\mathbb{E}_{P_{X,Y,Z}^{\mathcal{T}}}[\mathsf{S}_Z(X,Y)] < \infty$.

That $P_{X,Y,Z}^{\mathcal{T}}$ marginalizes to $P_{X,Y}^{\mathcal{T}}$ is more of an axiomatic property than an assumption, but we phrase it as so to emphasize that $P_{X,Y,Z}^{\mathcal{T}}$ is meant to describe the evaluation distribution. The assumption that the conditionals $P_{Z|X}$ and $P_{Z|X}^{\mathcal{T}}$ match almost surely represents the viewpoint that, after fixing an image $\boldsymbol{x}$, the latent caption $Z|X = \boldsymbol{x}$ follows the same relationship to $\boldsymbol{x}$ as seen during pre-training. Importantly, this does not require or imply that $P_X^{\mathcal{T}} = P_X$ or that $P_Z^{\mathcal{T}} = P_Z$. The marginal distribution $P_X^{\mathcal{T}}$ is supplied entirely by the evaluation distribution $P_{X,Y}^{\mathcal{T}}$, as for any measurable set $A \subseteq \mathcal{X}$, we have by definition that $P_X^{\mathcal{T}}(A) = P_{X,Y}^{\mathcal{T}}(A \times \mathcal{Y})$. On the other hand, the marginal $P_Z^{\mathcal{T}}$ can be defined using the Markov transition kernel $P_{Z|X=\boldsymbol{x}}^{\mathcal{T}}$, in that for any measurable $B \subseteq \mathcal{Z}$, it holds that

$$P_Z^{\mathcal{T}}(B) := \int_{\mathcal{X}_1} P_{Z|X=\boldsymbol{x}}^{\mathcal{T}}(B)\,\mathrm{d}P_X^{\mathcal{T}}(\boldsymbol{x}) = \int_{\mathcal{X}_1} P_{Z|X=\boldsymbol{x}}(B)\,\mathrm{d}P_X^{\mathcal{T}}(\boldsymbol{x}).$$

Finally, the absolute continuity condition, i.e., the existence of (4.17), rules out degeneracies such as Y being a deterministic function of X given $Z = \boldsymbol{z}$ (outside of a set of $P_Z^{\mathcal{T}}$-measure zero). This allows us to define the conditional dependence measure

$$I(X;Y|Z=\boldsymbol{z}) = \mathbb{E}_{P_{X|Z=\boldsymbol{z}}^{\mathcal{T}} \otimes P_{Y|Z=\boldsymbol{z}}^{\mathcal{T}}}\left[(1 - \mathsf{S}_{\boldsymbol{z}}(Y,X))^2\right], \tag{4.18}$$

which may be known in other contexts as the (square of the) conditional *mean square contingency* [Rényi, 1959, Eq. (13)]. This is the residual dependence measure that appears in the upcoming Theorem 4.4.1.

It is also worth pointing out that the first two conditions Assumption 4.4.1 do not contradict one another. For example, one can consider $P_{X,Y,Z}^{\mathcal{T}}$ that satisfies the Markov chain

$Y \to X \to Z$, where (X, Y) is drawn according to $P^{\mathcal{T}}_{X,Y}$, and Z and Y are conditionally independent given X. Then, informally, we have that $P^{\mathcal{T}}_{Z|X,Y} = P^{\mathcal{T}}_{Z|X} = P_{Z|X}$, so $P^{\mathcal{T}}_{X,Y,Z}$ is uniquely determined. While this example implies the existence of a valid joint probability measure $P^{\mathcal{T}}_{X,Y,Z}$, it is also, in a sense, showcasing the "least desirable" distribution for zero-shot prediction, as the dependence between X and Z does not provide any additional information about Y.

Next, we describe how to measure the incompatibility of the evaluation, pre-training, and prompting distributions, which will lead to a *prompt bias* term in the bound. Define the function

$$g_{P^{\mathcal{T}}_{Y,Z}}(\boldsymbol{z}) = \mathbb{E}_{P^{\mathcal{T}}_{Y,Z}}\left[f(Y)|Z\right](\boldsymbol{z}).$$

The prompt bias compares $g_{P^{\mathcal{T}}_{Y,Z}}$ to g_ρ from (4.14). This reflects the notion that if $P^{\mathcal{T}}_{X,Y,Z}$ agrees with two of the three fundamental distributions governing the problem, it may not be able to agree with the prompt distribution $\rho_{Y,Z}$ in general. Observe the following.

Theorem 4.4.1. *Assume that f is bounded in absolute value by B_f. Under Assumption 4.4.1, it holds that Then, it holds that*

$$\|\eta_\rho - \eta_\star\|^2_{\mathbf{L}^2(P^{\mathcal{T}}_X)} \leq 2\underbrace{\|g_\rho - g_{P^{\mathcal{T}}_{Y,Z}}\|^2_{\mathbf{L}^2(P^{\mathcal{T}}_Z)}}_{\textit{prompt bias}} + 2B_f^2\underbrace{\mathbb{E}_{P^{\mathcal{T}}_Z}\left[I(X;Y|Z)\right]}_{\textit{residual dependence}}. \tag{4.19}$$

Proof. We first establish a useful representation of the conditional mean of $f(Y)$ given $X = \boldsymbol{x}$, in terms of the (conditional) information density from Lemma B.4.1. Fix $\boldsymbol{x} \in \mathcal{X}_1$ and $\boldsymbol{z} \in \mathcal{Z}_1$, the sets on which the regular conditional distributions $P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$ and $P^{\mathcal{T}}_{X,Y|Z=\boldsymbol{z}}$ are defined (see Assumption 4.4.1). Because of the existence the Radon-Nikodym derivative $\mathsf{S}_{\boldsymbol{z}}$ from (4.17), we may apply Lemma B.4.1 with $U = Y$, $V = X$, and $h = f$ to write

$$\mathbb{E}_{P^{\mathcal{T}}_{X,Y|Z=\boldsymbol{z}}}\left[f(Y)|X\right](\boldsymbol{x}) = \underbrace{\mathbb{E}_{P^{\mathcal{T}}_{Y|Z=\boldsymbol{z}}}\left[f(Y)\mathsf{S}_{\boldsymbol{z}}(Y,\boldsymbol{x})\right]}_{=:r(\boldsymbol{x},\boldsymbol{z})} \text{ for all } (\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}_1 \times \mathcal{Z}_1.$$

We have denoted the right-hand side by the function $r(\boldsymbol{x}, \boldsymbol{z})$. Integrate both sides over

$P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$, then use the tower property of conditional expectation to achieve

$$\begin{aligned}\eta_\star(\boldsymbol{x}) = \mathbb{E}_{P^{\mathcal{T}}_{X,Y}}\left[f(Y)|X\right](\boldsymbol{x}) &= \int_{\mathcal{Z}} \mathbb{E}_{P^{\mathcal{T}}_{X,Y|Z=\boldsymbol{z}}}\left[f(Y)|X\right](\boldsymbol{x})\,\mathrm{d}P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}(\boldsymbol{z}) \\ &= \int_{\mathcal{Z}} r(\boldsymbol{x},\boldsymbol{z})\,\mathrm{d}P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}(\boldsymbol{z}) \\ &= \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[r(\boldsymbol{x},Z)\right]. \end{aligned} \tag{4.20}$$

Using the identity (4.20) and $P_{Z|X=\boldsymbol{x}} = P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$ on $\boldsymbol{x} \in \mathcal{X}_1$ (Assumption 4.4.1), we write

$$\begin{aligned}&\eta_\rho(\boldsymbol{x}) - \eta_\star(\boldsymbol{x}) \\ &= \mathbb{E}_{P_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z)\right] - \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[r(\boldsymbol{x},Z)\right] \\ &= \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z)\right] - \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[r(\boldsymbol{x},Z)\right] \\ &= \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z) - g_{P^{\mathcal{T}}_{Y,Z}}(Z)\right] + \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[g_{P^{\mathcal{T}}_{Y,Z}}(Z) - r(\boldsymbol{x},Z)\right].\end{aligned}$$

Taking the integral over $P^{\mathcal{T}}_X$, we have by the decomposition above that

$$\begin{aligned}\|\eta_\rho - \eta_\star\|^2_{\mathbf{L}^2(P^{\mathcal{T}}_X)} &= \int_{\mathcal{X}} \left(\eta_\rho(\boldsymbol{x}) - \eta_\star(\boldsymbol{x})\right)^2 \mathrm{d}P^{\mathcal{T}}_X(\boldsymbol{x}) \\ &\le 2\int_{\mathcal{X}_1} \left(\mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[g_\rho(Z) - g_{P^{\mathcal{T}}_{Y,Z}}(Z)\right]\right)^2 \mathrm{d}P^{\mathcal{T}}_X(\boldsymbol{x}) \qquad (4.21) \\ &\quad + 2\int_{\mathcal{X}_1} \left(\mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[g_{P^{\mathcal{T}}_{Y,Z}}(Z) - r(\boldsymbol{x},Z)\right]\right)^2 \mathrm{d}P^{\mathcal{T}}_X(\boldsymbol{x}). \qquad (4.22)\end{aligned}$$

To handle (4.21), we apply Jensen's inequality for each r.c.d. $P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$ to achieve

$$\begin{aligned}&\int_{\mathcal{X}_1} \left(\mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[\left(g_\rho(Z) - g_{P^{\mathcal{T}}_{Y,Z}}(Z)\right)\right]\right)^2 \mathrm{d}P^{\mathcal{T}}_X(\boldsymbol{x}) \\ &\le \int_{\mathcal{X}_1} \mathbb{E}_{P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}}\left[\left(g_\rho(Z) - g_{P^{\mathcal{T}}_{Y,Z}}(Z)\right)^2\right] \mathrm{d}P^{\mathcal{T}}_X(\boldsymbol{x}) \\ &= \mathbb{E}_{P^{\mathcal{T}}_Z}\left[\left(g_\rho(Z) - g_{P^{\mathcal{T}}_{Y,Z}}(Z)\right)^2\right] \\ &= \left\|g_\rho - g_{P^{\mathcal{T}}_{Y,Z}}\right\|^2_{\mathbf{L}^2(P^{\mathcal{T}}_Z)}.\end{aligned}$$

It remains to control (4.22). Applying Jensen's inequality for each r.c.d. $P^{\mathcal{T}}_{Z|X=\boldsymbol{x}}$ once again,

we have that

$$\int_{\mathcal{X}_1} \left(\mathbb{E}_{P^{\mathfrak{I}}_{Z|X=\boldsymbol{x}}} \left[\left(g_{P^{\mathfrak{I}}_{Y,Z}}(Z) - r(\boldsymbol{x}, Z) \right) \right] \right)^2 \mathrm{d}P^{\mathfrak{I}}_X(\boldsymbol{x}) \tag{4.23}$$

$$\leq \int_{\mathcal{X}_1} \left(\mathbb{E}_{P^{\mathfrak{I}}_{Z|X=\boldsymbol{x}}} \left[\left(g_{P^{\mathfrak{I}}_{Y,Z}}(Z) - r(\boldsymbol{x}, Z) \right)^2 \right] \right) \mathrm{d}P^{\mathfrak{I}}_X(\boldsymbol{x})$$

$$= \mathbb{E}_{P^{\mathfrak{I}}_{X,Z}} \left[\left(g_{P^{\mathfrak{I}}_{Y,Z}}(Z) - r(X, Z) \right)^2 \right]$$

$$= \int_{\mathcal{Z}_1} \mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}} \left[\left(g_{P^{\mathfrak{I}}_{Y,Z}}(\boldsymbol{z}) - r(X, \boldsymbol{z}) \right)^2 \right] \mathrm{d}P^{\mathfrak{I}}_Z(\boldsymbol{z}), \tag{4.24}$$

where the last step follows due to the existence of the r.c.d. $P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}$ for $\boldsymbol{z} \in \mathcal{Z}_1$, as $P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}(A) := P^{\mathfrak{I}}_{X,Y|Z=\boldsymbol{z}}(A \times \mathcal{Y})$ for every measurable $A \subseteq \mathcal{X}$, and the latter exists by assumption. Using the definition of $g_{P^{\mathfrak{I}}_{Y,Z}}$, write

$$g_{P^{\mathfrak{I}}_{Y,Z}}(\boldsymbol{z}) - r(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}_{P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}} \left[f(Y)(1 - \mathsf{S}_{\boldsymbol{z}}(Y, \boldsymbol{x})) \right].$$

We may substitute this expression into the integrand of (4.24) and apply Jensen's inequality to $P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}$ to achieve

$$\begin{aligned}
\mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}} \left[\left(g_{P^{\mathfrak{I}}_{Y,Z}}(\boldsymbol{z}) - r(X, \boldsymbol{z}) \right)^2 \right] &= \mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}} \left[\left(\mathbb{E}_{P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}} \left[f(Y)(1 - \mathsf{S}_{\boldsymbol{z}}(Y, X)) \right] \right)^2 \right] \\
&\leq \mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}} \left[\mathbb{E}_{P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}} \left[(f(Y)(1 - \mathsf{S}_{\boldsymbol{z}}(Y, X)))^2 \right] \right] \\
&\leq \|f\|^2_\infty \, \mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}}} \left[\mathbb{E}_{P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}} \left[(1 - \mathsf{S}_{\boldsymbol{z}}(Y, X))^2 \right] \right] \\
&= \|f\|^2_\infty \, \mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}} \otimes P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}} \left[(1 - \mathsf{S}_{\boldsymbol{z}}(Y, X))^2 \right],
\end{aligned}$$

where the final step follows by applying Fubini's theorem [Schilling, 2017, Corollary 14.9] to the product measure $P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}} \otimes P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}$ for fixed $\boldsymbol{z} \in \mathcal{Z}_1$. By the definition (4.18), it holds that

$$\mathbb{E}_{P^{\mathfrak{I}}_{X|Z=\boldsymbol{z}} \otimes P^{\mathfrak{I}}_{Y|Z=\boldsymbol{z}}} \left[(1 - \mathsf{S}_{\boldsymbol{z}}(Y, X))^2 \right] = I(X; Y | Z = \boldsymbol{z}). \tag{4.25}$$

After confirming that (4.25) is $P^{\mathfrak{I}}_Z$-integrable, substituting this expression back into (4.24) achieves the desired result. Expand the quadratic term and apply the Radon-Nikodym

theorem [Schilling, 2017, Theorem 20.1] to achieve

$$\begin{aligned} I(X;Y|Z=\boldsymbol{z}) &= 1 - 2\mathbb{E}_{P^{\mathfrak{T}}_{X|Z=\boldsymbol{z}}\otimes P^{\mathfrak{T}}_{Y|Z=\boldsymbol{z}}}\left[\mathsf{S}_{\boldsymbol{z}}(Y,X)\right] + \mathbb{E}_{P^{\mathfrak{T}}_{X|Z=\boldsymbol{z}}\otimes P^{\mathfrak{T}}_{Y|Z=\boldsymbol{z}}}\left[\mathsf{S}^2_{\boldsymbol{z}}(Y,X)\right] \\ &= 1 - 2\mathbb{E}_{P^{\mathfrak{T}}_{X,Y|Z=\boldsymbol{z}}}\left[1\right] + \mathbb{E}_{P^{\mathfrak{T}}_{X,Y|Z=\boldsymbol{z}}}\left[\mathsf{S}_{\boldsymbol{z}}(Y,X)\right] \\ &= \mathbb{E}_{P^{\mathfrak{T}}_{X,Y|Z=\boldsymbol{z}}}\left[\mathsf{S}_{\boldsymbol{z}}(Y,X)\right] - 1. \end{aligned}$$

Thus, by integrating against $P^{\mathfrak{T}}_Z$, we see that

$$\mathbb{E}_{P^{\mathfrak{T}}_Z}\left[I(X;Y|Z)\right] = \mathbb{E}_{P^{\mathfrak{T}}_{X,Y,Z}}\left[\mathsf{S}_Z(Y,X)\right] - 1,$$

where the expectation term is finite by Assumption 4.4.1. This completes the proof. □

To give context to Theorem 4.4.1, conditional independence relations have previously been used to describe the performance of multimodal contrastive SSL for FSL. We are particularly inspired by the *multi-view redundancy* theory of Tosh et al. [2021], which states informally that the population FSL predictor can approach the performance of the idealized direct predictor that is given *both* X and Z at test time, if $X \perp\!\!\!\perp Y|Z$ and $Z \perp\!\!\!\perp Y|X$ approximately hold. However, the theory of graphical models [Lauritzen, 1996, Proposition 3.1] asserts that both conditional independence relations hold only if $(X,Z) \perp\!\!\!\perp Y$, that is, neither view has information predictive of the label. This can be seen intuitively from Figure 4.2 by breaking the arrows $X \to Y$ and $Z \to Y$. Notice that we compare only to the Bayes optimal predictor (4.3) given X, so that we need only that $X \perp\!\!\!\perp Y|Z$ (i.e., X depends on Y through Z) to close the gap.

We now present several numerical illustrations to validate the hypotheses that inspired the results in both Section 4.3 and Section 4.4. Namely, we wish to see the effects of 1) balancing procedures during pre-training and 2) inference-time prompting on downstream model performance. This includes an example of *unbiased prompting*, or a choice of $\rho_{Y,Z}$ that renders the prompt bias term of (4.19) as zero.

4.5 Experiments

In our experiments, we illustrate the two subjects of this chapter in practice, that is, how data balancing and various prompt strategies (i.e., estimates of $P^{\mathcal{T}}_{Y,Z}$) affect the empirical performance of CLIP models. We focus on image classification tasks as a canonical testbed for ZSP. See Appendix B.5 for experiment details. Code to reproduce the data and experiments can be found at https://github.com/ronakdm/balancing and https://github.com/ronakdm/zeroshot.

Models, Datasets, and Evaluation. Throughout, we consider training variants of the CLIP model, which require a dataset of image-caption pairs. In the experiments that simulate pre-training, we train models on top of embedded representations of images and text. The pre-training set chosen is the ImageNet-Captions dataset [Fang et al., 2023], which pairs images from ImageNet [Deng et al., 2009] that were taken from Flickr with their original captions. For prompting-based experiments, we use three publicly available CLIP models from the OpenCLIP repository [Ilharco et al., 2022]: ResNet50 pre-trained on YFCC15M [Thomee et al., 2016], NLLB-CLIP pre-trained on a subset of LAION COCO [Visheratin, 2023], and ViT-B/32 pre-trained on the DataComp medium pool [Gadre et al., 2023].

We evaluate models based on zero-shot classification performance via top-k accuracy, in which a test example is considered to be classified correctly if the true class is contained within the elements of $\mathcal{T}$ with the k largest scores as computed by (1.7). Our evaluation datasets include five standard benchmarks: the Describable Textures Dataset or DTD [Cimpoi et al., 2014], Flowers 102 [Nilsback and Zisserman, 2008], FGVC Aircraft [Maji et al., 2013], SUN397 [Xiao et al., 2010], and ImageNet-1k [Deng et al., 2009]. For some prompting experiments, we also make use of the ImageNet-Captions dataset as a way to estimate prompt embeddings. Evaluation is done using tools from the CLIP Benchmark repository. In Figure 4.5 and Figure 4.6, "templates" refers to using all default community-curated prompts available in CLIP Benchmark. See Appendix B.5 for specific model tags and full experimental details.

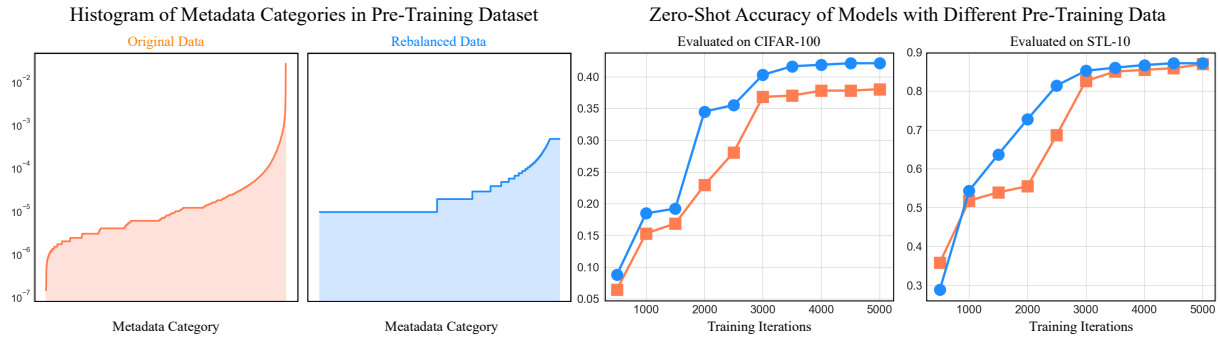

Figure 4.4: **Balancing as Data Curation.** Depiction of balancing and data curation on ImageNet-Captions dataset, in which $\mathcal{X}$ represents image-caption pairs and $\mathcal{Y}$ represents keywords. **Left:** Observed marginal $P_{n,Z}$ (orange) and P_Z (blue), which are sorted by order of increasing probability. **Right:** Zero-shot evaluation of an embedding model trained using the standard CLIP loss on the original versus the balanced training set.

4.5.1 Balanced Pre-Training Effects

We perform a data curation experiment exploring the use of balancing to adjust the entire pre-training set, in the spirit of Xu et al. [2024]. The target marginal P_Z is selected by choosing a threshold for which frequent keywords have their probability mass truncated, and the probability measure is normalized to sum to one. In Figure 4.4, we show the observed marginal $P_{n,Z}$ and the target marginal P_Z sorted in increasing order (left). The original marginal on $\mathcal{Y}$ has approximately 5 orders of magnitude of difference between the most and least probable keyword. After balancing, the target marginal has less than 2 orders of difference between the most frequent and least frequent keywords.. To see how this affects downstream performance, we plot the zero-shot classification accuracy over training iterations in Figure 4.4 (right) when using the original dataset (orange) and using the metadata-balanced dataset (blue). We observe moderate improvement, especially in the small batch regime ($m = 512$) when curating the dataset.

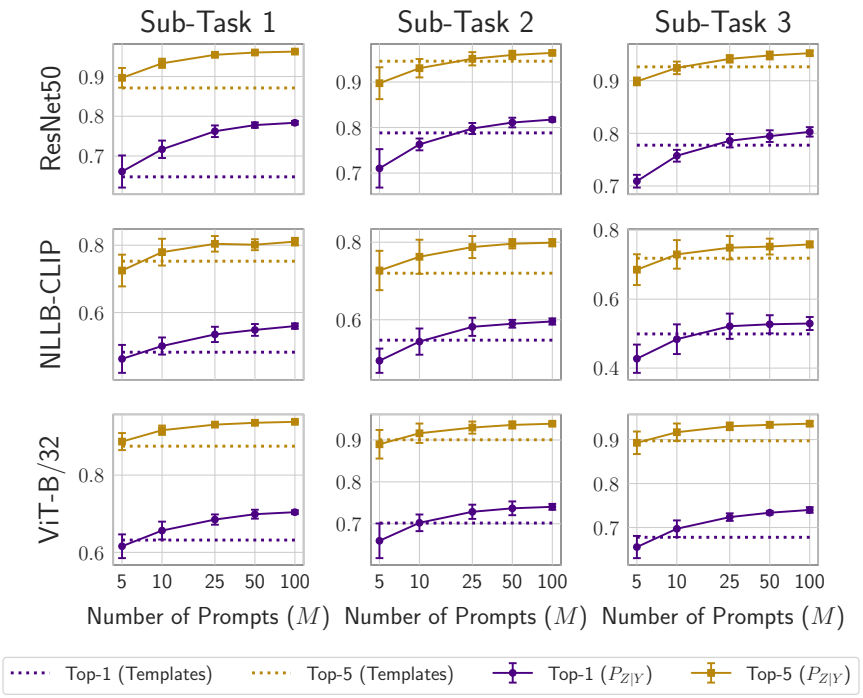

Figure 4.5: **Results: Unbiased Prompting.** Pre-trained models are varied along the rows, and sub-tasks (subsets of 50 ImageNet-1k classes) are varied along columns. In all plots, the x-axis denotes the number of prompts sampled for each class embedding, and the y-axis denotes top-k zero-shot classification accuracy. Error bars indicate standard deviations across 10 seeds for prompt sampling. In this setting $P_{Y,Z} = P^{\mathcal{T}}_{Y,Z}$.

4.5.2 Unbiased Prompting with Observations from $P^{\mathcal{T}}_{Z,Y}$

This is an illustrative experiment in which it is possible to use the unbiased prompting mechanism, that is, directly drawing samples from the distribution $P^{\mathcal{T}}_{Y,Z}$ to approximate the conditional mean $\mathbb{E}_{P^{\mathcal{T}}_{Y,Z}}[\boldsymbol{\beta}(Z)|Y = \boldsymbol{y}]$ (see Figure 4.3). Indeed, because the ImageNet-Captions dataset includes ImageNet images, their Flickr captions, and their label (i.e., the joint observation of (X, Y, Z)), we may directly observe the scaling of the variance with respect to M without prompt bias. We design three in-distribution sub-tasks by randomly selecting collections of 50 classes $(\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3)$ from each of 998 classes, reserving held-out *prompting examples* $(Z_1, Y_1), \ldots, (Z_{15{,}000}, Y_{15{,}000})$, 100 for each of 150 classes. Then, for task i, using M examples $j_1(\boldsymbol{y}), \ldots, j_M(\boldsymbol{y})$ selected randomly without replacement for $\boldsymbol{y} \in$

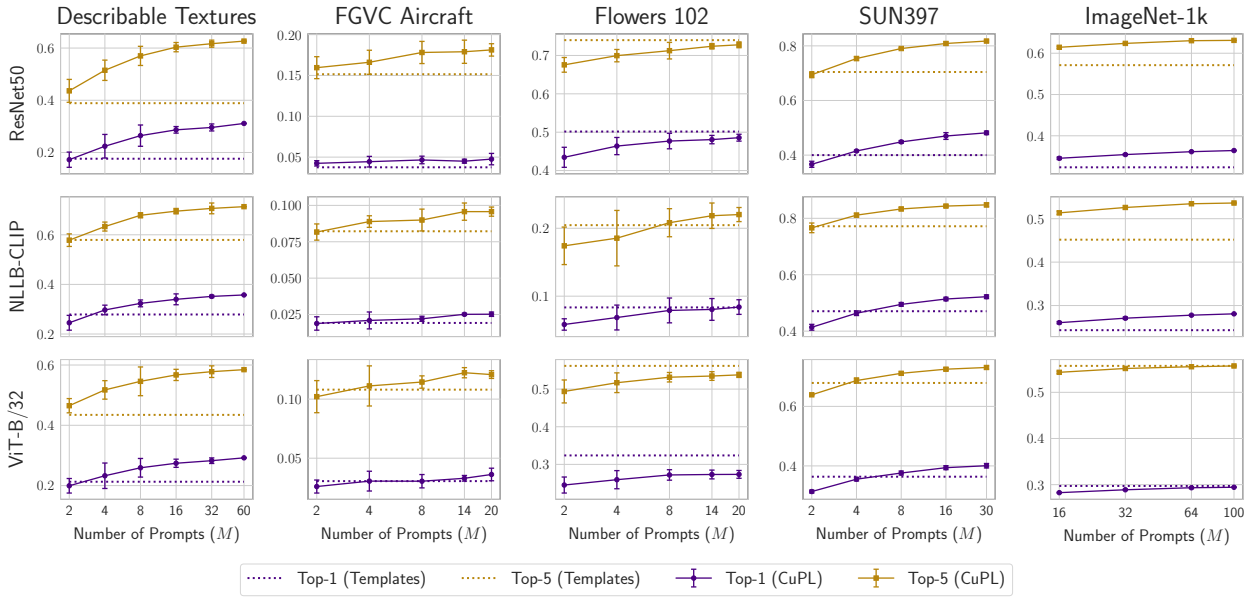

Figure 4.6: **Results: Class-Conditional Prompting.** Pre-trained models are varied along the rows, and evaluation datasets are varied along columns. In all plots, the x-axis denotes the number of prompts sampled for each class embedding, and the y-axis denotes top-k zero-shot classification accuracy. Error bars indicate standard deviations across 10 seeds for prompt sampling.

$\mathcal{Y}_i$, we use the vector $\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\beta}(Z_{j_m(y)})$ as the class embedding (projected to unit norm). Using an evaluation set of approximately 25,000 examples from each sub-task, we compute the classification accuracy of this approach and plot the results in Figure 4.5. Observe that the threshold at which unbiased prompting outperforms the 18 default templates is approximately $M = 10$ across tasks. However, the performance of the theoretically unbiased approach only saturates at $M = 100$ and can have enormous benefits (almost 15% absolute increase in top-1 accuracy for the ResNet50 on Sub-Task 1) in performance. Thus, for models that have not yet been saturated from pre-training, prompting can close surprisingly wide gaps in zero-shot classification accuracy.

4.5.3 Class-Conditional Prompting with Language Models

From both theoretical and empirical investigations, two facts are clear: class-agnostic templates are outperformed by captions that are attuned to the class being evaluated, and a potentially large number of prompts (50-100) is needed to saturate performance, as opposed to the 5-15 defaults used in OpenCLIP. Framed as so, this is a superb use case for modern large language models (LLMs), as they can generate class-aware captions with high throughput. As mentioned in Section 4.1, we investigate CuPL as a means to implement class-conditional prompting with LLMs. Our experimental setup and scientific goals differ from those used in Pratt et al. [2023]: 1) we use lightweight encoders that have not saturated their performance during pre-training, as opposed to their use of the large-scale ViT-L/14 architecture, 2) we quantify the variability of classification accuracy with respect to prompting by generating up to fifty times as many prompts per experiment, and 3) we employ LlaMA 3, which is free and accessible to other, as opposed to GPT-3 [Brown et al., 2020]. The results are shown in Figure 4.6, where we order the datasets in increasing number of classes per task: 47, 100, 102, 397, and 998. Similar phenomena as in the unbiased case are observed, although the approximate saturation threshold varies per dataset from 20 for Flowers 102 and FGVC Aircraft up to 60 for DTD. Note that the choice of defaults heavily influences the baseline performance. Shockingly, the Flowers 102 dataset uses a single default: "a photo of a __, a type of flower", and is often able to outperform the class-conditional LLM approach on average. On the other hand, the DTD templates of the form "a photo of a __ {texture, pattern, thing, object}" are dramatically outperformed by our LLM-generated captions such as "a gauzy surface is characterized by a thin, translucent, and wispy appearance that is soft and delicate in texture", with a nearly 20% increase in top-5 accuracy on the ResNet50 and ViT-B/32 architectures. Understanding, both theoretically and experimentally, the properties of the distribution $P_{Z|Y=\boldsymbol{y}}$ that explain such performance differences between hand-designed and model-generated captions is still a ripe and exciting topic for further research.

4.6 Possible Extensions

4.6.1 Misspecified Marginal Distributions

Throughout the chapter, we have assumed that the marginals (P_X, P_Z) supplied by the user are accurate. A natural question to consider is how the analysis and final result in terms of mean squared error change when exposed to some degree of marginal error. Indeed, we consider the balancing method given marginals $(\hat{P}_X, \hat{P}_Z)$ which satisfy the following structure.

Assumption 4.6.1. There exist fixed probability mass functions $\hat{P}_X$ and $\hat{P}_Z$ for some $\varepsilon \in [0, 1)$,

$$\hat{P}_{X,\varepsilon} = (1-\varepsilon)P_X + \varepsilon\hat{P}_X \text{ and } \hat{P}_{Z,\varepsilon} = (1-\varepsilon)P_Z + \varepsilon\hat{P}_Z.$$

We also have the existence of the positive quantity

$$\hat{p}_\star := \min\{\min_{\boldsymbol{x}} \hat{P}_X(\boldsymbol{x}), \min_{\boldsymbol{z}} \hat{P}_Z(\boldsymbol{z})\} > 0.$$

Given the existence of $\hat{p}_\star > 0$, we may also define

$$\hat{p}_{\star,\varepsilon} = \min\{\min_{\boldsymbol{x}} \hat{P}_{X,\varepsilon}(\boldsymbol{x}), \min_{\boldsymbol{z}} \hat{P}_{Z,\varepsilon}(\boldsymbol{z})\} \geq \varepsilon\hat{p}_\star + (1-\varepsilon)p_\star > 0.$$

To be precise, the iterations of balancing are initialized at $\hat{P}_n^{(0)} = P_n$ and

$$\hat{P}_n^{(k)}(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} \cdot \hat{P}_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ odd} \\ \frac{\hat{P}_{Z,\varepsilon}(\boldsymbol{z})}{\hat{P}_{n,Z}^{(k-1)}(\boldsymbol{z})} \cdot \hat{P}_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ even} \end{cases}. \tag{4.26}$$

While details are deferred to Appendix B.3.5, we give an overview of the proof and state the final bound in this section. Broadly, the proof will proceed by deriving analogous results to those in Appendix B.3.1, Appendix B.3.3, and Appendix B.3.4. As in Section 4.3, the backbone of the argument will be a recursive formula to relate the estimation error across

iteration counter $k \geq 1$. When doing so similarly to (4.8), we achieve the equality

$$[\hat{P}_n^{(k)} - P](h) = [\hat{P}_n^{(k-1)} - P](\mathcal{C}_k h) + \hat{V}_n^{(k-1)}(\mathcal{C}_k h) + \begin{cases} [\hat{P}_{X,\varepsilon} - P_X](\mu_X h) & \text{if } k \text{ odd} \\ [\hat{P}_{Z,\varepsilon} - P_Z](\mu_Z h) & \text{if } k \text{ even} \end{cases}. \quad (4.27)$$

as shown in (B.48). Because we must bound the error terms containing $(\hat{P}_{X,\varepsilon} - P_X)$ and $(\hat{P}_{Z,\varepsilon} - P_Z)$, unlike (4.8), our recursion will be stated in the form of an inequality. We measure the deviation of the marginals $(\hat{P}_X, \hat{P}_Z)$ from the true (P_X, P_Z) using the constant

$$c^2 = \max\left\{\chi^2(\hat{P}_X \| P_X), \chi^2(\hat{P}_Z \| P_Z)\right\},$$

and show in Proposition B.3.4 that

$$[\hat{P}_n^{(k)} - P](h) \leq \left|[\hat{P}_n^{(k-1)} - P](\mathcal{C}_k h)\right| + \left|\hat{V}_n^{(k-1)}(\mathcal{C}_k h)\right| + c\,\|h\|_{\mathbf{L}^2(P)}\,\sqrt{\varepsilon},$$

where, as one might expect, we defined

$$\hat{V}_n^{(k-1)}(h) := \begin{cases} \sum_{\boldsymbol{x},y} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1\right) h(\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) & k \text{ odd} \\ \sum_{\boldsymbol{x},y} \left(\frac{\hat{P}_{Z,\varepsilon}}{\hat{P}_{n,Z}^{(k-1)}}(\boldsymbol{z}) - 1\right) h(\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) & k \text{ even} \end{cases} \quad (4.28)$$

The inequality (4.28) is a form of the centered recursion formula in Proposition B.3.1, and we can perform the unrolling

$$[\hat{P}_n^{(k)} - P](h) \leq \underbrace{[P_n^{(0)} - P](h)(\mathcal{C}_1 \ldots \mathcal{C}_k h)}_{\text{first-order term}} + \underbrace{\sum_{\ell=1}^{k} \left|\hat{V}_n^{(\ell-1)}(\mathcal{C}_\ell \ldots \mathcal{C}_k h)\right|}_{\text{higher-order term}} + \underbrace{O(k\sqrt{\varepsilon})}_{\text{misspecification}}$$

and establish an analogous uncentered recursion formula to upper bound the higher-order terms. The remainder of the argument is purely algebraic; the calculations are contained in Appendix B.3.5, and are written in a way to reuse parts of the analysis from throughout Appendix B.3. For convenience, we state the mean squared error bound in terms of its dependence on $(\varepsilon, n, k, \hat{p}_{\star,\varepsilon})$ and remove absolute constants and other problem parameters.

Theorem 4.6.1. *Let Assumption 4.6.1 be true with error $\varepsilon \in [0,1)$. For a sequence of rebalanced distributions $(\hat{P}_n^{(k)})_{k\geq 1}$, there exists an absolute constant $C > 0$ such that when $n \geq C[\log_2(2n/\hat{p}_{\star,\varepsilon}) + m\log(n+1)]/\min\{p_\star, \hat{p}_{\star,\varepsilon}\}^2$, we have that*

$$\mathbb{E}_P\left[\left(\hat{P}_n^{(k)}(h) - P(h)\right)^2 \mathbb{1}_{\mathcal{S}}\right] + \mathbb{E}_P\left[(P_n(h) - P(h))^2 \mathbb{1}_{\mathcal{S}^c}\right] \leq \frac{\sigma_k^2}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$
$$+ \tilde{O}\left(\frac{k^4}{\hat{p}_{\star,\varepsilon}^2}\left(\sqrt{\frac{1}{n}\log\frac{1}{1-\varepsilon}} + \log\frac{1}{1-\varepsilon}\right)\left[\frac{k^2}{\hat{p}_{\star,\varepsilon}^2}\left(\sqrt{\frac{1}{n}\log\frac{1}{1-\varepsilon}} + \log\frac{1}{1-\varepsilon} + \frac{1}{n}\right) + \frac{1}{\sqrt{n}}\right]\right)$$
$$+ \tilde{O}\left(k^2\left[\sqrt{\varepsilon}\left(\frac{\hat{p}_{\star,\varepsilon}^4}{n^4} + \frac{1}{\sqrt{n}} + \frac{\hat{p}_{\star,\varepsilon}^2 k}{n^4}\left(n + \frac{k^2}{\hat{p}_{\star,\varepsilon}^2}\right) + \frac{k^2}{\hat{p}_{\star,\varepsilon}^2}\left[\frac{1}{n} + \sqrt{\frac{1}{n}\log\frac{1}{1-\varepsilon}} + \log\frac{1}{1-\varepsilon}\right]\right) + \varepsilon\right]\right).$$

Notably, we can no longer take $k \to \infty$ as $n \to \infty$ in the expression above unless we also have that $\varepsilon \to 0$ at the appropriate rate.

4.6.2 Alternative Marginal Rebalancing Approches

In Section 4.3, we considered an estimator which incorporated the marginals by approximating the projection (4.7), taking for granted that it is a reasonable way to use this side information. However, when dealing with continuous real-valued data, there is another rather simple method by using the generalized inverse of the cumulative distribution functions (CDFs) of P_X and P_Z. In the estimator discussed in Bickel et al. [1991, Section 4], we have that $\mathcal{X} = [0,1]$, $\mathcal{Z} = [0,1]$, and the respective target marginals are uniform. Thus, the joint distribution P is a copula and the goal is to estimate a particular linear functional of P: the probability of the event $\{X \geq s \cap Z \geq t\}$ for $(s,t) \in [0,1] \times [0,1]$.[3] By using the inverse CDF trick described below, one can simply transform the two modalities individually to (approximately) fit a particular marginal distribution in each variable. It is shown in the same work that this approach is in fact *inefficient* in the nonparametric model with (P_X, P_Z) known, while the raking/iterative proportional fitting method used in this chapter is efficient.

[3]We use the survival events $\{X \geq s\}$ and $\{Z \geq t\}$ instead of $\{X \leq s\}$ and $\{Z \leq t\}$ because calculations with inverse CDFs will be simpler.

One notable difference between this work and Bickel et al. [1991] is that we consider discrete data directly, whereas they consider continuous data that is discretized via partitions. Letting $\mathcal{A}_n$ be a partition of measurable subsets of $\mathcal{X}$ and $\mathcal{B}_n$ defined similarly for $\mathcal{Z}$, the condition Bickel et al. [1991, F1] requires that

$$P_X(A) \geq \frac{\lambda_n}{\sqrt{n}} \; \forall A \in \mathcal{A}_n \text{ and } P_Z(B) \geq \frac{\lambda_n}{\sqrt{n}} \; \forall B \in \mathcal{B}_n,$$

where the sequence $(\lambda_n)_{n\geq 1}$ satisfies $\lambda_n^2/\log(n) \to \infty$ and $\lambda_n/\sqrt{n} \to 0$ as $n \to \infty$. Moreover, they directly analyze the solution (4.7), so that the efficient influence function and asymptotic variance are expressed in terms of projections. As mentioned in Section 4.3, a key outcome of our analysis is a formula for the asymptotic variance reduction by constructing a sequence of probability measures and Markov operators associated to each iteration of (4.2), where every probability measure of the sequence can be expressed in closed form. Thus, the asymptotic variance can be stated in terms of the spectra of these operators.

Despite the inefficiency of the copula-based estimator for continuous data, it is conceptually interesting to construct such an estimator for discrete data, providing another possibly efficient estimator. Conceptually, the raking approach can be described as one in which the empirical measure P_n is computed from data, and then transformed using the auxiliary marginal information (P_X, P_Z). We may also consider a reversed approach, that is, to transform the data $(X_1, Z_1), \ldots, (X_n, Z_n)$ and then compute the empirical measure afterward, in a way that ensures that the resulting measure adheres to the marginal constraints. To be precise, "adhering" to the marginal constraints will be interpreted differently for continuous and discrete data, as explained in the rest of this section. We now describe a generalization of the copula-based estimator that preserves the essential aspects but can be used in our setting.

Generalizing the Copula Estimator Having observed univariate data $(X_1, Z_1), \ldots, (X_n, Z_n)$, let $F_{n,X}$ denote the empirical CDF of $\{X_i\}_{i=1}^n$, $F_{n,Z}$ of $\{Z_i\}_{i=1}^n$, and let $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Z_i)}$ denote the empirical measure. The target marginals can be represented by their respective

CDFs F_X and F_Z. Finally, for any CDF F on $\mathbb{R}$, define the quantile function $F^{-1}(s) := \inf\{\boldsymbol{x} \in \mathbb{R} : F(\boldsymbol{x}) \geq s\}$. The estimator in this case is

$$\begin{aligned} P_n\left(X \geq F_{n,X}^{-1}(s), Z \geq F_{n,Z}^{-1}(t)\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_i \geq F_{n,X}^{-1}(s), Z \geq F_{n,Z}^{-1}(t)\right\} \\ &\overset{(1)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{F_{n,X}(X_i) \geq s, F_{n,Z}(Z_i) \geq t\right\} \\ &\overset{(2)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{F_X^{-1}((F_{n,X}(X_i)) \geq s, F_Z^{-1}(F_{n,Z}(Z_i)) \geq t\right\}, \end{aligned}$$

where (1) follows by Bobkov and Ledoux [2019, Lemma A.3] and (2) follows because the inverse CDF of the uniform distribution $[0,1]$ is equal to the identity on $[0,1]$. Notably, we applied the maps $T_X = F_X^{-1} \circ F_{n,X}$ and $T_Z = F_Z^{-1} \circ F_{n,Z}$ to each of our data sources, and then estimated the empirical measure. As a result, we have that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{T_X(X_i)) \geq s\right\} &= 1 - F_X(s) \text{ for } s \in \left\{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n}\right\} \\ \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{T_Z(Z_i)) \geq t\right\} &= 1 - F_Z(t) \text{ for } t \in \left\{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n}\right\}. \end{aligned}$$

The restriction that s and t are selected at increments of $1/n$ is the reason why the transformed data only partially adheres to the marginal constraints. Indeed, it is unreasonable to require an empirical CDF to agree with the CDF of a continuous random variable on its entire domain. To complete the story, we notice that T_X and T_Z are the optimal transportation maps from $F_{n,X}$ to F_X and $F_{n,Z}$ to F_Z, respectively, with respect to the squared distance on $\mathbb{R}$ (see [Bobkov and Ledoux, 2019, Theorems 2.10 and 2.11]). This key observation will motivate the discrete version of this estimator.

The Transport Plan Estimator We return to the setting in which the sample spaces $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ and $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$ are finite. We adopt in this discussion the following assumption. In a full treatment, we argue that this assumption holds with high probability.

Assumption 4.6.2. The empirical marginals $(P_{n,X}, P_{n,Z})$ and the target marginals (P_X, P_Z) are positive on their domains.

Following suit from the previous section, we are interested in defining transformations $T_X : \mathcal{X} \mapsto \mathcal{X}$ and $T_Z : \mathcal{Z} \mapsto \mathcal{Z}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{T_X(X_i) = \boldsymbol{x}\right\} = P_X(\boldsymbol{x}) \text{ and } \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{T_Z(Z_i) = \boldsymbol{z}\right\} = P_Z(\boldsymbol{z}), \tag{4.29}$$

without altering the data too much. We first specify a ground cost by endowing $\mathcal{X}$ and $\mathcal{Z}$ with the discrete metrics $d_{\mathcal{X}}(\boldsymbol{z}, \boldsymbol{x}') = \mathbb{1}\left\{\boldsymbol{x} \neq \boldsymbol{x}'\right\}$ and $d_{\mathcal{Z}}(\boldsymbol{z}, \boldsymbol{z}') = \mathbb{1}\left\{\boldsymbol{z} \neq \boldsymbol{z}'\right\}$. This is a natural choice of metric, as we have not made any additional assumptions (such as ordinality) on $\mathcal{X}$ and $\mathcal{Z}$; they are purely categorical (although other domain-specific distances may be designed). Next, searching among the maps that satisfy (4.29) is equivalent to solving the Monge assignment problem, which may not be feasible. Thus, we consider the Kantorovich relaxation [Peyré and Cuturi, 2019, Section 2.3] and replace T_X and T_Z with probability measures $\pi_n^\star : \mathcal{X} \times \mathcal{X} \to [0, 1]$ and $\gamma_n^\star : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$ that solve the respective optimal transportation problems

$$\pi_n^\star = \underset{\pi \in \Pi(P_{n,X}, P_X)}{\arg\min} \sum_{i \neq j} \pi(\boldsymbol{x}_i, \boldsymbol{x}_j) \text{ and } \gamma_n^\star = \underset{\gamma \in \Pi(P_{n,Z}, P_Z)}{\arg\min} \sum_{i \neq j} \gamma(\boldsymbol{z}_i, \boldsymbol{z}_j), \tag{4.30}$$

which are always feasible and for which minimizers are guaranteed to exist. Here, Π denotes the set of couplings with the given marginals. For any solution pair $(\pi_n^\star, \gamma_n^\star)$ (called *transport plans*), we will design the estimator $P_n^\star$ of P in a way that is semantically similar to the transportation map interpretation of the previous section. Intuitively, we will define the estimator to be the joint distribution of random variables (U, V) on $\mathcal{X} \times \mathcal{Z}$, which are components of a quadruple (X, Z, U, V) satisfying the following properties:

- $(X, Z) \sim P_n$.
- $(X, U) \sim \pi_n^\star$ and $(Z, V) \sim \gamma_n^\star$.
- U and (Z, V) are conditionally independent given X.

- V and (X, U) are conditionally independent given Z.

These properties are enough to specify the expectation of a function h. Accordingly, we may retrieve probabilities by taking the expectation of the indicator. First, by the law of total expectation:

$$\mathbb{E}_{P_n^\star}\left[h(U,V)\right] = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{z}\in\mathcal{Z}} \mathbb{E}_{P_n^\star}\left[h(U,V)|(X,Z)=(\boldsymbol{x},\boldsymbol{z})\right] P_n(\boldsymbol{x},\boldsymbol{z}).$$

For an indicator random variable $h(\boldsymbol{x},\boldsymbol{z}) = \mathbb{1}\left\{\boldsymbol{x}=\boldsymbol{u}\right\}\mathbb{1}\left\{\boldsymbol{z}=\boldsymbol{v}\right\}$, we may then compute

$$\begin{aligned}
P_n^\star(\boldsymbol{u},\boldsymbol{v}) &= \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{z}\in\mathcal{Z}} \mathbb{E}_{P_n^\star}\left[\mathbb{1}\left\{U=\boldsymbol{u}\right\}\mathbb{1}\left\{V=\boldsymbol{v}\right\}|(X,Z)=(\boldsymbol{x},\boldsymbol{z})\right] P_n(\boldsymbol{x},\boldsymbol{z}) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{z}\in\mathcal{Z}} \mathbb{E}_{P_n^\star}\left[\mathbb{1}\left\{U=\boldsymbol{u}\right\}|X=\boldsymbol{x}\right]\mathbb{E}_{P_n^\star}\left[\mathbb{1}\left\{V=\boldsymbol{v}\right\}|Z=\boldsymbol{z}\right] P_n(\boldsymbol{x},\boldsymbol{z}) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{z}\in\mathcal{Z}} \frac{\pi_n^\star(\boldsymbol{x},\boldsymbol{u})}{P_{n,X}(\boldsymbol{x})}\cdot\frac{\gamma_n^\star(\boldsymbol{z},\boldsymbol{v})}{P_{n,Z}(\boldsymbol{z})} P_n(\boldsymbol{x},\boldsymbol{z}).
\end{aligned}$$

It is easy to verify that $P_n^\star$ satisfies the marginal constraints. However, given that this estimator relies on the optimal transport plans $\pi_n^\star$ and $\gamma_n^\star$, we find ourselves in a similar dilemma as Bickel et al. [1991]: it is challenging to compute the variance of estimators defined in terms of projections. Furthermore, while a solution $(\pi_n^\star, \gamma_n^\star)$ exists, it may not be unique. We hypothesize that a similar approach to the recursion from (4.8) can be taken by replacing $(\pi_n^\star, \gamma_n^\star)$ by the iterates of a procedure such as the Sinkhorn algorithm for solving entropy-regularized OT. The resulting couplings would also depend on a particular regularization parameter $\varepsilon \equiv \varepsilon_n > 0$. Computing the asymptotic variance of the estimator $P_n^\star(h)$ based on scaling the number of steps of this procedure and the regularization parameter offers an interesting path for future work, especially if the estimator is efficient in the discrete case.

4.7 Perspectives & Future Work

This chapter considers the statistical analysis of foundation models, from pre-training to zero-shot prediction. Inspired by data curation procedures used in practice, we study the effect of known marginals for estimating linear functionals of distributions over multiple modalities.

This yielded a surprising connection to the classical work of Bickel et al. [1991], quantifying the variance reduction afforded by this additional information in the form of non-asymptotic mean squared error bounds. We also provided a formal discussion of prompting-based zero-shot prediction, phrased in terms of the dependence relations between random variables.

An appealing theoretical follow-up to the variance reduction results in Section 4.3 would be to investigate the connections between the information projection viewpoint of the Sinkhorn iterations to an instance of mirror descent [Léger, 2021, Aubin-Frankowski et al., 2022, Deb et al., 2023]. In particular, the mirror descent viewpoint has yielded a sublinear $O(1/k)$ decay rate of the marginal violation in terms of Kullback-Leibler (KL) divergence (as opposed to the use of monotonicity of Sinkhorn iterations in Proposition B.2.4). The main benefit of such an analysis would be to meaningfully improve the dependence of the higher-order term of (4.10) on the iteration count k (as the current requirement for asymptotic efficiency is $k = o(n^{1/12})$). Part of the challenge of an approach like this is 1) that errors from each iteration accrue additively, as seen in the unrolled recursion (4.8), and 2) that the mirror descent approaches rely heavily on the use of the KL divergence. Noticing that for iteration k, the higher-order components include terms of the form $\left(\frac{P_X}{P_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1\right)$, which may suggest the use of χ^2-divergence as an alternative to KL. Owing to this, an exciting contribution would be to extend the mirror descent viewpoint to a broader class of information divergences, which would serve other communities such as statistical optimal transport as well.

Chapter 5

CONCLUSION

In this dissertation, we address several problems regarding statistical learning methods that exhibit out-of-distribution generalization. The resulting predictors may be applied to data from distributions other than the training distribution without a catastrophic increase in prediction error. As a methodological theme, we focus on the development and analysis of large-scale optimization procedures over the model parameters or even the distribution of the training data.

To summarize Chapter 2, we considered the distributionally robust optimization (DRO) problem, wherein the parameter of interest is estimated by minimizing the worst-case empirical risk achievable by reweighting the training examples. Using ideas such as progressive bias and variance reduction, we constructed stochastic algorithms with linear convergence guarantees for smoothed maximum-type problems with convex loss functions. We focused on several practical aspects of the problem, such as solving the dual maximization problem, selecting the uncertainty set/smoothing parameter, and extensions to changing supports and neural networks.

A question that can spawn many studies is the relationship between distributionally robust optimization and highly expressive, data-interpolating models such as neural networks. While linear models are amenable to fitting via convex optimization, one limitation to their use in distributionally robust optimization is the fact that the optimal primal solution might be close to (or even coincide with) the solution to the empirical risk minimization problem [Zhai et al., 2021]. On the other hand, while more expressive models may have fewer theoretical guarantees for the runtime of optimization procedures, it is of statistical interest whether the perturbations to the training distribution may increase the "effective" sample size of the

training set (akin to data augmentations). In fact, if a network is trained to zero loss on all training examples, then it will also minimize any distributionally robust objective. Because there are possibly many zero-loss parameter settings for overparametrized models, one may investigate whether distributionally robust objectives lead to such solutions with better generalization properties than the optimizers of the empirical risk minimization objective.

In Chapter 3, we augmented the algorithmic ideas of Chapter 2 with notions of adaptive sampling and historical regularization, which led to state-of-the-art algorithms for a broad class of "semilinear" min-max problems. This problem class is simultaneously a generalization of bilinearly coupled min-max problems and a special case of general nonbilinearly coupled min-max problems. When the dual feasible set is decomposable into separate feasible sets for each coordinate of the dual variable, we also apply coordinate-wise stochastic updates for improved complexity guarantees. While originally motivated by applications in distributionally robust optimization, a thorough experimental study of the proposed methods in the fully composite optimization and minimization with functional constraints (in the spirit of Chapter 2) would be a future direction of great practical interest.

Chapter 4 studies the effect of projecting an empirical measure onto the set of probability measures that satisfy particular marginal constraints. These marginal constraints come from prior knowledge of the data-generating distribution. We find that this procedure, which reflects data curation methods used in modern foundation modeling, provides a variance reduction for linear plug-in estimators based on this adjusted empirical measure. Other aspects of foundation modeling, such as downstream zero-shot prediction, are also discussed from the lens of conditional dependence relations between (unlabeled) pre-training and (labeled) task-specific data.

On downstream tasks, we focused on zero-shot prediction (often classification) as this is a canonical task for evaluating foundation models in terms of their pre-training data or architectures [Gadre et al., 2023]. However, the universal representation principle is used much more generally. Most commonly, the image and text embeddings can be applied for tasks such as retrieval or computing similarity/distance/kernel values between data points.

Deriving performance guarantees of zero-shot procedures against optimal predictors for these tasks would be an interesting direction for future work as well.

While many foundational questions remain, this dissertation aims to be a step toward out-of-distribution generalization being a common standard for statistical theory and methods across scientific disciplines.

Software

Links to all of the repositories below can be found at `https://ronakdm.github.io/software`.

Paper Reprodocubility:

- `lerm` [Mehta et al., 2023].
- `prospect` [Mehta et al., 2024b].
- `drago` [Mehta et al., 2024a].
- `balancing` [Liu et al., 2024].
- `zeroshot` [Mehta and Harchaoui, 2025].

Standalone Packages:

- `deshift`: Instance-level and group-level distributionally robust optimization for CPU/GPU PyTorch workflows, with support for data distributed computing.
- `drlearn`: Distributionally robust linear predictors in the scikit-learn interface.

BIBLIOGRAPHY

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 2002.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A Reductions Approach to Fair Classification. In *ICML*, 2018.

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *ICML*, 2019.

Medha Agarwal, Kasim Rafiq, Ronak Mehta, Briana Abrahms, and Zaid Harchaoui. Leveraging machine learning and accelerometry to classify animal behaviours with uncertainty. *Methods in Ecology and Evolution*, 2024.

Uchenna Akujuobi and Xiangliang Zhang. Delve: A Dataset-Driven Scholarly Search and Analysis System. *SIGKDD Explor. Newsl.*, 19, 2017.

Ahmet Alacaoglu and Yura Malitsky. Stochastic Variance Reduction for Variational Inequality Methods. In *COLT*, 2022.

Ahmet Alacaoglu, Quoc Tran Dinh, Olivier Fercoq, and Volkan Cevher. Smooth Primal-Dual Coordinate Descent Algorithms for Nonsmooth Convex Optimization. In *NeurIPS*, 2017.

Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. Random extrapolation for primal-dual coordinate descent. In *ICML*, 2020.

Ahmet Alacaoglu, Volkan Cevher, and Stephen J. Wright. On the Complexity of a Practical Primal-Dual Coordinate Method, 2022. arXiv Technical Report.

Mickael Albertus and Philippe Berthet. Auxiliary information: the raking-ratio empirical process. *Electronic Journal of Statistics*, 2019.

Zeinab Alizadeh, Erfan Yazdandoost Hamedani, and Afrooz Jalilzadeh. Variance-reduction for Variational Inequality Problems with Bregman Distance Function, 2024.

Philippe Artzner, Freddy Delbaen, Eber Jean-Marc, and David Heath. Coherent Measures of Risk. *Mathematical Finance*, 1999.

YM. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *CVPR*, 2023.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror Descent with Relative Smoothness in Measure Spaces, with application to Sinkhorn and EM. In *NeurIPS*, 2022.

Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024.

Michel Baes, Michael Burgisser, and Arkadi Nemirovski. A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *SIAM Journal on Optimization*, 2013.

Randall Balestriero and Yann LeCun. Contrastive and Non-Contrastive Self-Supervised Learning Recover Global and Local Spectral Embedding Methods. In *NeurIPS*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *ICLR*, 2022.

Sandi Baressi Segota, Nikola Andelic, Jan Kudlacek, and Robert Cep. Artificial Neural Network for Predicting Values of Residuary Resistance per Unit Weight of Displacement. *Journal of Maritime & Transportation Science*, 57, 2020.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 Competition Dataset. *arXiv preprint arXiv:2004.10340*, 2020.

Aharon Ben-Tal and Marc Teboulle. An Old-New Concept of Convex Risk Measures: The Optimized Certainty Equivalent. *Mathematical Finance*, 2007.

Dimitri P Bertsekas. Nonlinear Programming. *Journal of the Operational Research Society*, 1997.

Sanjay P. Bhat and L. A. Prashanth. Concentration of risk measures: A Wasserstein distance approach. In *NeurIPS*, 2019.

Peter J. Bickel, Ya'Acov Ritov, and Jon A. Wellner. Efficient Estimation of Linear Functionals of a Probability Measure P with Known Marginal Distributions. *The Annals of Statistics*, 1991.

David A. Binder. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 1983.

Sergey G. Bobkov and Michel Ledoux. One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances. *Memoirs of the American Mathematical Society*, 2019.

Digvijay Boob and Mohammad Khalafi. Optimal Primal-Dual Algorithm with Last iterate Convergence Guarantees for Stochastic Convex Optimization Problems, 2024. arXiv Technical Report.

Ekaterina Borodich, Georgiy Kormakov, Dmitry Kovalev, Aleksandr Beznosikov, and

Alexander Gasnikov. Near-Optimal Algorithm with Complexity Separation for Strongly Convex-Strongly Concave Composite Saddle Point Problems. In *ICOMP*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

Andreas Buja. Remarks on Functional Canonical Variates, Alternating Least Squares Methods and ACE. *The Annals of Statistics*, 1990.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *ACM Conference on Fairness, Accountability and Transparency*, 2018.

Jonathon Byrd and Zachary C. Lipton. What is the Effect of Importance Weighting in Deep Learning? In *ICML*, 2019.

Xufeng Cai, Ahmet Alacaoglu, and Jelena Diakonikolas. Variance Reduced Halpern Iteration for Finite-Sum Monotone Inclusions. In *ICLR*, 2024.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance Reduction for Matrix Games. In *NeurIPS*, 2019.

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. RECAPP: Crafting a More Efficient Catalyst for Convex Optimization. In *ICML*, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Antonin Chambolle and Thomas Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 2011.

Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications. *SIAM Journal on Optimization*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing Textures in the Wild. In *CVPR*, 2014.

Laurent Condat. Fast projection onto the simplex and the ℓ_1-ball. *Mathematical Programming*, 2016.

Peter Coppens and Panagiotis Patrinos. Ordered Risk Minimization: Learning More from Less Data. *IEEE Conference on Decision and Control (CDC)*, 2023.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 1995.

John Cotter and Kevin Dowd. Extreme Spectral Risk Measures: an Application to Futures Clearinghouse Margin Requirements. *Journal of Banking & Finance*, 2006.

Thomas M Cover. *Elements of Information Theory*. John Wiley & Sons, 1999.

Ying Cui, Jong-Shi Pang, and Bodhisattva Sen. Composite Difference-Max Programs for Modern Statistical Estimation Problems. *SIAM Journal on Optimization*, 2018.

Sebastian Curi, Kfir Y. Levy, Stefanie Jegelka, and Andreas Krause. Adaptive Sampling for Stochastic Risk-Averse Learning. In *NeurIPS*, 2020.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic Optimization under Distributional Drift. *Journal of Machine Learning Research*, 2023.

Abdelaati Daouia, Irène Gijbels, and Gilles Stupfler. Extremiles: A New Perspective on Asymmetric Least Squares. *Journal of the American Statistical Association*, 2019.

H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, 3nd edition, 2003.

Nabarun Deb, Young-Heon Kim, Soumik Pal, and Geoffrey Schiebinger. Wasserstein Mirror Gradient Flow as the limit of the Sinkhorn Algorithm. arXiv Technical Report, 2023.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *NeurIPS*, 2014.

W. Edwards Deming and Frederick F. Stephan. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, 1940.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-scale Hierarchical Image Database. In *CVPR*, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina" Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*, 2019b.

Jelena Diakonikolas. A Block Coordinate and Variance-Reduced Method for Generalized Variational Inequalities of Minty Type. *JDS*, 2025.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 2019.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *NeurIPS*, volume 34. Curran Associates, Inc., 2021.

Nikita Doikov and Yurii Nesterov. High-Order Optimization Methods for Fully Composite Problems. *SIAM Journal on Optimization*, 2022.

D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 2019.

John C. Duchi and Hongseok Namkoong. Variance-based Regularization with Convex Objectives. *Journal of Machine Learning Research*, 2019.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 2021.

John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 2021.

Alex Dytso, Martina Cardone, and Ian Zieder. Meta Derivative Identity for the Conditional Expectation. *IEEE Transactions on Information Theory*, 2023.

Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with Average Top-k Loss. In *NeurIPS*, 2017.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language-image pre-training (CLIP). In *ICML*, 2023.

Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking Importance Weighting for Deep Learning under Distribution Shift. In *NeurIPS*, 2020.

Olivier Fercoq. Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient. *Open Journal of Mathematical Optimization*, 2023.

Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance Stochastics*, 2002.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023.

I. Gohberg, S. Goldberg, and M.A. Kaashoek. *Classes of Linear Operators Vol. 1*. Springer, 1990.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain Adaptation with Conditional Transferable Components. In *ICML*, 2016.

Robert M. Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-Reduced Methods for Machine Learning. *Proceedings of the IEEE*, 2020.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate Shift by Kernel Mean Matching. In *Dataset Shift in Machine Learning*. The MIT Press, 2008.

Vincent Guigues and Claudia A. Sagastizábal. Risk-averse feasible policies for large-scale multistage stochastic linear programs. *Mathematical Programming*, 2013.

Michael U. Gutmann and Aapo Hyvärinen. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *JMLR*, 2012.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. In *NeurIPS*, 2021.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without Demographics in Repeated Loss Minimization. In *ICML*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Xue Dong He, Steven Kou, and Xianhua Peng. Risk Measures: Robustness, Elicitability, and Backtesting. *Annual Review of Statistics and Its Application*, 2022.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. arXiv Technical Report, 2022.

Matthew J. Holland and El Mehdi Haress. Spectral risk-based learning using unbounded losses. In *AISTATS*, 2022.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2013.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting Sample Selection Bias by Unlabeled Data. In *NeurIPS*, 2006.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. GitHub Repository, 2022.

C. T. Ireland and S. Kullback. Contingency tables with given marginals. *Biometrika*, 1968.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper Rates for Separable Minimax and Finite Sum Optimization via Primal-Dual Extragradient Methods. In *COLT*, 2022.

Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *NeurIPS*, 2013.

R. J. Johnston and C. J. Pattie. Entropy-maximizing and the iterative proportional fitting procedure. *The Professional Geographer*, 1993.

Corinne Jones, Vincent Roulet, and Zaid Harchaoui. Discriminative clustering with representation learning with any ratio of labeled to unlabeled data. *Statistics and Computing*, 2022.

Anatoli Juditsky, Arkadi Nemirovski, et al. First-Order Methods for Nonsmooth Convex Large-Scale Optimization, II: Utilizing Problem's Structure. *Optimization for Machine Learning*, 2011.

Anatoli Juditsky, Fatma Kılınç Karzan, and Arkadi Nemirovski. Randomized first order algorithm with applications to ℓ_1-minimization. *Mathematical Programming*, 2013.

H. Kahn and A. W. Marshall. Methods of Reducing Sample Size in Monte Carlo Computations. *Journal of the Operations Research Society of America*, 1953.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Computational complexity of kernel-based density-ratio estimation: a condition number analysis. *Machine Learning*, 2013.

Masahiro Kato and Takeshi Teshima. Non-Negative Bregman Divergence Minimization for Deep Direct Density Ratio Estimation. In *ICML*, 2021.

Kenji Kawaguchi and Haihao Lu. Ordered SGD: A New Stochastic Optimization Framework for EmpiricalvRisk Minimization. In *AISTATS*, 2020.

Justin Khim, Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. Uniform Convergence of Rank-weighted Learning. In *ICML*, 2020.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2021.

G. M. Korpelevich. An Extragradient Method for Finding Saddle Points and for Other Problems. *Ekonomika i Matematicheskie Metody*, 1976.

Dmitry Kovalev and Alexander Gasnikov. The First Optimal Algorithm for Smooth and Strongly-Convex-Strongly-Concave Minimax Optimization. In *NeurIPS*, 2022.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. University of Toronto Technical Report, 2009.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in

high-dimensional statistics: applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 2022.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. *Operations research & management science in the age of analytics*, 2019.

Ramnath Kumar, Kushal Alpesh Majmundar, Dheeraj Mysore Nagaraj, and Arun Suggala. Stochastic re-weighted gradient descent via distributionally robust optimization. *Transactions on Machine Learning Research*, 2024.

Yassine Laguel, Jerome Malick, and Zaid Harchaoui. First-Order Optimization for Superquantile-Based Supervised Learning. In *IEEE MLSP*, 2020.

Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at Work: Machine Learning Applications and Efficient Subgradient Computation. *Set-Valued and Variational Analysis*, 2021.

Henry Lam and Enlu Zhou. The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 2017.

Guanghui Lan and Yan Li. A Novel Catalyst Scheme for Stochastic Minimax Optimization, 2023. arXiv Technical Report.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Nicolas Le Roux, Mark Schmidt, and Francis Bach. A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets. In *NeurIPS*, 2012.

Thanh Quoc Trinh Le Thi Thanh Hai and Phan Tu Vuong. A refined convergence analysis of Popov's algorithm for pseudo-monotone variational inequalities. *Optimization*, 2025.

Jaeho Lee, Sejun Park, and Jinwoo Shin. Learning Bounds for Risk-sensitive Learning. In *NeurIPS*, 2020.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting What You Already Know Helps: Provable Self-Supervised Learning. In *NeurIPS*, 2021.

Flavien Léger. A Gradient Descent Perspective on Sinkhorn. *Applied Mathematics & Optimization*, 2021.

Liu Leqi, Adarsh Prasad, and Pradeep K Ravikumar. On Human-Aligned Risk Minimization. In *NeurIPS*, 2019.

Daniel Levy, Yair Carmon, John Duchi, and Aaron Sidford. Large-Scale Methods for Distributionally Robust Optimization. In *NeurIPS*, 2020.

Chris Junchi Li, Angela Yuan, Gauthier Gidel, Quanquan Gu, and Michael I. Jordan. Nesterov Meets Optimism: Rate-Optimal Separable Minimax Optimization. In *ICML*, 2023.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Kamal Mohamed Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Joshua P Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander T Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In search of the next generation of training sets for language models. In *NeurIPS*, 2024a.

Shuyao Li, Sushrut Karmalkar, Ilias Diakonikolas, and Jelena Diakonikolas. Learning a single neuron robustly to distributional shifts and adversarial label noise. In *NeurIPS*, 2024b.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 2018.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-Optimal Algorithms for Minimax Optimization. In *COLT*, 2020.

Lang Liu, Ronak Mehta, Soumik Pal, and Zaid Harchaoui. The Benefits of Balance: From Information Projections to Variance Reduction. In *NeurIPS*, 2024.

Cong Ma, Reese Pathak, and Martin J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 2023.

A. Rupam Mahmood, Hado P van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *NeurIPS*, 2014.

Julien Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 2014.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. arXiv Technical Report, 2013.

Patrice Marcotte. Application Of Khobotov's Algorithm To Variational Inequalities And Network Equilibrium Problems. *INFOR: Information Systems and Operational Research*, 1991.

Andreas Maurer, Daniela Angela Parletta, Andrea Paudice, and Massimiliano Pontil. Robust Unsupervised Learning via L-statistic Minimization. In *ICML*, 2021.

Ronak Mehta and Zaid Harchaoui. A Generalization Theory for Zero-Shot Prediction. In *ICML*, 2025.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, Lang Liu, and Zaid Harchaoui. Stochastic Optimization for Spectral Risk Measures. In *AISTATS*, 2023.

Ronak Mehta, Jelena Diakonikolas, and Zaid Harchaoui. Drago: Primal-Dual Coupled Variance Reduction for Faster Distributionally Robust Optimization. In *NeurIPS*, 2024a.

Ronak Mehta, Vincent Roulet, Krishna Pillutla, and Zaid Harchaoui. Distributionally Robust Optimization with Bias and Variance Reduction. In *ICLR*, 2024b.

Ronak Mehta, Jelena Diakonikolas, and Zaid Harchaoui. Min-Max Optimization with Dual-Linear Coupling, 2025. arXiv Technical Report.

Thomas Mikosch. Regular variation, subexponentiality and their applications in probability theory. *International Journal of Production Economics*, 1999.

Douglas J. Miller and Wei han Liu. On the recovery of joint distributions from limited information. *Journal of Econometrics*, 2002.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *AISTATS*, 2020.

Debarghya Mukherjee, Felix Petersen, Mikhail Yurochkin, and Yuekai Sun. Domain Adaptation meets Individual Fairness. And they get along. In *NeurIPS*, 2022.

Hongseok Namkoong and John C Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In *NeurIPS*, 2016.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz algorithm. In *NeurIPS*, 2014.

Arkadi Nemirovski. Prox-Method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems. *SIAM Journal on Optimization*, 2004.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 2007a.

Yurii Nesterov. Smoothing Technique and its Applications in Semidefinite Optimization. *Mathematical Programming*, 2007b.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Core Discussion Paper*, 2006.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-grained Aspects. In *EMNLP*, 2019.

Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

Marcel Nutz. Introduction to Entropic Optimal Transport. *Lecture notes, Columbia University*, 2021.

Ilsang Ohn and Yongdai Kim. Toward fast rates: A review of localization analysis for statistical learning, 2025. Technical report.

Kazusato Oko, Licong Lin, Yuhang Cai, and Song Mei. A Statistical Theory of Contrastive Pre-training and Multimodal Generative AI. arXiv Technical Report, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou,

Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *TMLR*, 2024.

Art Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 1990.

Art Owen. *Empirical Likelihood*. CRC Press, 1st edition, 2001.

Balamurugan Palaniappan and Francis Bach. Stochastic Variance Reduction Methods for Saddle-Point Problems. *NeurIPS*, 2016.

Ajay Kumar Pandey, L. A. Prashanth, and Sanjay P. Bhat. Estimation of Spectral Risk Measures. In *AAAI Conference on Artificial Intelligence*, 2019.

Giuditta Parolini. The Emergence of Modern Statistics in Agricultural Science: Analysis of Variance, Experimental Design and the Reshaping of Research at Rothamsted Experimental Station, 1919-1933. *Journal of the History of Biology*, 2015.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends in Machine Learning*, 2019.

Georg Ch Pflug and Andrzej Ruszczyński. Measuring Risk for Income Streams. *Computational Optimization and Applications*, 2005.

Alexander Pichugin, Maksim Pechin, Aleksandr Beznosikov, Vasilii Novitskii, and Alexander Gasnikov. Method with batching for stochastic finite-sum variational inequalities in non-Euclidean setting. *Chaos, Solitons & Fractals*, 2024.

L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 1980.

L. A. Prashanth and Sanjay P. Bhat. A Wasserstein Distance Approach for Concentration of Empirical Risk Estimates. *Journal of Machine Learning Research*, 2022.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 1959.

Ali Rizvi, Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 2014, 2014.

R. T. Rockafellar and J. O. Royset. Random Variables, Monotone Relations, and Convex Analysis. *Mathematical Programming*, 2014.

R Tyrrell Rockafellar and Johannes O Royset. Superquantiles and Their Applications to Risk, Random Variables, and Regression. In *Theory Driven by Influential Applications*. INFORMS, 2013.

R. Tyrrell Rockafellar and Stanislav Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18:33–53, 2013.

Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 1983.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *ICLR*, 2020.

René Schilling. *Measures, Integrals, and Martingales*. Springer, 2nd edition, 2017.

Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss. *Journal of Machine Learning Research*, 2013.

Jun Shao. Functional calculus and asymptotic theory for statistical analysis. *Statistics & Probability Letters*, 1989.

Alexander Shapiro. Distributionally Robust Stochastic Programming. *SIAM Journal on Optimization*, 2017.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

Galen Shorack. *Probability for Statisticians*. Springer, 2017.

Ravindra Singh and Naurang Singh Mangat. *Elements of Survey Sampling*. Springer, 1996.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly*, 74(4), 1967.

Chaobing Song, Stephen J Wright, and Jelena Diakonikolas. Variance Reduction via Primal-Dual Accelerated Dual Averaging for Nonsmooth Convex Finite-Sums. In *ICML*, 2021.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research*, 2007.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *NeurIPS*, 2008.

Adith Swaminathan and Thorsten Joachims. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research*, 2015.

Zhiquan Tan, Yifan Zhang, Jingqin Yang, and Yang Yuan. Contrastive Learning is Spectral Clustering on Similarity Graph. In *ICLR*, 2024.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the New Data in Multimedia Research. *Communications of the ACM*, 2016.

M. E. Thompson. *Theory of Sample Surveys*. Chapman & Hall, 2000.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *ALT*, 2021.

Athanasios Tsanas and Angeliki Xifara. Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools. *Energy and Buildings*, 49, 2012.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 1995.

Pınar Tüfekci. Prediction of Full Load Electrical Power Output of a Base Load Operated Combined Cycle Power Plant using Machine Learning Methods. *International Journal of Electrical Power & Energy Systems*, 60, 2014.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. arXiv Technical Report, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017.

Alexander Visheratin. NLLB-CLIP - train performant multilingual image retrieval model on a budget. In *NeurIPS Workshop: ENLSP-III*, 2023.

Maria-Luiza Vladarean, Nikita Doikov, Martin Jaggi, and Nicolas Flammarion. Linearization Algorithms for Fully Composite Optimization. In *COLT*, 2023.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions. *Stat*, 2020.

Hieu Vu, Toan Tran, Man-Chung Yue, and Viet Anh Nguyen. Distributionally Robust Fair Principal Components via Geodesic Descents. In *ICLR*, 2022.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Iyad Walwil and Olivier Fercoq. The smoothed duality gap as a stopping criterion. *Mathematical Programming Computation*, 2025.

Yuanhao Wang and Jian Li. Improved Algorithms for Convex-Concave Minimax Optimization. In *NeurIPS*, 2020.

Yunke Wang, Chang Xu, Bo Du, and Honglak Lee. Learning to Weight Imperfect Demonstrations. In *ICML*, 2021.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification. In *ICML*, 2014.

Robert Williamson and Aditya Menon. Fairness Risk Measures. In *ICML*, 2019.

Blake Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. In *NeurIPS*, 2016.

Xiaojing Xiang. A note on the bias of L-estimators and a bias reduction procedure. *Statistics & Probability Letters*, 1995.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Lin Xiao. Dual Averaging Method for Regularized Stochastic Learning and Online Optimization. In *NeurIPS*, 2009.

Da Xu, Yuting Ye, and Chuanwei Ruan. Understanding the role of importance weighting for deep learning. In *ICLR*, 2021.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *ICLR*, 2024.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A Catalyst Framework for Minimax Optimization. In *NeurIPS*, 2020.

Sitan Yang, Malcolm Wolff, Shankar Ramasubramanian, Vincent Quenneville-Belair, Ronak Mehta, and Michael Mahoney. GEANN: Scalable Graph Augmentations for Multi-Horizon Time Series Forecasting. In *KDD 2023 Workshop on Mining and Learning with Graphs*, 2023.

Ivy Yeh. Analysis of Strength of Concrete Using Design of Experiments and Neural Networks. *Journal of Materials in Civil Engineering*, 18, 2006.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *ICML*, 2021.

Runtian Zhai, Chen Dan, Arun Suggala, J. Zico Kolter, and Pradeep Ravikumar. Boosted CVaR Classification. In *NeurIPS*, 2021.

Huiming Zhang and Haoyu Wei. Sharper Sub-Weibull Concentrations. *Mathematics*, 2022.

Junyu Zhang, Minyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 2022.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain Adaptation under Target and Conditional Shift. In *ICML*, 2013.

Tianyi Zhang, Ikko Yamane, Nan Lu, and Masashi Sugiyama. A One-step Approach to Covariate Shift Adaptation. In *ACML*, 2020.

Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023.

Yu-Jie Zhang, Zhen-Yu Zhang, Peng Zhao, and Masashi Sugiyama. Adapting to Continuous Covariate Shift via Online Density Ratio Estimation. In *NeurIPS*, 2023.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Technical Background

In this section, we collect several results from convex analysis used throughout the thesis. In the following, let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ denote its associated dual norm.

A.1.1 Smooth and Strongly Convex Functions

The first set of results concerns L-smooth functions, or those with L-Lipschitz continuous gradient.

Theorem A.1.1. *[Nesterov, 2018, Theorem 2.1.5] The conditions below are considered for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\alpha \in [0, 1]$. The following are equivalent for a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$.*

1. *f is convex and L-smooth with respect to $\|\cdot\|$.*
2. $0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$.
3. $f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*^2 \leq f(\boldsymbol{y})$.
4. $\frac{1}{L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*^2 \leq \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.
5. $0 \leq \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq L \|\boldsymbol{x} - \boldsymbol{y}\|^2$.

Next, we detail the properties of strongly convex functions.

Theorem A.1.2. *[Nesterov, 2018, Theorem 2.1.10] If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and differentiable, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

- $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{1}{2\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*^2$.
- $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \leq \frac{1}{\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*^2$.
- $\mu \|\boldsymbol{x} - \boldsymbol{y}\| \leq \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_*$.

Finally, functions that are both smooth and strongly convex enjoy a number of relevant primal-dual properties.

Theorem A.1.3. *[Nesterov, 2018, Theorem 2.1.12] If f is both L-smooth and μ-strongly convex, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$-\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{y}\rangle = -\tfrac{\mu L}{\mu + L} \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \tfrac{1}{\mu + L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle . \quad \text{(A.1)}$$

Lemma A.1.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and M-smooth. Then, we have for any $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$,*

$$f(\boldsymbol{v}) \geq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\top (\boldsymbol{v} - \boldsymbol{w}) + \frac{1}{2(M + \mu)} \|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})\|_2^2 + \frac{\mu}{4} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2.$$

Proof. The function $g = f - \mu \|\cdot\|_2^2/2$ is convex and $M - \mu$ smooth. Hence, we have by line 3 of Theorem A.1.1 for any $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$,

$$g(\boldsymbol{v}) \geq g(\boldsymbol{w}) + \nabla g(\boldsymbol{w})^\top (\boldsymbol{v} - \boldsymbol{w}) + \frac{1}{2(M - \mu)} \|\nabla g(\boldsymbol{v}) - \nabla g(\boldsymbol{w})\|_2^2.$$

Expanding g and ∇g, we get

$$\begin{aligned} f(\boldsymbol{v}) \geq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\top (\boldsymbol{v} - \boldsymbol{w}) + \frac{1}{2(M - \mu)} \|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})\|_2^2 \\ + \frac{\mu M}{2(M - \mu)} \|\boldsymbol{w} - \boldsymbol{v}\|_2^2 - \frac{\mu}{M - \mu} (\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v}))^\top (\boldsymbol{w} - \boldsymbol{v}). \end{aligned}$$

Using Young's inequality, that is, $\boldsymbol{a}^\top \boldsymbol{b} \leq \frac{\alpha}{2}\|\boldsymbol{a}\|_2^2 + \frac{\alpha^{-1}}{2}\|\boldsymbol{b}\|_2^2$, we have

$$f(\boldsymbol{v}) \geq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\top (\boldsymbol{v} - \boldsymbol{w}) + \frac{1 - \alpha\mu}{2(M - \mu)}\|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{v})\|_2^2 + \frac{\mu(M - \alpha^{-1})}{2(M - \mu)}\|\boldsymbol{w} - \boldsymbol{v}\|_2^2.$$

Taking $\alpha = \frac{2}{\mu + M}$ gives the claim. □

A.1.2 f-Divergences

Let Q and P be two probability measures over $\mathcal{Z}$. Consider a convex function $f : [0, \infty) \to \mathbb{R} \cup \{+\infty\}$ such that $f(1) = 0$, $f(\boldsymbol{x})$ is finite for $x > 0$, and $\lim_{t \to 0^+} f(t) = 0$. The *f-divergence* from Q to P [Zhang, 2023, Appendix B] generated by this function f is

$$D_f(Q\|P) := \begin{cases} \int_{\mathcal{Z}} f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(z)\right) \mathrm{d}P(z) & \text{if } Q \ll P \\ +\infty & \text{otherwise} \end{cases}.$$

In the special case that $\boldsymbol{q}$ and $\boldsymbol{p}$ are two probability mass functions defined on atoms $\{1, \ldots, n\}$, we may use the abuse of notation

$$D_f(\boldsymbol{q}\|\boldsymbol{p}) := \sum_{i=1}^{n} f\left(\frac{q_i}{p_i}\right) p_i,$$

where we define $0 f\,(0/0) := 0$ in the formula above. If there is an i such that $p_i = 0$ but $q_i > 0$, we say $D_f(\boldsymbol{q}\|\boldsymbol{p}) = \infty$. The χ^2-divergence is generated by $f_{\chi^2}(t) = t^2 - 1$ and the KL divergence is generated by $f_{\mathrm{KL}}(t) = t \ln t + \iota_+(t)$ where ι_+ denotes the convex indicator that is zero for $t \geq 0$ and $+\infty$ otherwise, and we define $t \ln t = 0$ for all $t < 0$.

The convexity properties of the f-divergence in its first argument can be derived from similar properties of the function f on $\mathbb{R}$.

Proposition A.1.1. *Assume that $f : \mathbb{R} \to \mathbb{R}$ is α_n-strongly convex on $[0, n]$. Then, $\boldsymbol{q} \mapsto D_f(\boldsymbol{q}\|\mathbf{1}_n/n)$ is $(n\alpha_n)$-strongly convex with respect to $\|\cdot\|_2$.*

Proof. Due to the α_n-strong convexity of f, for any $\boldsymbol{q}, \boldsymbol{p} \in [0, 1]^n$ and any $\lambda \in (0, 1)$ and any

$i \in [n]$,

$$f\left(\lambda n q_i + (1-\lambda) n p_i\right) \leq \lambda f(n q_i) + (1-\lambda) f(n p_i) - \frac{\alpha_n}{2}\lambda(1-\lambda)(n q_i - n p_i)^2.$$

We average this inequality over i, yielding

$$\frac{1}{n}\sum_{i=1}^{n} f\left(n(\lambda q_i + (1-\lambda) p_i)\right) \leq \frac{\lambda}{n}\sum_{i=1}^{n} f(n q_i) + \frac{1-\lambda}{n}\sum_{i=1}^{n} f(n p_i) - \frac{\alpha_n}{2}\lambda(1-\lambda)\|n q_i - n p_i\|_2^2.$$

Defining $\mathrm{Reg}(\boldsymbol{q}) := D_f(\boldsymbol{q} \| \mathbf{1}_n/n)$, the statement above can be succinctly written as

$$\mathrm{Reg}(\lambda \boldsymbol{q} + (1-\lambda)\boldsymbol{p}) \leq \lambda\, \mathrm{Reg}(\boldsymbol{q}) + (1-\lambda)\, \mathrm{Reg}(\boldsymbol{p}) - \frac{\alpha_n n}{2}\lambda(1-\lambda)\|\boldsymbol{q} - \boldsymbol{p}\|_2^2 .$$

Therefore, Reg is $(\alpha_n n)$-strongly convex with respect to $\|\cdot\|_2$ on $[0,1]^n$, completing the proof.

□

A.1.3 Miscellaneous Results

Lemma A.1.2. *For a convex function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, if $x_1 \geq x_2$ and $y_2 \geq y_1$, then*

$$f(y_1 - x_1) + f(y_2 - x_2) \geq f(y_2 - x_1) + f(y_1 - x_2).$$

Proof. First, observe that

$$y_2 - x_2 \geq y_2 - x_1 \geq y_1 - x_1 \text{ and } y_2 - x_2 \geq y_1 - x_2 \geq y_1 - x_1.$$

Thus, $y_2 - x_1$ and $y_1 - x_2$ both lie between $y_2 - x_2$ and $y_1 - x_1$ and can be expressed as a convex combination of the two endpoints, that is

$$y_2 - x_1 = \alpha(y_2 - x_2) + (1-\alpha)(y_1 - x_1)$$
$$y_1 - x_2 = \beta(y_2 - x_2) + (1-\beta)(y_1 - x_1)$$

for some $\alpha, \beta \in [0,1]$. By solving for α we get $\alpha = 1 - \beta$. Apply the definition of convexity

to get

$$f(y_2 - x_1) \leq \alpha f(y_2 - x_2) + (1-\alpha) f(y_1 - x_1)$$
$$f(y_1 - x_2) \leq (1-\alpha) f(y_2 - x_2) + \alpha f(y_1 - x_1).$$

Sum both inequalities to achieve the desired result. □

Lemma A.1.1. *Consider a random variable X with c.d.f. F. If X satisfies $\mathbb{E}\left[|X|^p\right] < \infty$ for $p > 2$, then*

$$\int_{-\infty}^{+\infty} \sqrt{F(x)(1-F(x))}\, dx \leq \sqrt{2}\left(\frac{p}{p-2}\right) \mathbb{E}\left[|X|^p\right]^{\frac{1}{p}}.$$

Proof. By definition, $\int_{-\infty}^{\infty} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x = \lim_{a\to+\infty} \int_{-a}^{a} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x$. Denote $c = \mathbb{E}\left[|X|^p\right]^{1/p}$. For any constant $a \geq c > 0$, we have

$$\begin{aligned}
\int_{-a}^{a} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x &= \int_{-a}^{0} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x + \int_{0}^{a} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x \\
&\leq \int_{-a}^{0} \sqrt{F(x)}\,\mathrm{d}x + \int_{0}^{a} \sqrt{(1-F(x))}\,\mathrm{d}x \\
&= \int_{-a}^{0} \sqrt{\mathbb{P}(X \leq x)}\,\mathrm{d}x + \int_{0}^{a} \sqrt{\mathbb{P}(X > x)}\,\mathrm{d}x \\
&= \int_{0}^{a} \sqrt{\mathbb{P}(X \leq -x)} + \sqrt{\mathbb{P}(X > x)}\,\mathrm{d}x.
\end{aligned}$$

Then, use that for any $a, b \geq 0$,

$$(\sqrt{a} + \sqrt{b})^2 = a + b + 2\sqrt{ab} \leq 2(a+b) \implies \sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}.$$

Using this, and that $z \geq 0$, we have

$$\begin{aligned}
\sqrt{\mathbb{P}(X \leq -x)} + \sqrt{\mathbb{P}(X > x)} &\leq \sqrt{2(\mathbb{P}(X \leq -x) + \mathbb{P}(X > x))} \\
&= \sqrt{2(\mathbb{P}(|X| > x) + \mathbb{P}(X = -x))} \\
&\leq \sqrt{2(\mathbb{P}(|X| > x) + \mathbb{P}(|X| = x))} \\
&= \sqrt{2\mathbb{P}(|X| \geq x)}.
\end{aligned}$$

Combining with the first display, we have that

$$\begin{aligned}
\int_{-a}^{a} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x &\leq \int_{0}^{a} \sqrt{\mathbb{P}(X \leq -x)} + \sqrt{\mathbb{P}(X > x)}\,\mathrm{d}x \\
&\leq \sqrt{2}\int_{0}^{a} \sqrt{\mathbb{P}(|X| \geq x)}\,\mathrm{d}x \\
&\leq \sqrt{2}\int_{0}^{a} \sqrt{\min\left\{1, \frac{c^p}{z^p}\right\}}\,\mathrm{d}x && \text{Markov's inequality} \\
&= \sqrt{2}\left(c + c^{p/2}\int_{c}^{a} z^{-p/2}\,\mathrm{d}x\right).
\end{aligned}$$

Computing the integral yields

$$\int_{c}^{a} z^{-p/2}\,\mathrm{d}x = \frac{a^{1-p/2} - c^{1-p/2}}{1-p/2}.$$

Because $1 - p/2 < 0$, we have that $\lim_{a\to\infty} \int_{c}^{a} z^{-p/2}\,\mathrm{d}x = \frac{c^{1-p/2}}{p/2-1}$. Combining the steps above, we obtain

$$\begin{aligned}
\int_{-\infty}^{\infty} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x &= \lim_{a\to\infty} \int_{-a}^{a} \sqrt{F(x)(1-F(x))}\,\mathrm{d}x \\
&\leq \lim_{a\to\infty} \sqrt{2}\left(c + c^{p/2}\int_{c}^{a} z^{-p/2}\,\mathrm{d}x\right) \\
&= \sqrt{2}c\left(1 + \frac{1}{p/2-1}\right) \\
&= \sqrt{2}\frac{pc}{p-2}.
\end{aligned}$$

Resubstituting $c = \mathbb{E}\left[|X|^p\right]^{1/p}$ completes the proof. □

A.2 Convergence Analysis

The results of this section accompany the analysis in Section 2.6.

A.2.1 Intermediate Results

We first prove the generalized descent lemma, which forms the backbone of the argument in both the large and small shift cost settings.

Proposition 2.6.1 (Bias Bound)**.** *Consider any* $\boldsymbol{\theta} \in \mathbb{R}^d$, $\boldsymbol{l} \in \mathbb{R}^n$, *and* $\bar{\boldsymbol{q}} \in \mathcal{Q}$. *Define*

$$\boldsymbol{q} := q^{\mathrm{opt}}(\boldsymbol{l}) = \arg\max_{\boldsymbol{p} \in \mathcal{Q}} \langle \boldsymbol{p}, \boldsymbol{l} \rangle - \nu \operatorname{Reg}(\boldsymbol{p}).$$

For any $\alpha_1 \in [0, 1]$,

$$\begin{aligned}
&- (\nabla r(\boldsymbol{\theta})^\top \boldsymbol{q} - \nabla r(\boldsymbol{\theta}^\star)^\top \bar{\boldsymbol{q}})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\
&\leq -(\boldsymbol{q} - \bar{\boldsymbol{q}})^\top (\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}^\star)) - \frac{\mu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \\
&- \frac{\alpha_1}{4(M+\mu)\kappa_{\mathcal{Q}}} \frac{1}{n} \sum_{i=1}^n \|n q_i \nabla r_i(\boldsymbol{\theta}) - n q_i^\star \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{2\alpha_1 G^2}{\nu(M+\mu)\kappa_{\mathcal{Q}}} n (\boldsymbol{q} - \boldsymbol{q}^\star)^\top (\boldsymbol{l} - \boldsymbol{l}^\star).
\end{aligned}$$

Proof. First, for any $q_i > 0$, we have that $w \mapsto q_i r_i(\boldsymbol{\theta})$ is $(q_i M)$-smooth and $(q_i \mu)$-strongly convex. Define the notation $\sigma_n = \kappa_{\mathcal{Q}}/n$. By applying standard convex inequalities (Lemma A.1.1) we have that

$$\begin{aligned}
q_i r_i(\boldsymbol{\theta}^\star) &\geq q_i r_i(\boldsymbol{\theta}) + q_i \nabla r_i(\boldsymbol{\theta})^\top (\boldsymbol{\theta}^\star - \boldsymbol{\theta}) \\
&\quad + \frac{1}{2q_i(M+\mu)} \|q_i \nabla r_i(\boldsymbol{\theta}) - q_i \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{q_i \mu}{4} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 \\
&\geq q_i r_i(\boldsymbol{\theta}) + q_i \nabla r_i(\boldsymbol{\theta})^\top (\boldsymbol{\theta}^\star - \boldsymbol{\theta}) \\
&\quad + \frac{1}{2\sigma_n(M+\mu)} \|q_i \nabla r_i(\boldsymbol{\theta}) - q_i \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{q_i \mu}{4} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2
\end{aligned}$$

as $q_i \leq \sigma_n$. The second inequality holds for $q_i = 0$ as well, so by summing the inequality over i and using that $\sum_i q_i = 1$, we have that

$$\begin{aligned}
q^\top r(\boldsymbol{\theta}^\star) &\geq q^\top r(\boldsymbol{\theta}) + q^\top \nabla r(\boldsymbol{\theta})(\boldsymbol{\theta}^\star - \boldsymbol{\theta}) \\
&\quad + \frac{1}{2\sigma_n(M+\mu)} \sum_{i=1}^n \|q_i \nabla r_i(\boldsymbol{\theta}) - q_i \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{\mu}{4} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 .
\end{aligned}$$

Applying the same argument replacing q by $\bar{q}$ and swapping w and $\boldsymbol{\theta}^\star$ yields

$$\begin{aligned}
\bar{\boldsymbol{q}}^\top r(\boldsymbol{\theta}) &\geq \bar{\boldsymbol{q}}^\top r(\boldsymbol{\theta}^\star) + \bar{\boldsymbol{q}}^\top \nabla r(\boldsymbol{\theta}^\star)(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\
&\quad + \frac{1}{2\sigma_n(M+\mu)} \sum_{i=1}^n \|\bar{q}_i \nabla r_i(\boldsymbol{\theta}) - \bar{q}_i \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{\mu}{4} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2 .
\end{aligned}$$

Summing the two inequalities yields

$$\begin{aligned}
&-(\boldsymbol{q}-\bar{\boldsymbol{q}})^\top(r(\boldsymbol{\theta})-r(\boldsymbol{\theta}^\star)) \\
&\geq -\left(\nabla r(\boldsymbol{\theta})^\top \boldsymbol{q}-\nabla r(\boldsymbol{\theta}^\star)^\top \bar{q}\right)^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)+\frac{\mu}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right\|_2^2 \\
&+\frac{1}{2\sigma_n(M+\mu)}\left[\sum_{i=1}^n\left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2+\sum_{i=1}^n\left\|\bar{q}_i\nabla r_i(\boldsymbol{\theta})-\bar{q}_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2\right].
\end{aligned}$$

Dropping the $\sum_{i=1}^n \left\|\bar{q}_i\nabla r_i(\boldsymbol{\theta})-\bar{q}_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2$ term and applying a weight of $\alpha_1 \in [0,1]$ to $\sum_{i=1}^n \left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2$ still satisfies the inequality, which can equivalently be written as

$$\begin{aligned}
-\left(\nabla r(\boldsymbol{\theta})^\top \boldsymbol{q}-\nabla r(\boldsymbol{\theta}^\star)^\top \bar{q}\right)^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^\star) \leq -(\boldsymbol{q}-\bar{\boldsymbol{q}})^\top(r(\boldsymbol{\theta})-r(\boldsymbol{\theta}^\star))-\frac{\mu}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\right\|_2^2 \\
-\frac{\alpha_1}{2\sigma_n(M+\mu)}\sum_{i=1}^n\left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2. \quad \text{(A.2)}
\end{aligned}$$

Next, because

$$\left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i^\star\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2 \leq 2\left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2+2(q_i-q_i^\star)^2\left\|\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2,$$

we have that (by summing over i) that

$$-\sum_{i=1}^n\left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2 \leq -\frac{1}{2}\sum_{i=1}^n\left\|q_i\nabla r_i(\boldsymbol{\theta})-q_i^\star\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2^2+4G^2\left\|\boldsymbol{q}-\boldsymbol{q}^\star\right\|_2^2, \quad \text{(A.3)}$$

where we used that each $\left\|\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2 \leq 2G$. To see this, use that $\nabla r(\boldsymbol{\theta}^\star)^\top \boldsymbol{q}^\star = 0$ and $\nabla r(\boldsymbol{\theta}^\star) = \nabla \ell(\boldsymbol{\theta}^\star) + \mu\boldsymbol{\theta}^\star$, so

$$\left\|\nabla r_i(\boldsymbol{\theta}^\star)\right\|_2 = \left\|\nabla \ell_i(\boldsymbol{\theta}^\star)+\mu\boldsymbol{\theta}^\star\right\|_2 = \left\|\nabla \ell_i(\boldsymbol{\theta}^\star)-\textstyle\sum_{j=1}^n q_i^\star\nabla \ell_j(\boldsymbol{\theta}^\star)\right\|_2 \leq 2G.$$

Because the map q^{opt} is the gradient of a convex and $(1/\nu)$-smooth map, we also have that

$$\left\|\boldsymbol{q}-\boldsymbol{q}^\star\right\|_2^2 = \left\|q^{\text{opt}}(\boldsymbol{l})-q^{\text{opt}}(\ell(\boldsymbol{\theta}^\star))\right\|_2^2 \leq \frac{1}{\nu}(\boldsymbol{q}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}-\ell(\boldsymbol{\theta}^\star)), \quad \text{(A.4)}$$

so we apply the above to (A.3) to achieve

$$\begin{aligned}
&-\sum_{i=1}^{n} \|q_i \nabla r_i(\boldsymbol{\theta}) - q_i \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 \\
&\leq -\frac{1}{2}\sum_{i=1}^{n} \|q_i \nabla r_i(\boldsymbol{\theta}) - q_i^\star \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 + \frac{4G^2}{\nu}(\boldsymbol{q}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}-\ell(\boldsymbol{\theta}^\star)),
\end{aligned} \tag{A.5}$$

We also use (A.4) to claim non-negativity of $(\boldsymbol{q}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}-\ell(\boldsymbol{\theta}^\star))$. Finally, because $\sum_i q_i = \sum_i q_i^\star = 1$, we have that

$$\begin{aligned}
(\boldsymbol{q}-\bar{\boldsymbol{q}})^\top(r(\boldsymbol{\theta})-r(\boldsymbol{\theta}^\star)) &= (\boldsymbol{q}-\bar{\boldsymbol{q}})^\top\left(\ell(\boldsymbol{\theta}) + \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2\mathbf{1} - \ell(\boldsymbol{\theta}^\star) - \frac{\mu}{2}\|\boldsymbol{\theta}^\star\|_2^2\mathbf{1}\right) \\
&= (\boldsymbol{q}-\bar{\boldsymbol{q}})^\top\left(\ell(\boldsymbol{\theta})-\ell(\boldsymbol{\theta}^\star)\right) + (\boldsymbol{q}-\bar{\boldsymbol{q}})^\top\mathbf{1}\left(\|\boldsymbol{\theta}\|_2^2 - \|\boldsymbol{\theta}^\star\|_2^2\right) \\
&= (\boldsymbol{q}-\bar{\boldsymbol{q}})^\top\left(\ell(\boldsymbol{\theta})-\ell(\boldsymbol{\theta}^\star)\right).
\end{aligned} \tag{A.6}$$

Combine (A.2), (A.5), and (A.6) along with $\kappa_{\mathcal{Q}} = n\sigma_n$ to achieve the claim. □

The upcoming results provide the upper bounds for the Lyapunov function terms introduced in (2.26).

Lemma 2.6.3. *For any value of $\alpha_2 > 0$, we have that*

$$\begin{aligned}
\mathbb{E}_k\left[U^{(k+1)}\right] \leq{}& \eta^2(1+\alpha_2)Q^{(k)} + \eta^2(1+\alpha_2^{-1})S^{(k)} \\
&+ \frac{\eta M^2}{\mu n}\left(1-\frac{1}{n}\right)T^{(k)} + \left(1-\frac{1}{n}\right)\frac{G^2}{2\nu\mu n}R^{(k)} + \left(1-\frac{1}{n}\right)U^{(k)}.
\end{aligned}$$

Proof. First, we apply Condition 2.5.1 to the functions

$$h_i(\boldsymbol{u},\boldsymbol{x}) = \frac{1}{n}\|\boldsymbol{u}-\boldsymbol{x}\|_2^2$$

to achieve the equality

$$\begin{aligned}
\mathbb{E}_k\left[U^{(k+1)}\right] &= \frac{1}{n}\mathbb{E}_k\left[\|\boldsymbol{\theta}^{(k+1)}-\boldsymbol{\theta}^{(k)}\|_2^2\right] + \left(1-\frac{1}{n}\right)\mathbb{E}_k\left[\frac{1}{n}\sum_{j=1}^{n}\|\boldsymbol{\theta}^{(k+1)}-\hat{\boldsymbol{\theta}}_j^{(k)}\|_2^2\right] \\
&= \frac{\eta^2}{n}\mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] + \left(1-\frac{1}{n}\right)\mathbb{E}_k\left[\frac{1}{n}\sum_{j=1}^{n}\|\boldsymbol{\theta}^{(k+1)}-\hat{\boldsymbol{\theta}}_j^{(k)}\|_2^2\right].
\end{aligned}$$

Next, we expand the second term.

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_k\left[\sum_{j=1}^{n}\|\boldsymbol{\theta}^{(k+1)}-\hat{\boldsymbol{\theta}}_j^{(k)}\|_2^2\right] \\
&=\frac{1}{n}\mathbb{E}_k\left[\sum_{j=1}^{n}\|\boldsymbol{\theta}^{(k+1)}-\boldsymbol{\theta}^{(k)}\|_2^2\right]+\frac{2}{n}\mathbb{E}_k\left[\sum_{j=1}^{n}(\boldsymbol{\theta}^{(k+1)}-\boldsymbol{\theta}^{(k)})^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)})\right]+\frac{1}{n}\mathbb{E}_k\left[\sum_{j=1}^{n}\|\hat{\boldsymbol{\theta}}_j^{(k)}-\boldsymbol{\theta}^{(k)}\|_2^2\right] \\
&=\eta^2\mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right]-\frac{2\eta}{n}\sum_{j=1}^{n}\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^{(k)})^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)})+\frac{1}{n}\sum_{j=1}^{n}\|\hat{\boldsymbol{\theta}}_j^{(k)}-\boldsymbol{\theta}^{(k)}\|_2^2.
\end{aligned}$$

The first term is simply the noise term that appears in Lemma 2.6.1, whereas the last term is $U^{(k)}$. Next, we have

$$\begin{aligned}
-2\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^{(k)})^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}) &= -2(\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^{(k)})-\nabla({\boldsymbol{q}^{(k)}}^\top r)(\hat{\boldsymbol{\theta}}_j^{(k)}))^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}) \\
&\quad -2(\nabla({\boldsymbol{q}^{(k)}}^\top r)(\hat{\boldsymbol{\theta}}_j^{(k)})-\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^\star))^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}) \\
&\quad -2(\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^\star)-\nabla({\boldsymbol{q}^\star}^\top r)(\boldsymbol{\theta}^\star))^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}),
\end{aligned}$$

where the last term is introduced because $\nabla({\boldsymbol{q}^\star}^\top r)(\boldsymbol{\theta}^\star)=0$. We bound each of the three terms. First,

$$-2(\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^{(k)})-\nabla({\boldsymbol{q}^{(k)}}^\top r)(\hat{\boldsymbol{\theta}}_j^{(k)}))^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)})\le -2\mu\left\|\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}\right\|_2^2$$

because ${\boldsymbol{q}^{(k)}}^\top r$ is μ-strongly convex [Nesterov, 2018, Theorem 2.1.9]. Second,

$$\begin{aligned}
&-2(\nabla({\boldsymbol{q}^{(k)}}^\top r)(\hat{\boldsymbol{\theta}}_j^{(k)})-\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^\star))^\top(\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}) \\
&\le \alpha_4\|\nabla({\boldsymbol{q}^{(k)}}^\top r)(\hat{\boldsymbol{\theta}}_j^{(k)})-\nabla({\boldsymbol{q}^{(k)}}^\top r)(\boldsymbol{\theta}^\star)\|_2^2+\alpha_4^{-1}\|\hat{\boldsymbol{\theta}}_j^{(k)}-\boldsymbol{\theta}^{(k)}\|_2^2 \\
&\le \alpha_4 M^2\|\hat{\boldsymbol{\theta}}_j^{(k)}-\boldsymbol{\theta}^\star\|_2^2+\alpha_4^{-1}\|\hat{\boldsymbol{\theta}}_j^{(k)}-\boldsymbol{\theta}^{(k)}\|_2^2
\end{aligned}$$

by Young's inequality with parameter α_4 and the M-Lipschitz continuity of $\nabla(\boldsymbol{q}^{(k)\top} r)$. Third,

$$\begin{aligned}
&- 2(\nabla(\boldsymbol{q}^{(k)\top} r)(\boldsymbol{\theta}^\star) - \nabla(\boldsymbol{q}^{\star\top} r)(\boldsymbol{\theta}^\star))^\top (\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_j^{(k)}) \\
&= -2(\nabla((\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top \ell)(\boldsymbol{\theta}^\star))^\top (\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_j^{(k)}) \\
&\leq \alpha_5 \|\nabla((\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top \ell)(\boldsymbol{\theta}^\star)\|_2^2 + \alpha_5^{-1} \|\hat{\boldsymbol{\theta}}_j^{(k)} - \boldsymbol{\theta}^{(k)}\|_2^2 \\
&\leq \alpha_5 G^2 \|\boldsymbol{q}^{(k)} - \boldsymbol{q}\|_2^2 + \alpha_5^{-1} \|\hat{\boldsymbol{\theta}}_j^{(k)} - \boldsymbol{\theta}^{(k)}\|_2^2,
\end{aligned}$$

by Young's inequality with parameter α_5 and the G-Lipschitz continuity of each ℓ_i. Combining with the above, we have

$$\begin{aligned}
-2\sum_{j=1}^n \nabla(\boldsymbol{q}^{(k)\top} r)(\boldsymbol{\theta}^{(k)})^\top (\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_j^{(k)}) &\leq \alpha_4 M^2 T^{(k)} + (\alpha_4^{-1} + \alpha_5^{-1} - 2\mu) U^{(k)} + \alpha_5 G^2 n \|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\|_2^2 \\
&\leq \mu^{-1} M^2 T^{(k)} + \mu^{-1} G^2 n \|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\|_2^2
\end{aligned}$$

when we set $\alpha_4 = \alpha_5 = \mu^{-1}$. Hence, we get

$$\begin{aligned}
&\mathbb{E}_k\left[U^{(k+1)}\right] \\
&= \frac{\eta^2}{n} \mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] + \left(1 - \frac{1}{n}\right) \mathbb{E}_k\left[\frac{1}{n}\sum_{j=1}^n \|\boldsymbol{\theta}^{(k+1)} - \hat{\boldsymbol{\theta}}_j^{(k)}\|_2^2\right] \\
&\leq \eta^2 \mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] - \frac{\eta}{n}\left(1 - \frac{1}{n}\right) 2\sum_{j=1}^n \nabla(\boldsymbol{q}^{(k)\top} r)(\boldsymbol{\theta}^{(k)})^\top (\boldsymbol{\theta}^{(k)} - \hat{\boldsymbol{\theta}}_j^{(k)}) + \left(1 - \frac{1}{n}\right) U^{(k)} \\
&\leq \eta^2 \mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] + \frac{\eta}{n}\left(1 - \frac{1}{n}\right)\left[\mu^{-1} M^2 T^{(k)} + \mu^{-1} G^2 n \|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\|_2^2\right] + \left(1 - \frac{1}{n}\right) U^{(k)} \\
&= \eta^2 \mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] + \left(1 - \frac{1}{n}\right)\frac{\eta M^2}{\mu n} T^{(k)} + \left(1 - \frac{1}{n}\right)\frac{G^2}{2n\mu} 2n\eta \|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\|_2^2 + \left(1 - \frac{1}{n}\right) U^{(k)} \\
&= \eta^2 \mathbb{E}_k\left[\|\boldsymbol{v}^{(k)}\|_2^2\right] + \left(1 - \frac{1}{n}\right)\frac{\eta M^2}{\mu n} T^{(k)} + \left(1 - \frac{1}{n}\right)\frac{G^2}{2n\mu\nu} R^{(k)} + \left(1 - \frac{1}{n}\right) U^{(k)} \\
&\leq \eta^2(1 + \alpha_2) Q^{(k)} + \eta^2(1 + \alpha_2^{-1}) S^{(k)} \\
&+ \frac{\eta M^2}{\mu n}\left(1 - \frac{1}{n}\right) T^{(k)} + \left(1 - \frac{1}{n}\right)\frac{G^2}{2\nu\mu n} R^{(k)} + \left(1 - \frac{1}{n}\right) U^{(k)},
\end{aligned}$$

where the two last steps follows Lemma 2.4.1 and Theorem A.1.1 to claim $\|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\|_2^2 \leq$

$\frac{1}{\bar{\nu}}(\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)(\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star)$. □

Lemma 2.6.4. *For any $\alpha_3 > 0$, it holds that*

$$\mathbb{E}_k\left[R^{(k+1)}\right] \leq 2\eta(\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top(\ell(\boldsymbol{\theta}^{(k)}) - \boldsymbol{l}^\star) + \left(1 - \frac{1}{n}\right) R^{(k)} + \frac{\eta G^2 n}{2\nu}\alpha_3^{-1} T^{(k)} + \frac{2\eta G^2 n}{\nu}(1 + \alpha_3) U^{(k)}.$$

Proof. First, decompose

$$(\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k+1)} - \boldsymbol{l}^\star) = (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k+1)} - \boldsymbol{l}^\star) + (\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)})^\top(\boldsymbol{l}^{(k+1)} - \boldsymbol{l}^{(k)}) \tag{A.7}$$

$$+ (\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)})^\top(\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star). \tag{A.8}$$

Because $\boldsymbol{q}^{(k)} = q^{\mathrm{opt}}(\boldsymbol{l}^{(k)})$ for all t, and $q^{\mathrm{opt}}\cdot)$ is the gradient of a convex and $(1/\nu)$-smooth function, we have for the second term of (A.8) that

$$(\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)})^\top(\boldsymbol{l}^{(k+1)} - \boldsymbol{l}^{(k)}) \leq \frac{1}{\bar{\nu}}\|\boldsymbol{l}^{(k+1)} - \boldsymbol{l}^{(k)}\|_2^2.$$

Next, using Young's inequality, that is, $a^\top b \leq \frac{\alpha_3}{2}\|a\|_2^2 + \frac{\alpha_3^{-1}}{2}\|b\|_2^2$ for any $\alpha_3 > 0$, we have for the third term of (A.8) that

$$\begin{aligned}(\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)})^\top(\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star) &\leq \frac{\alpha_3}{2}\|\boldsymbol{q}^{(k+1)} - \boldsymbol{q}^{(k)}\|_2^2 + \frac{\alpha_3^{-1}}{2}\|\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star\|_2^2 \\ &\leq \frac{\alpha_3}{2\bar{\nu}^2}\|\boldsymbol{l}^{(k+1)} - \boldsymbol{l}^{(k)}\|_2^2 + \frac{\alpha_3^{-1}}{2}\|\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star\|_2^2.\end{aligned}$$

Note that we have

$$\mathbb{E}_k\left[\boldsymbol{l}^{(k+1)}\right] = \frac{1}{n}\ell(\boldsymbol{\theta}^{(k)}) + \left(1 - \frac{1}{n}\right)\boldsymbol{l}^{(k)}.$$

Hence, we get,

$$\begin{aligned}
\frac{1}{2\eta n}\mathbb{E}_k\left[R^{(k+1)}\right] &= \frac{1}{n}(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)^\top(l(\boldsymbol{\theta}^{(k)})-\boldsymbol{l}^\star)+\left(1-\frac{1}{n}\right)(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k)}-\boldsymbol{l}^\star)\\
&\quad+\mathbb{E}_k\left[(\boldsymbol{q}^{(k+1)}-\boldsymbol{q}^{(k)})^\top(\boldsymbol{l}^{(k+1)}-\boldsymbol{l}^{(k)})\right]+\mathbb{E}_k\left[(\boldsymbol{q}^{(k+1)}-\boldsymbol{q}^{(k)})^\top(\boldsymbol{l}^{(k)}-\boldsymbol{l}^\star)\right]\\
&\le \frac{1}{n}(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)^\top(l(\boldsymbol{\theta}^{(k)})-\boldsymbol{l}^\star)+\left(1-\frac{1}{n}\right)(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k)}-\boldsymbol{l}^\star)\\
&\quad+\left(\frac{1}{\bar\nu}+\frac{\alpha_3}{2\bar\nu^2}\right)\mathbb{E}_k\left[\|\boldsymbol{l}^{(k+1)}-\boldsymbol{l}^{(k)}\|_2^2\right]+\frac{\alpha_3^{-1}}{2}\|\boldsymbol{l}^{(k)}-\boldsymbol{l}^\star\|_2^2\\
&= \frac{1}{n}(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)(\ell(\boldsymbol{\theta}^{(k)})-\boldsymbol{l}^\star)+\left(1-\frac{1}{n}\right)(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k)}-\boldsymbol{l}^\star)\\
&\quad+\frac{1}{n\bar\nu}\left(1+\frac{\alpha_3}{2\bar\nu}\right)\sum_{j=1}^n(\ell_j(\boldsymbol{\theta}^{(k)})-\ell_j(\hat{\boldsymbol{\theta}}_j))^2\\
&\quad+\frac{\alpha_3^{-1}}{2}\sum_{j=1}^n(\ell_j(\hat{\boldsymbol{\theta}}_j)-\ell_j(\boldsymbol{\theta}^\star))^2.
\end{aligned}$$

Then, apply the G-Lipschitz continuity of each ℓ_i to achieve

$$\begin{aligned}
\frac{1}{2\eta n}\mathbb{E}_k\left[R^{(k+1)}\right] &\le \frac{1}{n}(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)(\ell(\boldsymbol{\theta}^{(k)})-\boldsymbol{l}^\star)+\left(1-\frac{1}{n}\right)(\boldsymbol{q}^{(k)}-\boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k)}-\boldsymbol{l}^\star)\\
&\quad+\frac{G^2}{n\bar\nu}\left(1+\frac{\alpha_3}{2\bar\nu}\right)\sum_{j=1}^n\|\boldsymbol{\theta}^{(k)}-\hat{\boldsymbol{\theta}}_j^{(k)}\|_2^2\\
&\quad+\frac{G^2\alpha_3^{-1}}{2}\sum_{j=1}^n\|\hat{\boldsymbol{\theta}}_j^{(k)}-\boldsymbol{\theta}^\star\|_2^2.
\end{aligned}$$

Replacing α_3 by $2\bar\nu\alpha_3$ gives the claim. □

A.2.2 Proof of Main Results

The forthcoming theorems provide the complete convergence analyses for the settings in which 1) the shift cost ν satisfies a particular lower bound and 2) when that bound is violated, respectively.

Theorem 2.5.2. *Assume that $n \ge 2$ and that the shift cost $\nu \le 8nG^2/\mu$. The sequence of*

iterates produced by Algorithm 1 with

$$\eta = \frac{1}{16n\mu} \min\left\{ \frac{1}{6[8\delta + (\kappa+1)\kappa_{\mathcal{P}}]}, \frac{1}{4\delta^2 \max\{2n\kappa^2, \delta\}} \right\}$$

achieves

$$\mathbb{E}_0 \left\| \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^\star \right\|_2^2 \leq \left(5 + 16\delta + \frac{6\kappa^2}{\sigma_n} \right) \exp\left(-t/\tau\right) \left\| \boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star \right\|_2^2$$

for

$$\tau = 32n \max\left\{ 6[8\delta + (\kappa+1)\kappa_{\mathcal{P}}], 4\delta^2 \max\left\{2n\kappa^2, \delta\right\}, 1/16 \right\}.$$

Proof. First, invoke Lemma 2.6.1 with $\boldsymbol{q}' = \boldsymbol{q}^{(k)}$ and $\alpha_1 = 1$ to obtain

$$\mathbb{E}_k \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^\star\|_2^2 \leq (1 - \eta\mu) \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \tag{A.9}$$

$$- 2\eta (\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^\star) \boldsymbol{q}^{(k)} + \frac{2G^2}{\bar{\nu}(M+\mu)\kappa_\sigma} R^{(k)} \tag{A.10}$$

$$- \eta \left(\frac{1}{2(M+\mu)\kappa_\sigma} - \eta(1+\alpha_2) \right) Q^{(k)} + \eta^2 (1 + \alpha_2^{-1}) S^{(k)}. \tag{A.11}$$

We will first bound (A.10), by using that $\nabla r(\boldsymbol{\theta}^\star)\boldsymbol{q}^\star = 0$ and Young's inequality with parameter $a > 0$ to write

$$\begin{aligned} \left| (\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^\star) \boldsymbol{q}^{(k)} \right| &= \left| (\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^\star) (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star) \right| \\ &\leq \frac{a}{2} \left\| \nabla r(\boldsymbol{\theta}^\star)^\top (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star) \right\|_2^2 + \frac{1}{2a} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\ &\leq \frac{aG^2\gamma_\star^2}{2\nu^2} T^{(k)} + \frac{1}{2a} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2, \end{aligned}$$

where we used in the second inequality that:

$$\begin{aligned} \left\| \nabla r(\boldsymbol{\theta}^\star)^\top (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star) \right\|_2^2 = \left\| \nabla \ell(\boldsymbol{\theta}^\star)^\top (\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star) \right\|_2^2 &\leq \gamma_\star^2 \|\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star\|_2^2 \leq \frac{\gamma_\star^2}{\nu^2} \|\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star\|_2^2 \\ &\leq \frac{G^2\gamma_\star^2}{\nu^2} \sum_{i=1}^n \|\hat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}^\star\|_2^2 = \frac{G^2\gamma_\star^2}{\nu^2} T^{(k)}. \end{aligned}$$

We also have by Cauchy-Schwartz and Lipschitz continuity that

$$R^{(k)} = 2\eta n(\boldsymbol{q}^{(k)} - \boldsymbol{q}^\star)^\top(\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star) \leq \frac{2\eta n}{\nu}\left\|\boldsymbol{l}^{(k)} - \boldsymbol{l}^\star\right\|_2^2 \leq \frac{2\eta n G^2}{\nu} T^{(k)}.$$

Combining the above displays yields

$$\begin{aligned}
&- 2\eta(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star)^\top \nabla r(\boldsymbol{\theta}^\star)\boldsymbol{q}^{(k)} + \frac{2G^2}{\bar{\nu}(M+\mu)\kappa_\sigma} R^{(k)} \\
&\leq \frac{\eta G^2}{\nu^2}\left[a\gamma_\star^2 + \frac{4nG^2}{(M+\mu)\kappa_\sigma}\right] T^{(k)} + \frac{\eta}{a}\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2.
\end{aligned}$$

We take $\alpha_2 = 2$, $c_3 = c_4 = 0$, and apply Lemma 2.6.2 to achieve

$$\begin{aligned}
\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq{}& \left[\tau^{-1} - \eta\mu + \eta a^{-1} + c_2\right]\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2 \\
&+ \left[\tau^{-1} + \frac{3\eta^2}{2c_1} - \frac{1}{n}\right] c_1 S^{(k)} \\
&+ \left[\tau^{-1} + \frac{\eta G^2}{\nu^2 c_2}\left(a\gamma_\star^2 + \frac{4nG^2}{(M+\mu)\kappa_\sigma}\right) - \frac{1}{n}\right] c_2 T^{(k)} \\
&+ \left[-\frac{\eta}{2(M+\mu)\kappa_\sigma} + 3\eta^2 + \frac{c_1}{n}\right] Q^{(k)},
\end{aligned}$$

where $\tau > 0$ is a to-be-specified rate constant. We now need to set the various free parameters a, c_1, c_2, and η to make each of the squared bracketed terms be non-positive. We enforce $\tau \geq 2n$ throughout. By setting

$$\eta = \frac{1}{12(\mu + M)\kappa_\sigma} \text{ and } c_1 = \frac{n\eta}{4(\mu+M)\kappa_\sigma},$$

we have that the bracketed constants before $c_1 S^{(k)}$ and $Q^{(k)}$ vanish. Then, setting

$$a^{-1} = \frac{\mu}{2} \text{ and } c_2 = \frac{1}{48(\kappa+1)\kappa_\sigma}$$

make the bracketed constant before $\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2$, assuming that we enforce

$$\tau \geq 48(\kappa+1)\kappa_\sigma.$$

We turn to the final constant after substituting the values of a, c_2, and η. We need that

$$\frac{\eta G^2}{\nu^2 c_2}\left(a\gamma_\star^2 + \frac{8nG^2}{(M+\mu)\kappa_\sigma}\right) = \frac{8G^2}{\nu^2\mu^2}\left(\gamma_\star^2 + \frac{2nG^2}{(\kappa+1)\kappa_\sigma}\right) \leq \frac{1}{2n},$$

which occurs when

$$\nu^2 \geq \frac{16nG^2}{\mu^2}\left[\gamma_\star^2 + \frac{2nG^2}{(\kappa+1)\kappa_\sigma}\right].$$

Because $\gamma_\star^2 \leq nG^2 \leq 2nG^2$, this is achieved when

$$\nu \geq \frac{8nG^2}{\mu},$$

completing the proof of the claim

$$\mathbb{E}_k\left[V^{(k+1)}\right] \leq (1-\tau^{-1})V^{(k)}.$$

To complete the proof, we bound the initial terms. Because $c_3 = c_4 = 0$, we need only to bound $S^{(0)}$ and $T^{(0)}$.

$$\begin{aligned}
S^{(0)} &= \frac{1}{n}\sum_{i=1}^{n} \|n\rho_i^{(0)}\nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(0)}) - nq_i^\star \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|nq_i^{(0)}\nabla r_i(\boldsymbol{\theta}^{(0)}) - nq_i^* \nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n} \|nq_i^{(0)}\nabla(r_i(\boldsymbol{\theta}^{(0)}) - \nabla r_i(\boldsymbol{\theta}^\star))\|_2^2 + \frac{2}{n}\sum_{i=1}^{n}\|n(q_i^{(0)} - q_i^\star)\nabla r_i(\boldsymbol{\theta}^\star)\|_2^2 \\
&\leq 2n\sum_{i=1}^{n}(q_i^{(0)})^2 M^2\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 + 8nG^2\|\boldsymbol{q}^{(0)} - \boldsymbol{q}^\star\|_2^2 \\
&\leq \left[2n\left\|\sigma\right\|_2^2 M^2 + \frac{8n^2G^4}{\nu^2}\right]\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 \\
&\leq \left[2n\left\|\sigma\right\|_2^2 M^2 + \mu^2/8\right]\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 \leq 3nM^2\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2.
\end{aligned}$$

This means ultimately that

$$c_1 S^{(0)} \leq \frac{n^2}{16(1+\kappa^{-1})^2\kappa_\sigma^2}\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2.$$

Next, we have

$$c_2 T^{(0)} = \frac{n}{48(\kappa+1)\kappa_\sigma} \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 .$$

Thus, we can write

$$\begin{aligned} V^{(0)} &\leq \left[1 + \frac{n^2}{16(1+\kappa^{-1})^2\kappa_\sigma^2} + \frac{n}{48(\kappa+1)\kappa_\sigma}\right] \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 \\ &\leq (1 + \sigma_n^{-1} + \sigma_n^{-2}) \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 , \end{aligned}$$

completing the proof. □

Theorem 2.5.1. *Suppose the shift cost satisfies*

$$\nu \geq 8nG^2/\mu.$$

Then, the sequence of iterates produced by Algorithm 1 with $\eta = 1/(12(\mu+M)\kappa_{\mathcal{P}})$ *achieves*

$$\mathbb{E}_0 \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^\star\|_2^2 \leq (1 + \sigma_n^{-1} + \sigma_n^{-2}) \exp(-t/\tau) \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\|_2^2 .$$

with

$$\tau = 2\max\{n, 24\kappa_{\mathcal{P}}(\kappa+1)\}.$$

Proof. First, we apply Lemma 2.6.1 with $\boldsymbol{q}' = \boldsymbol{q}^\star$, as well as Lemma 2.6.4, Lemma 2.6.2, and

Lemma 2.6.3, set $c_4 = 1$, and consolidate all constants to write

$$\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq (\tau^{-1} - \eta\mu + c_2)\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2 \tag{A.12}$$

$$+ \left[\tau^{-1} - \frac{1}{n} + \frac{2\alpha_1 G^2}{\nu(M+\mu)\kappa_\sigma} + \left(1 - \frac{1}{n}\right)\frac{G^2 c_3}{2\nu\mu n}\right] R^{(k)} \tag{A.13}$$

$$+ \left[\tau^{-1} + \frac{1+c_3}{c_1}\eta^2(1+\alpha_2^{-1}) - \frac{1}{n}\right] c_1 S^{(k)} \tag{A.14}$$

$$+ \left[\tau^{-1} + \frac{\eta G^2 n}{2c_2\nu}\alpha_3^{-1} + \frac{c_3\eta M^2}{c_2\mu n}\left(1 - \frac{1}{n}\right) - \frac{1}{n}\right] c_2 T^{(k)} \tag{A.15}$$

$$+ \left[\tau^{-1} + \frac{2\eta G^2 n}{c_3\nu}(1+\alpha_3) - \frac{1}{n}\right] c_3 U^{(k)} \tag{A.16}$$

$$+ \left[-\frac{\eta\alpha_1}{2(M+\mu)\kappa_\sigma} + \eta^2(1+c_3)(1+\alpha_2) + \frac{c_1}{n}\right] Q^{(k)}. \tag{A.17}$$

We first set $c_1 = \frac{n\eta\alpha_1}{4(M+\mu)\kappa_\sigma}$ and $c_2 = \eta\mu/2$ to clean up (A.12) and (A.17). We also drop the terms $(1 - 1/n) \leq 1$. Then, we notice in (A.13) that to achieve

$$\frac{2\alpha_1 G^2}{\nu(M+\mu)\kappa_\sigma} \leq \frac{1}{4n},$$

we need that $\alpha_1 \leq ((M+\mu)\kappa_\sigma)/(8nG^2/\nu)$. Combined with the requirement that $\alpha_1 \in [0,1]$, we set $\alpha_1 = ((M+\mu)\kappa_\sigma)/(8nG^2/\nu + (M+\mu)\kappa_\sigma)$. We set $\alpha_2 = 2$, and can rewrite the expression above.

$$\begin{aligned}
\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq{}& \left(\tau^{-1} - \frac{\eta\mu}{2}\right)\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\right\|_2^2 \\
&+ \left[\tau^{-1} - \frac{3}{4n} + \frac{G^2 c_3}{2\nu\mu n}\right] R^{(k)} \\
&+ \left[\tau^{-1} + \frac{6(1+c_3)(M+\mu)\kappa_\sigma}{n\alpha_1}\eta - \frac{1}{n}\right] c_1 S^{(k)} \\
&+ \left[\tau^{-1} + \frac{G^2 n}{\mu\nu}\alpha_3^{-1} + \frac{c_3 M^2}{\mu^2 n} - \frac{1}{n}\right] c_2 T^{(k)} \\
&+ \left[\tau^{-1} + \frac{2\eta G^2 n}{c_3\nu}(1+\alpha_3) - \frac{1}{n}\right] c_3 U^{(k)} \\
&+ \left[-\frac{\eta\alpha_1}{4(M+\mu)\kappa_\sigma} + 3\eta^2(1+c_3)\right] Q^{(k)}.
\end{aligned}$$

Next, set the learning rate to be

$$\eta \leq \frac{\alpha_1}{12(1+c_3)(M+\mu)\kappa_\sigma} \tag{A.18}$$

to cancel out $Q^{(k)}$ and achieve

$$\begin{aligned}
\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq & \left(\tau^{-1} - \frac{\eta\mu}{2}\right)\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\
& + \left[\tau^{-1} - \frac{3}{4n} + \frac{G^2 c_3}{2\nu\mu n}\right] R^{(k)} \\
& + \left[\tau^{-1} - \frac{1}{2n}\right] c_1 S^{(k)} \\
& + \left[\tau^{-1} + \frac{G^2 n}{\mu\nu}\alpha_3^{-1} + \frac{c_3 M^2}{\mu^2 n} - \frac{1}{n}\right] c_2 T^{(k)} \\
& + \left[\tau^{-1} + \frac{2\eta G^2 n}{c_3\nu}(1+\alpha_3) - \frac{1}{n}\right] c_3 U^{(k)}.
\end{aligned}$$

Requiring now that $\tau \geq 2n$, we may also cancel the $S^{(k)}$ term. We substitute $\delta = nG^2/(\mu\nu)$ to achieve

$$\begin{aligned}
\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq & \left(\tau^{-1} - \frac{\eta\mu}{2}\right)\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\
& + \left[-\frac{1}{4n} + \frac{c_3\delta}{2n^2}\right] R^{(k)} \\
& + \left[-\frac{1}{2n} + \frac{\delta}{\alpha_3} + \frac{c_3 M^2}{\mu^2 n}\right] c_2 T^{(k)} \\
& + \left[-\frac{1}{2n} + \frac{2\mu\eta\delta}{c_3}(1+\alpha_3)\right] c_3 U^{(k)}.
\end{aligned}$$

It remains to select c_3 and α_3. As such, we set $\alpha_3 = 4n\delta$ and use that $1 + 4n\delta \leq 8n\delta$ when

$n \geq 2$ and $\delta \geq 1/8$ as assumed, and so

$$\begin{aligned}\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq &\left(\tau^{-1} - \frac{\eta\mu}{2}\right)\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\ &+ \left[-\frac{1}{4n} + \frac{c_3\delta}{2n^2}\right] R^{(k)} \\ &+ \left[-\frac{1}{4n} + \frac{c_3\kappa^2}{n}\right] c_2 T^{(k)} \\ &+ \left[-\frac{1}{2n} + \frac{16n\mu\eta\delta^2}{c_3}\right] c_3 U^{(k)}.\end{aligned}$$

We require now that

$$c_3 = \frac{1}{2}\min\left\{\frac{1}{2\kappa^2}, \frac{n}{\delta}\right\},$$

which cancels $T^{(k)}$ and $R^{(k)}$, leaving

$$\begin{aligned}\mathbb{E}_k\left[V^{(k+1)}\right] - (1-\tau^{-1})V^{(k)} \leq &\left(\tau^{-1} - \frac{\eta\mu}{2}\right)\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2^2 \\ &+ \left[-\frac{1}{2n} + 32\mu\eta\delta^2\max\left\{2n\kappa^2, \delta\right\}\right] c_3 U^{(k)}.\end{aligned}$$

From the above, we retrieve the requirement that

$$\eta \leq \frac{1}{64n\mu\delta^2\max\left\{2n\kappa^2, \delta\right\}}. \tag{A.19}$$

It now remains to set η. By substituting in the values for α_1 and c_3 into (A.18), we have that

$$\begin{aligned}\eta \overset{\text{want}}{\leq} \frac{\alpha_1}{12(1+c_3)(M+\mu)\kappa_\sigma} &= \frac{1}{12(1+c_3)[8\mu\delta + (M+\mu)\kappa_\sigma]} \\ &\geq \frac{1}{(12+6n/\delta)[8\mu\delta + (M+\mu)\kappa_\sigma]} \\ &\geq \frac{1}{(12+48n)[8\mu\delta + (M+\mu)\kappa_\sigma]} \\ &\geq \frac{1}{96n[8\mu\delta + (M+\mu)\kappa_\sigma]}.\end{aligned}$$

The combination of (A.19) and the above display yields

$$\begin{aligned}\eta &= \min\left\{\frac{1}{96n[8\mu\delta+(M+\mu)\kappa_\sigma]}, \frac{1}{64n\mu\delta^2\max\{2n\kappa^2,\delta\}}\right\}\\ &= \frac{1}{16n\mu}\min\left\{\frac{1}{6[8\delta+(\kappa+1)\kappa_\sigma]}, \frac{1}{4\delta^2\max\{2n\kappa^2,\delta\}}\right\}.\end{aligned}$$

We need finally that $\tau \geq 2/(\mu\eta)$, resulting in the requirement

$$\tau \geq 32n\max\left\{6[8\delta+(\kappa+1)\kappa_\sigma], 4\delta^2\max\left\{2n\kappa^2,\delta\right\}\right\}.$$

This is achieved by setting

$$\tau = 32n\max\left\{6[8\delta+(\kappa+1)\kappa_\sigma], 4\delta^2\max\left\{2n\kappa^2,\delta\right\}, 1/16\right\}.$$

completing the proof of the claim

$$\mathbb{E}_k\left[V^{(k+1)}\right] \leq (1-\tau^{-1})V^{(k)}.$$

Next, we bound the initial terms to achieve the final rate. First, we bound η which is used in all of the terms. Because $\delta \geq 1/8$,

$$\eta \leq \frac{1}{16n\mu}\cdot\frac{1}{4\delta^2\max\{2n\kappa^2,\delta\}} \leq \frac{1}{64n\mu\delta^3} \leq \frac{8}{n\mu}. \tag{A.20}$$

Then,

$$
\begin{aligned}
S^{(0)} &= \frac{1}{n}\sum_{i=1}^{n} \|n\rho_i^{(0)}\nabla r_i(\tilde{\boldsymbol{\theta}}_i^{(0)}) - nq_i^{\star}\nabla r_i(\boldsymbol{\theta}^{\star})\|_2^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \|nq_i^{(0)}\nabla r_i(\boldsymbol{\theta}^{(0)}) - nq_i^{*}\nabla r_i(\boldsymbol{\theta}^{\star})\|_2^2 \\
&\le \frac{2}{n}\sum_{i=1}^{n} \|nq_i^{(0)}\nabla (r_i(\boldsymbol{\theta}^{(0)}) - \nabla r_i(\boldsymbol{\theta}^{\star}))\|_2^2 + \frac{2}{n}\sum_{i=1}^{n} \|n(q_i^{(0)} - q_i^{\star})\nabla r_i(\boldsymbol{\theta}^{\star})\|_2^2 \\
&\le 2n\sum_{i=1}^{n}(q_i^{(0)})^2 M^2 \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2 + 8nG^2\|\boldsymbol{q}^{(0)} - \boldsymbol{q}^{\star}\|_2^2 \\
&\le \left[2n\,\|\sigma\|_2^2\, M^2 + \frac{8n^2G^2}{\nu^2}\right]\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2 \\
&\le \left[2n\,\|\sigma\|_2^2\, M^2 + \mu^2/8\right]\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2 \le 3nM^2\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2.
\end{aligned}
$$

Continuing with $\alpha_1 \le 1$ and (A.20),

$$
\begin{aligned}
c_1 S^{(0)} &= \frac{n\eta\alpha_1}{4(M+\mu)\kappa_\sigma} S^{(0)} \\
&\le \frac{2}{\mu(M+\mu)\kappa_\sigma}\cdot 3nM^2\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2 \\
&\le \frac{6n\kappa^2}{(1+\kappa)\kappa_\sigma}\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2 \\
&\le \frac{6\kappa^2}{\sigma_n}\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2.
\end{aligned}
$$

Next, we have $T^{(0)} = n\,\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2$ and by (A.20),

$$
\begin{aligned}
c_2 T^{(0)} &= \frac{\eta\mu}{2}\cdot n\,\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2 \\
&\le 4\,\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{\star}\|_2^2\,.
\end{aligned}
$$

Because $U^{(0)} = 0$, it is bounded trivially. For $R^{(0)}$, with $c_4 = 1$ we have

$$\begin{aligned} R^{(0)} &= 2n\eta(q^{\text{opt}}\ell(\boldsymbol{\theta}^{(0)})) - q^{\text{opt}}\ell(\boldsymbol{\theta}^\star)))^\top(\ell(\boldsymbol{\theta}^{(0)}) - \ell(\boldsymbol{\theta}^\star)) \\ &\leq \frac{2n\eta}{\nu} \left\|\ell(\boldsymbol{\theta}^{(0)}) - \ell(\boldsymbol{\theta}^\star)\right\|_2^2 \\ &\leq \frac{2n^2\eta G^2}{\nu} \left\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\right\|_2^2 \\ &\leq \frac{16nG^2}{\mu\nu} \left\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\right\|_2^2 \\ &= 16\delta \left\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\right\|_2^2. \end{aligned}$$

Combining each of these terms together, we have that

$$V^{(0)} \leq \left(5 + 16\delta + \frac{6\kappa^2}{\sigma_n}\right) \left\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^\star\right\|_2^2,$$

completing the proof. □

A.3 Implementation Details

In this section, we describe Prospect including computational details, in a way that is amenable to implementation. Particular attention is given to the case when $\mathcal{Q} \equiv \mathcal{Q}(\sigma)$ is the spectral risk measure uncertainty set and the penalty is an f-divergence.

Efficient Implementation We exactly solve the maximization problem

$$\boldsymbol{q} = q^{\text{opt}}(\boldsymbol{l}) = \operatorname*{arg\,max}_{\boldsymbol{q} \in \mathcal{Q}(\sigma)} \left\{ \langle \boldsymbol{q}, \boldsymbol{l} \rangle - (\nu/n) \sum_{i=1}^{n} f(nq_i) \right\}. \tag{A.21}$$

by a sequence of three steps:

- **Sorting:** Find π such that $l_{\pi(1)} \leq \ldots \leq l_{\pi(n)}$.
- **Isotonic regression:** Apply Pool Adjacent Violators (PAV) (Subroutine 1) to solve the isotonic regression minimization problem (2.28), yielding solution $\boldsymbol{z} = z^{\text{opt}}(\boldsymbol{l})$.
- **Conversion:** Use (2.29) to convert $\boldsymbol{z}$ back to $\boldsymbol{q} = q^{\text{opt}}(\boldsymbol{l})$.

The sorting step runs in $O(n \ln n)$ elementary operations, whereas the isotonic regression and conversion steps run in $O(n)$ operations. Crucially, retrieving $\boldsymbol{q}$ from the output $\boldsymbol{z} = z^{\text{opt}}(\boldsymbol{l})$ in the third step can be done by a single $O(n)$-time pass by setting

$$q_{\pi(i)} = \frac{1}{n}[f^*]'\left(\tfrac{1}{\nu}(l_{\pi(i)} - z_i)\right)$$

for $i = 1, \ldots, n$, as opposed to computing the inverse π^{-1} and using (2.29) directly, which in fact requires another sorting operation and can be avoided. Because only one element of l changes on every iteration, we may sort it by simply bubbling the value of the index that changed into its correct position to generate the newly sorted version. The full algorithm is given Algorithm 6. We give a brief explanation on the PAV algorithm for general f-divergences below.

Pool Adjacent Violators (PAV) Algorithm First, recall the optimization problem we wish to solve:

$$\min_{\substack{\boldsymbol{z}\in\mathbb{R}^n \\ \boldsymbol{z}_1 \le \ldots \le \boldsymbol{z}_n}} \sum_{i=1}^{n} g_i(z_i; \boldsymbol{l}), \quad \text{where} \quad g_i(z_i; \boldsymbol{l}) := \sigma_i z_i + \frac{\nu}{n} f^*\left(\frac{l_{\pi(i)} - z_i}{\nu}\right). \tag{A.22}$$

The objective can be thought of as fitting a real-valued monotonic function to the points $(1, l_{\pi(1)}), \ldots, (n, l_{\pi(n)})$, which would require specifying its values $(c_1, \ldots, c_n)$ on $(1, \ldots, n)$ and defining the function as any $x \in [c_j, c_{j+1}]$ on $(j, j+1)$. Because $l_{\pi(1)} \le \ldots \le l_{\pi(n)}$, if we evaluated our function $(c_1, \ldots, c_n)$ on a loss such as $\sum_{i=1}^{n}(l_{\pi(i)} - z_i)^2$, we may easily solve the problem by returning $c_1 = l_{\pi(1)}, \ldots, c_n = l_{\pi(n)}$. However, by specifying functions $g_1, \ldots, g_n$ we allow our loss function to change in different regions of the input space $\{1, \ldots, n\}$. In such cases, the monotonicity constraint $c_1 \le \ldots \le c_n$ is often violated because individually minimizing $g_i(z_i)$ for each z_i has no guarantee of yielding a function that is monotonic.

The idea behind the PAV algorithm is to attempt a pass at minimizing each g_i individually, and correcting *violations* as they appear. To provide intuition, define $c_i^* \in \arg\min_{z_i \in \mathbb{R}} g_i(z_i)$, and consider $i < j$ such that $c_i^* > c_j^*$. If f^* is strictly convex, then $g_i(x) > g_i(c_i^*)$ for any $x < c_i^*$ and similarly $g_j(x) > g_j(c_j^*)$ for any $x > c_j^*$. Thus, to correct

the violation, we decrease c_i^* to $\bar{c}_i$ and increase c_j^* to $\bar{c}_j$ until $\bar{c}_i = \bar{c}_j$. We determine this midpoint precisely by

$$\bar{c}_i = \bar{c}_j = \operatorname*{arg\,min}_{x \in \mathbb{R}} g_i(x) + g_j(x)$$

as these are exactly the contributions made by these terms in the overall objective. The computation above is called *pooling* the indices i and j. We may generalize this viewpoint to *violating chains*, that is collections of contiguous indices $(i, i+1, \ldots, i+m)$ such that $c_j^* < c_i^*$ for all $j < i$ and $c_j^* > c_{i+m}^*$ for all $j > i+m$, but $c_i^* > c_{i+m}^*$. One approach is to use dynamic programming to identify such chains and then compute the pooled quantities

$$\bar{c}_i = \operatorname*{arg\,min}_{x \in \mathbb{R}} \sum_{j=1}^{m} g_{i+j}(x).$$

This requires two passes through the vector: one for identifying violators and the other for pooling. The Pool Adjacent Violators algorithm, on the other hand, is able to perform both operations in one pass by greedily pooling violators as they appear. This can be viewed as a meta-algorithm, as it hinges on the notion that the solution of "larger" pooling problems can be easily computed from solutions of "smaller" pooling problems. Precisely, for indices $S \subseteq [n] = \{1, \ldots, n\}$ define

$$\mathrm{Sol}(S) = \operatorname*{arg\,min}_{x \in \mathbb{R}} \sum_{i \in S} g_i(x).$$

We rely on the existence of an operation Pool, such that for any $S, T \subseteq [n]$ such that $S \cap T = \emptyset$, we have that

$$\mathrm{Sol}(S \cup T) = \mathrm{Pool}\left(\mathrm{Sol}(S), \mathrm{Met}(S), \mathrm{Sol}(T), \mathrm{Met}(T)\right), \tag{A.23}$$

where $\mathrm{Met}(S)$ denotes "metadata" associated to S, and that the number of elementary operations in the Pool function is $O(1)$ with respect to $|S| + |T|$. We review our running examples.

For the χ^2-divergence, we have that $f_{\chi^2}(x) = x^2 - 1$ and $f^*_{\chi^2}(y) = y^2/4 + 1$, so

$$\begin{aligned}
\operatorname{Sol}(S) &= \operatorname*{arg\,min}_{x\in\mathbb{R}} \left\{ x \left(\sum_{i\in S} \sigma_i \right) + |S| + \frac{1}{4n\nu} \sum_{i\in S} (l_{\pi(i)} - x)^2 \right\} \\
&= \frac{1}{|S|} \left[\sum_{i\in S} l_{\pi(i)} - 2n\nu \sum_{i\in S} \sigma_i \right] \\
\operatorname{Sol}(S\cup T) &= \frac{1}{|S|+|T|} \left[\sum_{i\in S\cup T} l_{\pi(i)} - 2n\nu \sum_{i\in S\cup T} \sigma_i \right] \\
&= \frac{|S|\operatorname{Sol}(S) + |T|\operatorname{Sol}(T)}{|S|+|T|}.
\end{aligned}$$

Thus, the metadata $\operatorname{Met}(S) = |S|$ used in the pooling step (A.23) is the size of each subset.

For the KL divergence, $f_{\mathrm{KL}}(x) = x\ln x$ and $f^*_{\mathrm{KL}}(y) = e^{-1}\exp(y)$, so so

$$\begin{aligned}
\operatorname{Sol}(S) &= \operatorname*{arg\,min}_{x\in\mathbb{R}} \left\{ x \left(\sum_{i\in S} \sigma_i \right) + \frac{\nu}{ne} \sum_{i\in S} \exp\left(l_{\pi(i)}/\nu\right) \exp\left(-x/\nu\right) \right\} \\
&= \nu \left[\ln \sum_{i\in S} \exp\left(l_{\pi(i)}/\nu\right) - \ln \sum_{i\in S} \sigma_i - \ln n - 1 \right] \\
\operatorname{Sol}(S\cup T) &= \nu \left[\ln \sum_{i\in S\cup T} \exp\left(l_{\pi(i)}/\nu\right) - \ln \sum_{i\in S\cup T} \sigma_i - \ln n - 1 \right] \\
&= \nu \left[\ln \left(\sum_{i\in S} \exp\left(l_{\pi(i)}/\nu\right) + \sum_{i\in T} \exp\left(l_{\pi(i)}/\nu\right) \right) - \ln \left(\sum_{i\in S} \sigma_i + \sum_{i\in T} \sigma_i \right) - \ln n - 1 \right].
\end{aligned}$$

Here, we carry the metadata $\operatorname{Met}(S) = \left(\ln \sum_{i\in S} \exp\left(l_{\pi(i)}/\nu\right), \ln \sum_{i\in S} \sigma_i\right)$, which can easily be combined and plugged into the function

$$(m_1, m_2), (m_1', m_2') \mapsto \nu \left[\ln\left(\exp m_1 + \exp m_1'\right) - \ln\left(\exp m_2 + \exp m_2'\right) - \ln n - 1 \right]. \quad \text{(A.24)}$$

for two instances of metadata (m_1, m_2) and (m_1', m_2'). We carry the "logsumexp" instead of just the sum of exponential quantities for numerical stability, and Equation (A.24) applies this operation as well. It might be that $\sum_{i\in S} \sigma_i = 0$, e.g. for the superquantile. In this case, we may interpret $\operatorname{Sol}(S) = -\infty$ and evaluate $\exp(-\infty) = 0$ in the conversion formula (A.22). Two examples of the PAV algorithm are given in Subroutine 1 and Subroutine 2,

Algorithm 6 Prospect (with exact implementation details)

Inputs: Initial points $\boldsymbol{\theta}_0(= \mathbf{0})$, spectrum $\sigma(=$ 2-extremile), stepsize $\eta(= 0.01)$, number of iterations $t(= 1000)$, tolerance $\varepsilon(= 10^{-4})$, regularization parameter $\mu(= 1/n)$, shift cost $\nu(= 0.05)$, loss/gradient oracles $\ell_1, \ldots, \ell_n$ and $\nabla\ell_1, \ldots, \nabla\ell_n$.

$\boldsymbol{l} \leftarrow \ell(\boldsymbol{\theta}_0) \in \mathbb{R}^n$.
$\boldsymbol{g} \leftarrow (\nabla\ell_i(\boldsymbol{\theta}_0) + \mu\boldsymbol{\theta}_0)_{i=1}^n \in \mathbb{R}^{n\times d}$.
$\pi \leftarrow \text{argsort}(\boldsymbol{l})$.
$\boldsymbol{c} \leftarrow \text{PAV}(\boldsymbol{l}, \pi, \sigma)$. ▷ Subroutine 1 or Subroutine 2
$\boldsymbol{q} \leftarrow \text{Convert}(\boldsymbol{c}, \boldsymbol{l}, \pi, \nu, \mathbf{0}_n)$. ▷ Subroutine 3
$\boldsymbol{\rho} \leftarrow \boldsymbol{q}$.
$\bar{\boldsymbol{g}} \leftarrow \sum_{i=1}^n \rho_i \boldsymbol{g}_i \in \mathbb{R}^d$.
for $k = 1, \ldots, t$ **do**
    If n divides k, then check if certificate $\leq \varepsilon$ (Section 3.7.1). If so, terminate.
    Sample $i, j \sim \text{Unif}[n]$.
    $\boldsymbol{v} \leftarrow nq_i(\nabla\ell_i(\boldsymbol{\theta}) + \mu\boldsymbol{\theta}) - n\rho_i\boldsymbol{g}_i - \bar{\boldsymbol{g}}$. ▷ **Iterate Update**
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\boldsymbol{v}$.
    $l_j \leftarrow \ell_j(\boldsymbol{\theta})$. ▷ **Bias Reducing Update**
    $\pi \leftarrow \text{BubbleArgSort}(\pi, \boldsymbol{l})$. ▷ Subroutine 4
    $\boldsymbol{c} \leftarrow \text{PAV}(\boldsymbol{l}, \pi, \sigma)$.
    $\boldsymbol{q} \leftarrow \text{Convert}(\boldsymbol{c}, \boldsymbol{l}, \pi, \nu, \boldsymbol{q})$.
    $\bar{\boldsymbol{g}} \leftarrow \bar{\boldsymbol{g}} - \rho_i\boldsymbol{g}_i + q_i(\nabla\ell_i(\boldsymbol{\theta}) + \mu\boldsymbol{\theta})$. ▷ **Variance Reducing Update**
    $\boldsymbol{g}_i \leftarrow \nabla\ell_i(\boldsymbol{\theta}) + \mu\boldsymbol{\theta}$.
    $\rho_i \leftarrow q_i$.

Output: Final point $\boldsymbol{\theta}$.

respectively. These operate by selecting the unique values of the optimizer and partitions of indices that achieve that value.

Hardware Acceleration Finally, note that all of the subroutines in Algorithm 6 (Subroutine 1/Subroutine 2, Subroutine 3, and Subroutine 4) all require primitive operations such as control flow and linear scans through vectors. Because these steps are outside of the purview of oracle calls or matrix multiplications, they benefit from just-in-time compilation on the CPU. We accelerate these subroutines using the Numba package in Python and are able to achieve an approximate 50%-60% decrease in runtime across benchmarks.

Subroutine 1 Pool Adjacent Violators (PAV) Algorithm for χ^2 divergence

Inputs: Losses $(\ell_i)_{i\in[n]}$, argsort π, and spectrum $(\sigma_i)_{i\in[n]}$.

Initialize partition endpoints $(b_0, b_1) = (0, 1)$, partition value $v_1 = l_{\pi(1)} - 2n\nu\sigma_1$, number of parts $K = 1$.
for $i = 2, \dots, n$ **do**
  Add part $K = K + 1$.
  Compute $v_K = l_{\pi(i)} - 2n\nu\sigma_i$.
  while $K \geq 2$ and $v_{K-1} \geq v_K$ **do**
    $v_{K-1} = \frac{(b_K - b_{K-1})v_{K-1} + (i - b_K)v_K}{i - b_{K-1}}$.
    Set $K = K - 1$.
  $b_K = i$.

Output: Vector c containing $z_i = v_K$ for $b_{K-1} < i \leq b_K$.

Subroutine 2 Pool Adjacent Violators (PAV) Algorithm for KL divergence

Inputs: Losses $(\ell_i)_{i\in[n]}$, argsort π, and spectrum $(\sigma_i)_{i\in[n]}$.

Initialize partition endpoints $(b_0, b_1) = (0, 1)$, number of parts $K = 1$.
Initialize partition value $v_1 = \nu\left(l_{\pi(1)}/\nu - \ln\sigma_1 - \ln n - 1\right)$.
Initialize metadata $m_1 = \ell_{\pi(1)}/\nu$ and $t_1 = \ln\sigma_1$.
for $i = 2, \dots, n$ **do**
  Add part $K = K + 1$.
  Compute $v_K = \nu\left(l_{\pi(i)}/\nu - \ln\sigma_i - \ln n - 1\right)$.
  Compute $m_K = \ell_{\pi(i)}/\nu$ and $t_K = \ln\sigma_i$
  while $k \geq 2$ and $v_{K-1} \geq v_K$ **do**
    $m_{K-1} = \mathrm{logsumexp}(m_{K-1}, m_K)$ and $t_{K-1} = \mathrm{logsumexp}(t_{K-1}, t_K)$.
    $v_{K-1} = \nu\,(m_{K-1} - t_{K-1} - \ln n - 1)$.
    Set $K = K - 1$.
  $b_K = i$.

Output: Vector c containing $z_i = v_K$ for $b_{K-1} < i \leq b_K$.

Subroutine 3 Convert

Require: Sorted vector $c \in \mathbb{R}$, vector $\boldsymbol{l} \in \mathbb{R}^n$, argsort π of $\boldsymbol{l}$, shift cost $\nu \geq 0$, vector $\boldsymbol{q} \in \mathbb{R}^n$.

for $i = 1, \dots, n$ **do**
  Set $q_{\pi(i)} = (1/n)[f^*]'\left((l_{\pi(i)} - z_i)/\nu\right)$.
return $\boldsymbol{q}$.

Subroutine 4 BubbleArgSort

Require: Index j_{init}, sorting permutation π, loss table $\boldsymbol{l}$.

$j = j_{\text{init}}$. ▷ If $l_{\pi(j_{\text{init}})}$ too small, bubble left.
while $j > 1$ and $l_{\pi(j)} < l_{\pi(j-1)}$ **do**
    Swap $\pi(j)$ and $\pi(j-1)$.
$j = j_{\text{init}}$. ▷ If $l_{\pi(j_{\text{init}})}$ too large, bubble right.
while $j < n$ and $l_{\pi(j)} > l_{\pi(j+1)}$ **do**
    Swap $\pi(j)$ and $\pi(j+1)$.
return π

Dataset	d	n_{train}	n_{test}	Task	Source
`yacht`	6	244	62	Regression	UCI
`energy`	8	614	154	Regression	UCI
`concrete`	8	824	206	Regression	UCI
`kin8nm`	8	6,553	1,639	Regression	OpenML
`power`	4	7,654	1,914	Regression	UCI
`diabetes`	33	4,000	1,000	Binary Classification	Fairlearn
`acsincome`	202	4,000	1,000	Regression	Fairlearn
`amazon`	535	10,000	10,000	Multiclass Classification	WILDS
`iwildcam`	9420	20,000	5,000	Multiclass Classification	WILDS

Table A.1: Dataset attributes and dimensionality d, train sample size n_{train}, and test sample size n_{test}.

A.4 Experimental Details

A.4.1 Tasks & Objectives

In all settings, we consider supervised learning tasks specified by losses of the form

$$\ell_i(\boldsymbol{\theta}) = h(y_i, \langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}_i) \rangle),$$

where we consider an input $\boldsymbol{x}_i \in \mathcal{X}$, a feature map $\varphi : \mathcal{X} \to \mathbb{R}^d$, and a label $y_i \in \mathcal{Y}$. The function $h : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ measures the error between the true label and another value which is the prediction in regression and the logit probabilities of the associated classes in classification. In the regression tasks, $\mathcal{Y} = \mathbb{R}$ and we used the squared loss

$$\ell_i(\boldsymbol{\theta}) = \frac{1}{2}(y_i - \langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}_i) \rangle)^2 .$$

For binary classification, we have $\mathcal{Y} = \{-1, 1\}$, denoting a negative and positive class. We used the binary logistic loss

$$\ell_i(\boldsymbol{\theta}) = -y_i \langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}_i) \rangle + \ln(1 + e^{\langle \boldsymbol{\theta}, \varphi(\boldsymbol{x}_i) \rangle}) .$$

For multiclass classification, $\mathcal{Y} = \{1, \dots, C\}$ where C is the number of classes. We used the multinomial logistic loss:

$$\ell_i(\boldsymbol{\theta}) = -\ln p_{y_i}(\boldsymbol{x}_i; \boldsymbol{\theta}), \text{ where } p_{y_i}(\boldsymbol{x}_i; \boldsymbol{\theta}) := \frac{\exp\left(\langle \boldsymbol{\theta}_{\cdot y}, \varphi(\boldsymbol{x}_i) \rangle\right)}{\sum_{y'=1}^{C} \exp\left(\langle \boldsymbol{\theta}_{\cdot y'}, \varphi(\boldsymbol{x}_i) \rangle\right)}, \ \boldsymbol{\theta} \in \mathbb{R}^{d \times C}$$

The design matrix $(\varphi(\boldsymbol{x}_1), \dots, \varphi(\boldsymbol{x}_n)) \in \mathbb{R}^{n \times d}$ is standardized to have columns with zero mean and unit variance, and the estimated mean and variance from the training set is used to standardize the test sets as well. Our final objectives are of the form

$$\mathcal{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{q} \in \mathcal{Q}(\sigma)} \sum_{i=1}^{n} q_i \ell_i(\boldsymbol{\theta}) - \nu n \left\| \boldsymbol{q} - \mathbf{1}/n \right\|_2^2 + \frac{\mu}{2} \left\| \boldsymbol{\theta} \right\|_2^2$$

for shift cost $\nu \geq 0$ and regularization constant $\mu \geq 0$.

A.4.2 Datasets

We detail the datasets used in the experiments. If not specified below, the input space $\mathcal{X} = \mathbb{R}^d$ and φ is the identity map. The sample sizes, dimensions, and source of the datasets are summarized in Table A.1, where d refers to the dimension of each $\varphi(\boldsymbol{x}_i)$.

(a) `yacht`: prediction of the residuary resistance of a sailing yacht based on its physical attributes [Tsanas and Xifara, 2012].

(b) `energy`: prediction of the cooling load of a building based on its physical attributes Baressi Segota et al. [2020].

(c) `concrete`: prediction of the compressive strength of a concrete type based on its physical and chemical attributes [Yeh, 2006].

(d) `kin8nm`: prediction of the distance of an 8-link all-revolute robot arm to a spatial end-point [Akujuobi and Zhang, 2017].

(e) `power`: prediction of net hourly electrical energy output of a power plant given environmental factors [Tüfekci, 2014].

(f) `diabetes`: prediction of readmission for diabetes patients based on 10 years' worth of clinical care data at 130 US hospitals [Rizvi et al., 2014].

(g) `acsincome`: prediction of income of US adults given features compiled from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) [Ding et al., 2021].

(h) `amazon`: prediction of the review score of a sentence taken from Amazon products. Each input $x \in \mathcal{X}$ is a sentence in natural language and the feature map $\varphi(\boldsymbol{x}) \in \mathbb{R}^d$ is generated by the following steps:

- A BERT neural network [Devlin et al., 2019a] (fine-tuned on $10,000$ held-out examples) is applied to the text $\boldsymbol{x}_i$, resulting in vector x_i'.
- The $x_1', \ldots, x_n'$ are normalized to have unit norm.
- Principle Components Analysis (PCA) is applied, resulting in 105 components that explain 99% of the variance, resulting in vectors $x_i'' \in \mathbb{R}^{105}$. The d in Table A.1 refers to the total dimension of the parameter vectors for all 5 classes.

(i) `iwildcam`: prediction of an animal or flora in an image from wilderness camera traps, with heterogeneity in illumination, camera angle, background, vegetation, color, and relative animal frequencies [Beery et al., 2020]. Each input $x \in \mathcal{X}$ is an image the feature map $\varphi(\boldsymbol{x}) \in \mathbb{R}^d$ is generated by the following steps:

- A ResNet50 neural network [He et al., 2016] that is pretrained on ImageNet [Deng et al., 2009] is applied to the image $\boldsymbol{x}_i$, resulting in vector $\boldsymbol{x}_i'$.
- The $\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n'$ are normalized to have unit norm.
- Principle Components Analysis (PCA) is applied, resulting in $d = 157$ components that explain 99% of the variance. The d in Table A.1 refers to the total dimension of the parameter vectors for all 60 classes.

A.4.3 Hyperparameter Selection

We fix a minibatch size of 64 SGD and SRDA and an epoch length of $N = n$ for LSVRG. For SaddleSAGA we consider three schemes for selecting the primal and dual learning rates that reduce the search to a single parameter $\eta > 0$ by tuning a scaling of the primal and dual learning rates that performs well across experiments. In practice, the regularization parameter μ and shift cost ν are tuned by a statistical metric, i.e., generalization error as measured on a validation set. We study the optimization performance of the methods for multiple values of each in Appendix A.4.5.

For the tuned hyperparameters, we use the following method. Let $k \in \{1, \ldots, K\}$ be a seed that determines algorithmic randomness. This corresponds to sampling a minibatch without replacement for SGD and SRDA and a single sampled index for SaddleSAGA, LSVRG, and Prospect. Letting $\mathcal{L}_k(\eta)$ denote the average value of the training loss of the last ten passes using learning rate η and seed k, the quantity $\mathcal{L}(\eta) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_k(\eta)$ was minimized to select η. The learning rate η is chosen in the set $\{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}, 3 \times 10^{-2}, 1 \times 10^{-1}, 3 \times 10^{-1}, 1 \times 10^{0}, 3 \times 10^{0}\}$, with two orders of magnitude lower numbers used in `acsincome` due to its sparsity. We discard any learning rates that cause the optimizer to diverge for any seed.

A.4.4 Compute Environment

No GPUs were used in the study; Experiments were run on a CPU workstation with an Intel i9 processor, a clock speed of 2.80GHz, 32 virtual cores, and 126G of memory. The code used in this project was written in Python 3 using the PyTorch and Numba packages for automatic differentiation and just-in-time compilation, respectively.

A.4.5 Additional Experiments

Varying Risk Parameters We study the effect of varying the risk parameters, that is (τ, r, γ) for the τ-CVaR (Equation (2.13)), r-extremile (Equation (2.14)), γ-ESRM (Equa-

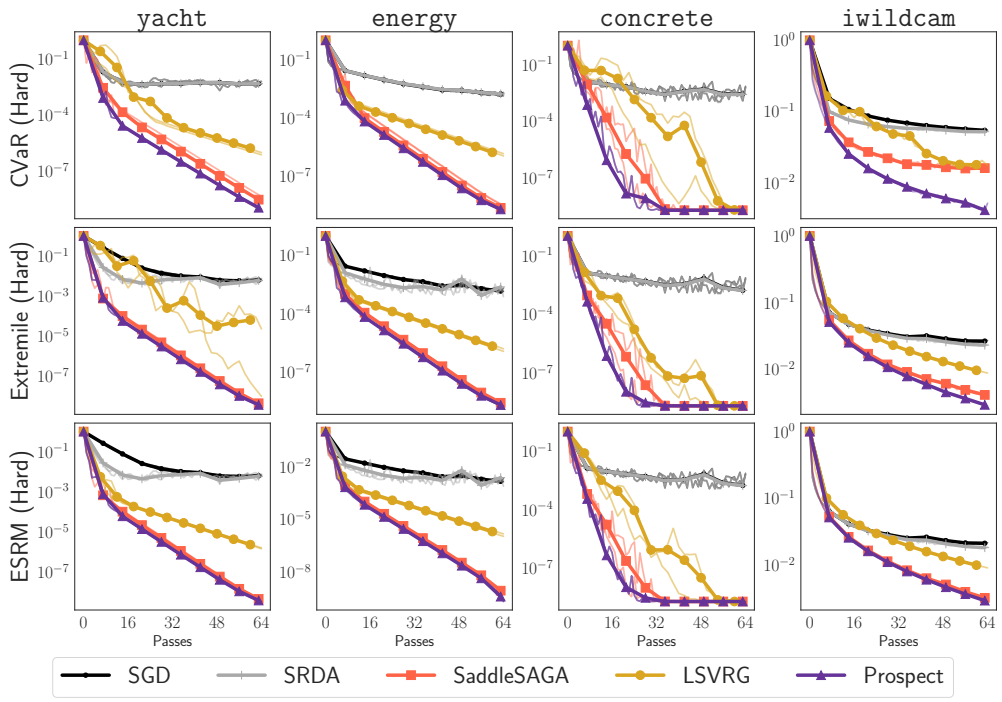

Figure A.1: **Harder Risk Parameter Settings.** Each row represents a different "hard" variant of the superquantile, extremile, and ESRM spectra. Columns represent different datasets. Suboptimality (2.49) is measured on the y-axis while the x-axis measures the total number of gradient evaluations made divided by n, i.e., the number of passes through the training set.

tion (2.15)), choosing the spectrum to increase the condition number $\kappa_\sigma = n\sigma_n$ compared to the experiments in the main text. We use $\tau = 0.75$, $r = 2.5$, and $\gamma = 1/e^{-2}$ to generate "hard" version of the superquantile, extremile, and ESRM. Figure A.1 plots the corresponding training curves for four datasets of varying sample sizes: `yacht`, `energy`, `concrete`, and `iwildcam`. We see that the comparison of methods is the same as the original methods, that is that Prospect performs the best or close to best in terms of optimization trajectories. Except on `concrete`, SaddleSAGA generally matches the performance of Prospect. The trajectory of LSVRG is noticeably noisier than on the original settings; we hypothesize that the bias accrued by this epoch-based algorithm is exacerbated by the skewness in the spectrum, as mentioned in Mehta et al. [2023, Proposition 1].

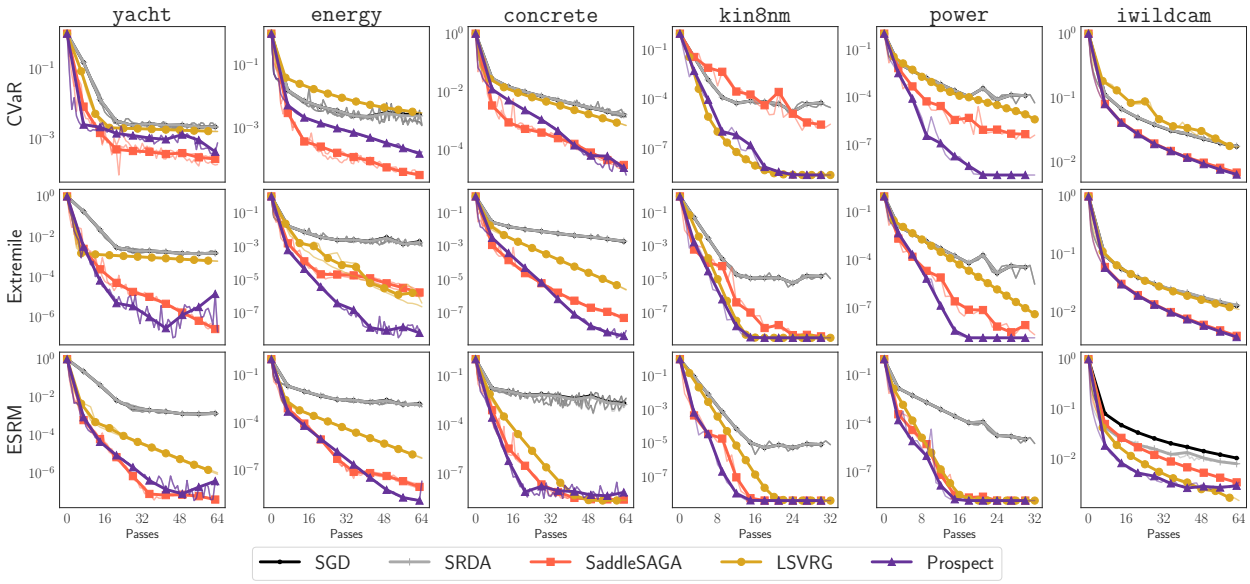

Figure A.2: **No Shift Cost Settings.** Each row represents a different spectral risk objective with $\nu = 0$ (instead of $\nu = 1$) while each column represents a different datasets. Suboptimality (2.49) is measured on the y-axis while the x-axis measures the total number of gradient evaluations made divided by n, i.e., the number of passes through the training set.

Removing Shift Cost A relevant setting is the low or no shift cost regime ($\nu = 0$), as this allows the adversary to make arbitrary distribution shifts (while still constrained to $\mathcal{Q}(\sigma)$). Figure A.2 displays these curves for this no-cost experiment. When $\nu = 0$, the optimization problem can equivalently be written as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left[\max_{\boldsymbol{q} \in \mathcal{Q}(\sigma)} \langle \boldsymbol{q}, \ell(\boldsymbol{\theta}) \rangle + \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2 = \sum_{i=1}^{n} \sigma_i \ell_{(i)}(\boldsymbol{\theta}) + \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2 \right].$$

In this case, we always have that $\boldsymbol{q}(\boldsymbol{l}) = (\sigma_{\pi^{-1}(1)}, \ldots, \sigma_{\pi^{-1}(n)})$, where π sorts $\boldsymbol{l}$. Here, $\boldsymbol{\theta}$ is chosen to optimize a linear combination of order statistics of the losses. In the low shift cost settings, performance trends are qualitatively similar to those seen from $\nu = 1$. Interestingly, for the no-cost setting, SaddleSAGA, LSVRG, and Prospect seem to converge linearly empirically even without smoothness of the objective.

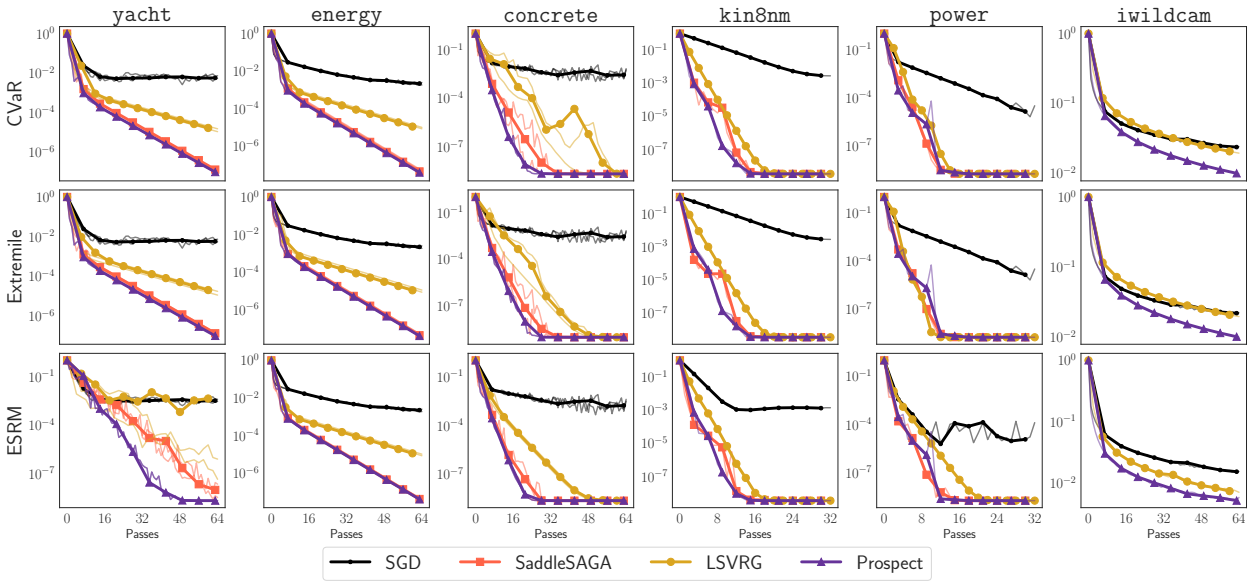

Figure A.3: **Reduced ℓ_2-regularization settings ($\mu = 1/(10n)$.** Each row represents a different spectral risk objective with $\mu = 1/(10n)$ (instead of $\mu = 1/n$) while each column represents a different dataset. Suboptimality (2.49) is measured on the y-axis while the x-axis measures the total number of gradient evaluations made divided by n, i.e., the number of passes through the training set.

Lowering Regularization Next, we decrease the ℓ_2-regularization from $\mu = 1/n$ to $\mu = 1/(10n)$. These settings are plotted in Figure A.3. Performance rankings among methods reflect those of the original parameters. For five of the six datasets, that is `yacht`, `energy`, `concrete`, `kin8nm`, and `power`, the regression tasks involve optimizing the squared error. This function is already strongly convex, with a constant depending on the smallest eigenvalue of the empirical second-moment matrix. When assuming that the input data vectors are bounded, this function is also G-Lipschitz. Thus, if the problem is already well-conditioned, we may observe similar behavior even at negligible regularization ($\mu = 5 \cdot 10^{-7}$ for `iwildcam`, for example).

Appendix B

APPENDIX TO CHAPTER 4

B.1 Linear Operators and Variance Reduction

This section is dedicated to establishing the variance reduction result in Corollary 4.3.1 by employing properties of the conditional mean operators introduced in Section 4.3. In the first part, we establish Proposition B.1.1, the singular value decomposition that defines the quantities appearing in Corollary 4.3.1. In the second part, we quantify the difference between σ_0^2 and σ_k^2 for even and odd iterations of k.

B.1.1 Singular Value Decomposition

Recall the conditional mean operators μ_X and μ_Z from Section 4.3,

$$[\mu_X h](\boldsymbol{x}) := \mathbb{E}\left[h(X,Z)|X\right](\boldsymbol{x}) \text{ and } [\mu_Z h](\boldsymbol{z}) := \mathbb{E}\left[h(X,Z)|Z\right](\boldsymbol{z}),$$

with the corresponding debiasing (a.k.a. centering) operators defined by $\mathcal{C}_X = I - \mu_X$ and $\mathcal{C}_Z = I - \mu_Z$.

Proposition B.1.1. *There exists a basis* $\{\alpha_j\}_{j=1}^m$ *of* $\mathbf{L}^2(P_X)$*, a basis* $\{\beta_j\}_{j=1}^m$ *of* $\mathbf{L}^2(P_Z)$*, and real values* $\{s_j\}_{j=1}^m$*, which satisfy:*

$$\mu_Z \alpha_j = s_j \beta_j \text{ and } \mu_X \beta_j = s_j \alpha_j \text{ for } j \in \{1,\ldots,m\}, \tag{B.1}$$

$\alpha_1 = \mathbf{1}_{\mathcal{X}}$, $\beta_1 = \mathbf{1}_{\mathcal{Z}}$, $s_1 = 1$ *and* s_j *is non-negative and non-increasing in* j*.*

Proof. When μ_X is restricted to $\mathbf{L}^2(P_Z)$ and μ_Z is restricted to $\mathbf{L}^2(P_X)$, these operators are

in fact adjoint in $\mathbf{L}^2(P)$, as by the tower property we have the relation

$$\langle f, \mu_X g \rangle_{\mathbf{L}^2(P_X)} = \mathbb{E}\left[f(X)\mathbb{E}\left[g(Z)|X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[f(X)|Z\right]g(Z)\right] = \langle \mu_Z f, g \rangle_{\mathbf{L}^2(P_Z)}.$$

Since $\mu_Z : \mathbf{L}^2(P_X) \to \mathbf{L}^2(P_Z)$ is a compact linear operator, by Gohberg et al. [1990, Section IV.1 Theorem 1.1] and Gohberg et al. [1990, Section IV.1 Corollary 1.2], we have that μ_Z admits a singular value decomposition satisfying (B.1). Next, we show that $s_1 \leq 1$ and that $\mathbf{1}_{\mathcal{X}}$ is an eigenvector of $\mu_X \mu_Z : \mathbf{L}^2(P_X) \to \mathbf{L}^2(P_X)$ with eigenvalue 1, which confirms that $s_1 = 1$ and $\alpha_1 = \mathbf{1}_{\mathcal{X}}$ by the definition of singular values (arguing symmetrically achieves $\beta_1 = \mathbf{1}_{\mathcal{Z}}$). By the variational representation of singular values [Gohberg et al., 1990, Section IV.1 Equation (2)], we have that

$$\sup_{f:\|f\|_{\mathbf{L}^2(P_X)}=1} \|\mu_Z f\|_{\mathbf{L}^2(P_Z)} = s_1.$$

Consider any $f \in \mathbf{L}^2(P_X)$ such that $\|f\|_{\mathbf{L}^2(P_X)} = 1$. Define the conditional probability $P_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) = P(\boldsymbol{x}, \boldsymbol{z})/P_Z(\boldsymbol{z})$ which is well-defined by assumption. Then, by the Cauchy-Schwarz inequality in $\mathbf{L}^2(P_{X|Z})$,

$$\begin{aligned}
\|\mu_Z f\|^2_{\mathbf{L}^2(P_Z)} &= \sum_{\boldsymbol{z}\in\mathcal{Z}} \left(\sum_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) P_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) \right)^2 P_Z(\boldsymbol{z}) \\
&\leq \sum_{\boldsymbol{z}\in\mathcal{Z}} \sum_{\boldsymbol{x}\in\mathcal{X}} f^2(\boldsymbol{x}) P_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) P_Z(\boldsymbol{z}) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} f^2(\boldsymbol{x}) \sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{x}, \boldsymbol{z}) \\
&= \|f\|^2_{\mathbf{L}^2(P_X)} = 1.
\end{aligned}$$

This proves that $s_1 \leq 1$. For equality, notice that $\mu_X \mu_Z \mathbf{1}_{\mathcal{X}} = \mu_X \mathbf{1}_{\mathcal{Z}} = \mathbf{1}_{\mathcal{X}}$, completing the proof. □

B.1.2 Proof of Main Results

From Proposition B.1.1, we establish two bases $\{\alpha_j\}_{j=1}^m$ and $\{\beta_j\}_{j=1}^m$ of $\mathbf{L}^2(P_X)$ and $\mathbf{L}^2(P_Z)$, respectively. These bases span the range of the operators μ_X and μ_Z. We will consider the repeated application of the operator $\mathcal{C}_Z\mathcal{C}_X$, a sequence of two centering operations on some function $h \in \mathbf{L}^2(P)$, and compare

$$\mathbb{E}\left[((\mathcal{C}_Z\mathcal{C}_X)^t\bar{h})^2\right] \text{ against } \mathbb{E}\left[\bar{h}^2\right]$$

for $\bar{h} = h - \mathbb{E}_P[h]$. We establish the main result by measuring the reduction in variance from a single application, in terms of the coordinates of the function of interest on each of the two subspaces. We will then observe how these coordinates change iteration-to-iteration to give the final result.

Lemma B.1.1. *For any $h \in \mathbf{L}^2(P)$ such that $\mathbb{E}_P[h] = 0$, let*

$$\mu_X h = \sum_{j=1}^m u_j\alpha_j \text{ and } \mu_Z h = \sum_{j=1}^m v_j\beta_j.$$

Then, we have that

$$\mathbb{E}\left[(\mathcal{C}_Z\mathcal{C}_X h)^2\right] = \mathbb{E}\left[h^2\right] - \sum_{j=2}^m u_j^2 - \sum_{j=2}^m (v_j - s_j u_j)^2.$$

Proof. By orthogonality, we have that

$$\begin{aligned}
\mathbb{E}\left[(\mathcal{C}_Z\mathcal{C}_X h)^2\right] &= \mathbb{E}\left[((I-\mu_Z)\mathcal{C}_X h)^2\right] \\
&= \mathbb{E}\left[(\mathcal{C}_X h)^2\right] - 2\mathbb{E}\left[(\mathcal{C}_X h)(\mu_Z\mathcal{C}_X h)\right] + \mathbb{E}\left[(\mu_Z\mathcal{C}_X h)^2\right] \\
&= \mathbb{E}\left[(\mathcal{C}_X h)^2\right] - 2P_Z((\mu_Z\mathcal{C}_X h)^2) + P_Z((\mu_Z\mathcal{C}_X h)^2) \\
&= \mathbb{E}\left[(\mathcal{C}_X h)^2\right] - P_Z((\mu_Z\mathcal{C}_X h)^2) \\
&= \mathbb{E}\left[h^2\right] - P_X((\mu_X h)^2) - P_Z((\mu_Z\mathcal{C}_X h)^2).
\end{aligned}$$

Because $P(h) = 0$, it holds by the tower property of conditional expectation that $P_X(\mu_X h) =$

0, which implies that

$$u_1 = \langle \mu_X h, \alpha_1 \rangle_{\mathbf{L}^2(P_X)} = 0 \implies P_X((\mu_X h)^2) = \sum_{j=2}^{m} u_j^2.$$

For the second term, observe that $P_X(\mathcal{C}_X h) = 0$, so it holds by the tower property that $P_Z(\mu_Z \mathcal{C}_X h) = 0$, so

$$P_Z((\mu_Z \mathcal{C}_X h)^2) = \sum_{j=2}^{m} \left(\langle \mu_Z \mathcal{C}_X h, \beta_j \rangle_{\mathbf{L}^2(P_Z)} \right)^2.$$

Next, we compute the term in the square by applying Proposition B.1.1:

$$\begin{aligned}
\langle \mu_Z \mathcal{C}_X h, \beta_j \rangle_{\mathbf{L}^2(P_Z)} &= \langle \mu_Z h, \beta_j \rangle_{\mathbf{L}^2(P_Z)} - \langle \mu_Z \mu_X h, \beta_j \rangle_{\mathbf{L}^2(P_Z)} \\
&= v_j - \left\langle \mu_Z \sum_{k=1}^{m} u_k \alpha_k, \beta_j \right\rangle_{\mathbf{L}^2(P_Z)} \\
&= v_j - \left\langle \sum_{k=1}^{m} u_k s_k \beta_k, \beta_j \right\rangle_{\mathbf{L}^2(P_Z)} \\
&= v_j - s_j u_j,
\end{aligned}$$

which completes the proof. □

Lemma B.1.1 ensures that we have a reduction on each iteration, with a formula that depends on the coordinates of the function on each subspace. Because these coordinates change every iteration, we track them in the next lemma. Define $h_0 = \bar{h}$ and $h_{t+1} = (\mathcal{C}_Z \mathcal{C}_X) h_t$, along with the constants $\{u_{t,j}\}_{j=1}^m$ and $\{v_{t,j}\}_{j=1}^m$ given by

$$\mu_X h_t = \sum_{j=1}^{m} u_{t,j} \alpha_j \text{ and } \mu_Z h_t = \sum_{j=1}^{m} v_{t,j} \beta_j.$$

We have the following.

Lemma B.1.2. *For all $t \geq 0$, it holds that*

$$u_{t+1,j} = s_j^2 u_{t,j} - s_j v_{t,j},$$
$$v_{t+1,j} = 0.$$

Proof. Fix any $j \in [m]$, and use Proposition B.1.1 to write

$$\begin{aligned}
u_{t+1,j} &= \langle \mu_X \mathcal{C}_Z \mathcal{C}_X h_t, \alpha_j \rangle_{\mathbf{L}^2(P_X)} \\
&= \langle \mu_X (I - \mu_X - \mu_Z + \mu_Z \mu_X) h_t, \alpha_j \rangle_{\mathbf{L}^2(P_X)} \\
&= \langle \mu_X \mu_Z \mu_X h_t, \alpha_j \rangle_{\mathbf{L}^2(P_X)} - \langle \mu_X \mu_Z h_t, \alpha_j \rangle_{\mathbf{L}^2(P_X)} \\
&= \left\langle \mu_X \mu_Z \sum_{k=1}^m u_{t,k} \alpha_k, \alpha_j \right\rangle_{\mathbf{L}^2(P_X)} - \left\langle \mu_X \sum_{k=1}^m v_{t,k} \beta_k, \alpha_j \right\rangle_{\mathbf{L}^2(P_X)} \\
&= s_j^2 u_{t,j} - s_j v_{t,j},
\end{aligned}$$

which proves the first part of the claim. For the second part, note that $\mu_Z \mathcal{C}_Z = 0$, so $\langle \mu_Z \mathcal{C}_Z \mathcal{C}_X h_t, \alpha_j \rangle_{\mathbf{L}^2(P_Z)} = 0$. □

Using Lemma B.1.1 and Lemma B.1.2, we can simply accumulate the reduction incurred on every iteration.

Proposition B.1.2. *Define the constants $(u_j)_{j=1}^m$ and $(v_j)_{j=1}^m$ by*

$$\mu_X \bar{h} = \sum_{j=1}^m u_j \alpha_j \text{ and } \mu_Z \bar{h} = \sum_{j=1}^m v_j \beta_j.$$

Then, we may quantify the variance reduction achieved by $t+1$ iterations of the $\mathcal{C}_Z \mathcal{C}_X$ operator as

$$\begin{aligned}
\mathbb{E}\left[\bar{h}^2\right] - \mathbb{E}\left[((\mathcal{C}_Z \mathcal{C}_X)^{t+1} \bar{h})^2\right] &= \sum_{j=2}^m \left\{ u_j^2 + (v_j - s_j u_j)^2 \left[1 + \frac{s_j^2 (1 - s_j^{4t})}{1 - s_j^2} \right] \right\} \\
&\to \sum_{j=2}^m \left[u_j^2 + \frac{(v_j - s_j u_j)^2}{1 - s_j^2} \right]
\end{aligned}$$

as $t \to \infty$.

Proof. Apply Lemma B.1.1 $(t+1)$-times so that

$$\begin{aligned}\mathbb{E}\left[((\mathcal{C}_Z\mathcal{C}_X)^{t+1}\bar{h})^2\right] &= \mathbb{E}\left[\bar{h}^2\right] - \sum_{j=2}^{m}\sum_{\tau=0}^{t}\left[(1+s_j^2)u_{\tau,j}^2 + v_{\tau,j}^2 - 2s_j u_{\tau,j}v_{\tau,j}\right] \\ &= \mathbb{E}\left[\bar{h}^2\right] - \sum_{j=2}^{m}\left[v_{0,j}^2 - 2s_j u_{0,j}v_{0,j} + \sum_{\tau=0}^{t}(1+s_j^2)u_{\tau,j}^2\right]\end{aligned}$$

as by Lemma B.1.2, we have that $v_{\tau,j} = 0$ for $\tau > 0$. Next, we unroll the definition of $u_{\tau,j}$ so that

$$\begin{aligned}u_{\tau,j} &= s_j^2 u_{\tau-1,j} - s_j v_{\tau-1,j} \\ &= s_j^2(s_j^2 u_{\tau-2,j} - s_j v_{\tau-2,j}) - s_j v_{\tau-1,j} \\ &= s_j^{2\tau-2}(s_j^2 u_{0,j} - s_j v_{0,j})\end{aligned}$$

for $\tau > 0$, yielding

$$\begin{aligned}&\mathbb{E}\left[\bar{h}^2\right] - \mathbb{E}\left[((\mathcal{C}_Z\mathcal{C}_X)^{t+1}\bar{h})^2\right] \\ &= \sum_{j=2}^{m}\left[u_{0,j}^2 + (v_{0,j} - s_j u_{0,j})^2 + (1+s_j^2)(s_j^2 u_{0,j} - s_j v_{0,j})^2\sum_{\tau=1}^{t}(s_j^4)^{\tau-1}\right] \\ &= \sum_{j=2}^{m}\left[u_{0,j}^2 + (v_{0,j} - s_j u_{0,j})^2 + (1+s_j^2)(s_j^2 u_{0,j} - s_j v_{0,j})^2\sum_{\tau=0}^{t-1}(s_j^4)^{\tau}\right] \\ &= \sum_{j=2}^{m}\left[u_{0,j}^2 + (v_{0,j} - s_j u_{0,j})^2 + \frac{s_j^2(1+s_j^2)(v_{0,j} - s_j u_{0,j})^2(1-s_j^{4t})}{1-s_j^4}\right] \\ &= \sum_{j=2}^{m}\left[u_{0,j}^2 + (v_{0,j} - s_j u_{0,j})^2 + \frac{s_j^2(v_{0,j} - s_j u_{0,j})^2(1-s_j^{4t})}{1-s_j^2}\right].\end{aligned}$$

Substitute $u_{0,j} = u_j$ and $v_{0,j} = v_j$ to complete the proof. □

We also present the corresponding result for k odd. The proof follows similarly by repeated application of the operator $\mathcal{C}_Z\mathcal{C}_X$. However, the iterations will be compared to

$\sigma_1^2 = \mathbb{E}_P\left[(\mathcal{C}_X\bar{h})^2\right]$, as we consider $\mathcal{C}_X\bar{h}$ as the "first" iteration to this process.

Proposition B.1.3. *Define the constants* $(u_j)_{j=1}^m$ *by*

$$\mu_Z\mathcal{C}_X\bar{h} = \sum_{j=1}^m u_j\beta_j.$$

Then, we may quantify the variance reduction achieved by $t+1$ *iterations of the* $\mathcal{C}_X\mathcal{C}_Z$ *operator as*

$$\begin{aligned}\mathbb{E}\left[(\mathcal{C}_X\bar{h})^2\right] - \mathbb{E}\left[((\mathcal{C}_X\mathcal{C}_Z)^{t+1}\mathcal{C}_X\bar{h})^2\right] &= \sum_{j=2}^m \left\{u_j^2 + (s_j u_j)^2\left[1 + \frac{s_j^2(1-s_j^{4t})}{1-s_j^2}\right]\right\} \\ &\to \sum_{j=2}^m \left(\frac{1+s_j^2}{1-s_j^2}\right)u_j^2\end{aligned}$$

as $t \to \infty$.

In order to have full monotonicity, we also need that $\sigma_0^2 \geq \sigma_1^2$. This follows by orthogonality, as

$$\sigma_0^2 = \mathbb{E}\left[\bar{h}^2\right] = \mathbb{E}\left[(\mathcal{C}_X\bar{h})^2\right] + \mathbb{E}\left[(\mu_X\bar{h})^2\right] = \sigma_1^2 + \mathbb{E}\left[(\mu_X\bar{h})^2\right] \geq \sigma_1^2. \tag{B.2}$$

Thus, we can combine Proposition B.1.3 and (B.2) to fully quantify the relationship between σ_0^2 and σ_k^2 for k odd.

B.2 Information Projections

This section is dedicated to deriving three representations of the balancing procedure as projections in various statistical divergences, as shown in Figure 4.1.

We consider two sets of probability measures denoted by $\Pi_X = \{Q : Q_X = P_X\}$ and $\Pi_Z = \{Q : Q_Z = P_Z\}$. The marginal matching steps are written as projections in terms of a statistical divergence D (precisely, an f-divergence) in the form

$$\frac{P_X}{P_{n,X}^{(k-1)}} \otimes P_n^{(k-1)} = \underset{Q\in\Pi_X}{\arg\min}\, D(Q\|P_n^{(k-1)}), \qquad \frac{P_Z}{P_{n,Z}^{(k-1)}} \otimes R = \underset{Q\in\Pi_Z}{\arg\min}\, D(Q\|P_n^{(k-1)}).$$

We provide the derivations for three common choices of D: Kullback-Leibler (KL), reverse KL, and χ^2. Using this viewpoint, and simply assuming the positivity of the marginal measures P_X and P_Z, we derive an upper bound in Proposition B.2.5 that is *constant* in k. This is an improvement over the recent work of Albertus and Berthet [2019], in which they show an upper bound that scales *exponentially* in k.

The KL representation will be used in the proof of Proposition B.2.5, which (recalling the sequence $(P_n^{(k)})_{k\geq 1}$ from (4.2)), controls the error between $P_{n,Z}^{(k)}$ and P_Z for k odd and $P_{n,X}^{(k)}$ and P_X for k even.

B.2.1 Balancing as Information Projections

The arguments for three information divergences (KL, reverse KL, and χ^2) are contained in the following propositions.

Proposition B.2.1 (Projection in KL-Divergence)**.** *Assume that $P_X \ll R_X$ and $P_Z \ll R_Z$, and define*

$$Q^\star := \operatorname*{arg\,min}_{Q\in\Pi_X} \mathrm{KL}(Q\|R), \quad P^\star := \operatorname*{arg\,min}_{Q\in\Pi_Z} \mathrm{KL}(Q\|R). \tag{B.3}$$

Then, it holds that

$$Q^\star(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} P_X(\boldsymbol{x})R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x})>0 \\ 0 & \text{if } R_X(\boldsymbol{x})=0 \end{cases} \tag{B.4}$$

and

$$P^\star(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} P_Z(\boldsymbol{z})R_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) & \text{if } R_Z(\boldsymbol{z})>0 \\ 0 & \text{if } R_Z(\boldsymbol{z})=0 \end{cases}. \tag{B.5}$$

Proof. In the case that $Q(\boldsymbol{x},\boldsymbol{z})=0$, we apply the convention that $0\log 0 = 0$. Consider the

case $Q^\star$, the projection of R onto Π_X. Write

$$\begin{aligned}
\mathrm{KL}(Q\|R) &= \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{z}\in\mathcal{Z}} Q(\boldsymbol{x},\boldsymbol{z}) \log \tfrac{Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x})Q_X(\boldsymbol{x})}{R_{Z|X}(\boldsymbol{z}|\boldsymbol{x})R_X(\boldsymbol{x})} \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} Q_X(\boldsymbol{x}) \left[\sum_{\boldsymbol{z}\in\mathcal{Z}} Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \log \tfrac{Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x})Q_X(\boldsymbol{x})}{R_{Z|X}(\boldsymbol{z}|\boldsymbol{x})R_X(\boldsymbol{x})}\right] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} Q_X(\boldsymbol{x}) \left[\sum_{\boldsymbol{z}\in\mathcal{Z}} Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \log \tfrac{Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x})}{R_{Z|X}(\boldsymbol{z}|\boldsymbol{x})} + \sum_{\boldsymbol{z}\in\mathcal{Z}} Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \log \tfrac{Q_X(\boldsymbol{x})}{R_X(\boldsymbol{x})}\right] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} Q_X(\boldsymbol{x}) \left[\sum_{\boldsymbol{z}\in\mathcal{Z}} Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \log \tfrac{Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x})}{R_{Z|X}(\boldsymbol{z}|\boldsymbol{x})}\right] + \sum_{\boldsymbol{x}\in\mathcal{X}} Q_X(\boldsymbol{x}) \log \tfrac{Q_X(\boldsymbol{x})}{R_X(\boldsymbol{x})} \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} Q_X(\boldsymbol{x})\, \mathrm{KL}(Q_{Z|X}(\cdot|\boldsymbol{x})\|R_{Z|X}(\cdot|\boldsymbol{x})) + \mathrm{KL}(Q_X\|R_X) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} P_X(\boldsymbol{x})\, \mathrm{KL}(Q_{Z|X}(\cdot|\boldsymbol{x})\|R_{Z|X}(\cdot|\boldsymbol{x})) + \mathrm{KL}(P_X\|R_X),
\end{aligned}$$

where the last line is due to the marginal constraint $Q \in \Pi_X$. For the above to be well defined, we need that $P_X \ll R_X$ so that $\mathrm{KL}(P_X\|R_X) < +\infty$. The above is minimized when $Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) = R_{Z|X}(\boldsymbol{z}|\boldsymbol{x})$ for all $(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}\times\mathcal{Z}$ such that $Q_X(\boldsymbol{x}) = P_X(\boldsymbol{x}) > 0$. The case of $P^\star$ follows analogously when using that $P_Z \ll R_Z$. □

Proposition B.2.2 (Projection in Reverse KL-Divergence)**.** *Assume that $P_Z \ll R_X$ and $P_Z \ll R_Z$, and define*

$$Q^\star := \operatorname*{arg\,min}_{Q\in\Pi_X} \mathrm{KL}(R\|Q), \qquad P^\star := \operatorname*{arg\,min}_{Q\in\Pi_Z} \mathrm{KL}(R\|Q). \tag{B.6}$$

Then, it holds that

$$Q^\star(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} P_X(\boldsymbol{x})R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x}) > 0 \\ 0 & \text{if } R_X(\boldsymbol{x}) = 0 \end{cases} \tag{B.7}$$

and

$$P^\star(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} P_Z(\boldsymbol{z})R_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) & \text{if } R_Z(\boldsymbol{z}) > 0 \\ 0 & \text{if } R_Z(\boldsymbol{z}) = 0 \end{cases}. \tag{B.8}$$

Proof. In the case that $R(\boldsymbol{x},\boldsymbol{z}) = 0$, we apply the convention that $0 \log 0 = 0$. Note that minimizing $\mathrm{KL}(R\|Q)$ over Q is equivalent to minimizing $-\sum_{\boldsymbol{x},y} R(\boldsymbol{x},\boldsymbol{z}) \log Q(\boldsymbol{x},\boldsymbol{z})$ (i.e. the cross entropy). Consider the case $Q^\star$, the projection of R onto Π_X. Because $R \ll Q$ for $\mathrm{KL}(R\|Q) < +\infty$ to hold, we have that $R(\boldsymbol{x}) > 0 \implies Q(\boldsymbol{x}) > 0$, so that $Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x})$ is well-defined. Write

$$\begin{aligned} &-\sum_{\boldsymbol{x},y} R(\boldsymbol{x},\boldsymbol{z}) \log Q(\boldsymbol{x},\boldsymbol{z}) \\ &= -\sum_{\boldsymbol{x}\in\mathcal{X}} R_X(\boldsymbol{x}) \log Q_X(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\mathcal{X}} R(\boldsymbol{x}) \sum_{\boldsymbol{z}\in\mathcal{Z}} R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \log Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \\ &= -\sum_{\boldsymbol{x}\in\mathcal{X}} R_X(\boldsymbol{x}) \log P_X(\boldsymbol{x}) + \sum_{\boldsymbol{x}\in\mathcal{X}} R_X(\boldsymbol{x}) \left[-\sum_{\boldsymbol{z}\in\mathcal{Z}} R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \log Q_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) \right]. \end{aligned}$$

The first term does not depend on Q due to the marginal constraint $Q \in \Pi_X$. The second term is the expectation of the cross entropy from $R_{Z|X}$ to $Q_{Z|X}$ over R_X, which is minimized if $R_{Z|X} = Q_{Z|X}$. We have specified $Q_{Z|X}$ and Q_X, completing the proof. □

The projection result for χ^2-divergence requires a few more intermediate steps. Let $\mathbf{1}$ denote the function that is identically equal to 1. Consider the following optimization problem, which is the subject of the subsequent lemmas:

$$\min_{\zeta\in\mathcal{A}_X} \|\mathbf{1} - \zeta\|^2_{\mathbf{L}^2(R)}, \tag{B.9}$$

where

$$\mathcal{A}_X := \left\{ f : \mathcal{X}\times\mathcal{Z} \to \mathbb{R} \text{ satisfying } \sum_{\boldsymbol{z}\in\mathcal{Z}} f(\boldsymbol{x},\boldsymbol{z}) R(\boldsymbol{x},\boldsymbol{z}) = P_X(\boldsymbol{x}) \text{ for any } x \in \mathcal{X} \right\}.$$

Lemma B.2.1. *Assume that $P_X \ll R_X$, and define The problem (B.9) is feasible, and its solution can be written as*

$$\zeta^\star = \mathcal{C}_X^R(\mathbf{1} - f) + f$$

for any $f \in \mathbf{L}^2(R)$, where the linear operator $\mathcal{C}_X^R$ is specified by

$$[\mathcal{C}_X^R g](\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}, \boldsymbol{z}) - \sum_{\boldsymbol{z}' \in \mathcal{Z}} g(\boldsymbol{x}, y') R_{Z|X}(\boldsymbol{z}'|\boldsymbol{x}).$$

Proof. First, we establish feasibility by letting

$$f(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} P_X(\boldsymbol{x})/R_X(\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x}) > 0 \\ 1 & \text{otherwise} \end{cases}.$$

This function does not depend on the second input $\boldsymbol{z}$. Because we assumed that $P_X \ll R_X$, we have that the terms of $f(\boldsymbol{x}, \boldsymbol{z})$ for which $R_X(\boldsymbol{x}) = 0$ do not affect whether $\sum_{\boldsymbol{z} \in \mathcal{Z}} f(\boldsymbol{x}, \boldsymbol{z}) R(\boldsymbol{x}, \boldsymbol{z}) = P_X(\boldsymbol{x})$, because $P_X(\boldsymbol{x}) = 0$ in these cases. In the remainder of this proof, we will show that (B.9) is an affine projection problem, and find its solution by converting it to a subspace projection problem. Indeed, consider $f_1, \ldots, f_r \in \mathcal{A}_X$, and $\alpha_1, \ldots, \alpha_r \in \mathbb{R}$ such that $\sum_{j=1}^r \alpha_j = 1$. Then,

$$\sum_{\boldsymbol{z} \in \mathcal{Z}} \left[\sum_{j=1}^r \alpha_j f_j(\boldsymbol{x}, \boldsymbol{z}) \right] \cdot R(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^r \alpha_j \left[\sum_{\boldsymbol{z} \in \mathcal{Z}} f_j(\boldsymbol{x}, \boldsymbol{z}) R(\boldsymbol{x}, \boldsymbol{z}) \right] = P_X(\boldsymbol{x}),$$

indicating that $\sum_{j=1}^r \alpha_j f_j(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{A}_X$ and $\mathcal{A}_X$ is an affine subset of $\mathbf{L}^2(R)$. Define

$$\mathcal{S}_X := \left\{ g : \mathcal{X} \times \mathcal{Z} \to \mathbb{R} \text{ satisfying } \sum_{\boldsymbol{z} \in \mathcal{Z}} g(\boldsymbol{x}, \boldsymbol{z}) R(\boldsymbol{x}, \boldsymbol{z}) = 0 \text{ for any } x \in \mathcal{X} \right\}.$$

Then, for any $f \in \mathcal{A}_X$, we have that $g \in \mathcal{S}_X$ if and only if $g + f \in \mathcal{A}_X$. Taking any $f \in \mathcal{A}_X$, letting $\phi^\star$ be the solution of

$$\min_{\phi \in \mathcal{S}_X} \|\mathbf{1} - f - \phi\|^2_{\mathbf{L}^2(R)}, \tag{B.10}$$

we will have that $\phi^\star + f$ will be the solution of (B.9). The remainder of the proof is showing that $\phi^\star = \mathcal{C}_X^R(\mathbf{1} - f)$.

First, define the operator μ_X^R by $[\mu_X g](\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z}' \in \mathcal{Z}} g(\boldsymbol{x}, y') R_{Z|X}(\boldsymbol{z}'|\boldsymbol{x})$, and note (by factoring out $R_X(\boldsymbol{x})$) that $g \in \mathcal{S}_X$ if and only if $\mu_X^R g = 0$. In addition, $\mu_X^R g$ is linear and idempotent as $\mu_X^R \mu_X^R g = \mu_X^R g$, so it is a projection operator in $\mathbf{L}^2(R)$. Thus, $\mathcal{S}_X$ is the orthogonal complement of range(μ_X^R), and the solution of (B.10) is given by $(I - \mu_X^R)(\mathbf{1} - f) = \mathcal{C}_X^R(\mathbf{1} - f)$, because $\mathcal{C}_X^R = I - \mu_X^R$. The claim is proved. □

Lemma B.2.2. *Assume that $P_X \ll R_X$. Define*

$$Q^\star := \arg\min_{Q \in \Pi_X} \chi^2(Q \| R). \tag{B.11}$$

and let $\zeta^\star$ be the solution of problem (B.9). Then,

$$Q^\star(\boldsymbol{x}, \boldsymbol{z}) = \zeta^\star(\boldsymbol{x}, \boldsymbol{z}) R(\boldsymbol{x}, \boldsymbol{z}) = \begin{cases} P_X(\boldsymbol{x}) R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x}) > 0 \\ 0 & \text{if } R_X(\boldsymbol{x}) = 0 \end{cases}. \tag{B.12}$$

Proof. First, by reparametrizing the problem (B.11) as finding ζ such that $Q(\boldsymbol{x}, \boldsymbol{z}) = \zeta(\boldsymbol{x}, \boldsymbol{z}) R(\boldsymbol{x}, \boldsymbol{z})$, we can compute its solution by solving

$$\min_{\zeta \in \mathcal{A}_X, \zeta \geq 0} \|\mathbf{1} - \zeta\|^2_{\mathbf{L}^2(R)}, \tag{B.13}$$

Notice that we also have a non-negativity constraint, as opposed to (B.9). If $\zeta^\star$ solves (B.9) and happens to be non-negative, then we have that $\zeta^\star$ solves (B.13) as well and the first equality of (B.12) is satisfied by definition. We show the second equality of (B.12) by direct computation, which also establishes the non-negativity of $\zeta^\star$ simultaneously.

Apply Lemma B.2.1 with

$$f(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} P_X(\boldsymbol{x}) / R_X(\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x}) > 0 \\ 1 & \text{otherwise} \end{cases}.$$

so that

$$\begin{aligned}\zeta^\star(\boldsymbol{x},\boldsymbol{z}) &= \mathcal{C}_X^R\left(\mathbf{1}-f\right)(\boldsymbol{x},\boldsymbol{z}) + f(\boldsymbol{x},\boldsymbol{z}) \\ &= \left[\sum_{z\in\mathcal{Z}} f(\boldsymbol{x},z)R_{Z|X}(z|\boldsymbol{x}) - f(\boldsymbol{x},\boldsymbol{z})\right] + f(\boldsymbol{x},\boldsymbol{z}) \\ &= f(\boldsymbol{x},y')\end{aligned}$$

for any $\boldsymbol{z}' \in \mathcal{Z}$. Thus, the likelihood ratio of $Q^\star$ with respect to R is a marginal reweighting. Accordingly,

$$Q^\star(\boldsymbol{x},\boldsymbol{z}) = \zeta^\star(\boldsymbol{x},\boldsymbol{z})R(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} P_X(\boldsymbol{x})R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x}) > 0 \\ 0 & \text{if } R_X(\boldsymbol{x}) = 0 \end{cases},$$

completing the proof. □

Proposition B.2.3 (Projection in χ^2-Divergence)**.** *Assume that $P_X \ll R_X$ and $P_Z \ll R_Z$. Define*

$$Q^\star := \operatorname*{arg\,min}_{Q\in\Pi_X} \chi^2(Q\|R), \quad P^\star := \operatorname*{arg\,min}_{Q\in\Pi_Z} \chi^2(Q\|R). \tag{B.14}$$

Then, it holds that

$$\begin{aligned} Q^\star(\boldsymbol{x},\boldsymbol{z}) &= \begin{cases} P_X(\boldsymbol{x})R_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) & \text{if } R_X(\boldsymbol{x}) > 0 \\ 0 & \text{if } R_X(\boldsymbol{x}) = 0 \end{cases} \\ P^\star(\boldsymbol{x},\boldsymbol{z}) &= \begin{cases} P_Z(\boldsymbol{z})R_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) & \text{if } R_Z(\boldsymbol{z}) > 0 \\ 0 & \text{if } R_Z(\boldsymbol{z}) = 0 \end{cases}. \end{aligned} \tag{B.15}$$

Proof. The first equality of (B.15) follows by the claim of Lemma B.2.2. The second equality follows by repeating the argument of Lemma B.2.1 and Lemma B.2.2 with $(X,\boldsymbol{x})$ and $(Z,\boldsymbol{z})$ swapped. □

B.2.2 Proof of Main Results

We may now control the errors of the ratio of marginals using the projection interpretation established in the previous sections. Recall the event $\mathcal{S}$. The following result, the monotonicity of the marginal violation terms in terms of KL, will be useful in the bound.

Proposition B.2.4. *[Nutz, 2021, Proposition 6.10] Under the event $\mathcal{S}$, it holds that*

$$\mathrm{KL}(P_{n,X}^{(0)} \| P_X) \geq \mathrm{KL}(P_Z \| P_{n,Z}^{(1)}) \geq \mathrm{KL}(P_{n,X}^{(2)} \| P_X) \geq \dots$$

We give the following result for $\mathcal{X}$, and the analogous claim holds on $\mathcal{Z}$.

Proposition B.2.5. *Assume that $P_{n,X}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$. It holds that*

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left| \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq \begin{cases} \max\{n-1, 1\} & \text{if } k = 1 \\ \max\{1/p_\star^2 - 1, 1\} & \text{if } k > 1. \end{cases} \tag{B.16}$$

In addition, we have that

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left| \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq \begin{cases} n\sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X} \| P_X)} & \text{if } k = 1 \\ \frac{1}{p_\star^2}\sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X} \| P_X)} & \text{if } k > 1 \end{cases}.$$

Moreover, when $\mathrm{KL}(P_{n,X} \| P_X) \leq p_\star^2/2$, we have

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left| \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq \frac{2}{p_\star}\sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X} \| P_X)}. \tag{B.17}$$

Proof. We first show that $P_n^{(k-1)}(\boldsymbol{x}) \geq 1/n$ for $k = 1$ and $P_n^{(k-1)}(\boldsymbol{x}) \geq p_\star^2$ for $k > 1$. In the case that $k = 1$, the result follows directly from the event $\mathcal{S}$. For $k > 1$ such that k is odd, we have that for $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned} P_{n,X}^{(k-1)}(\boldsymbol{x}) &= \sum_{\boldsymbol{z} \in \mathcal{Z}} P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \frac{P_Z(\boldsymbol{z})}{P_{n,Z}^{(k-2)}(\boldsymbol{z})} P_n^{(k-2)}(\boldsymbol{x}, \boldsymbol{z}) \\ &\geq p_\star \sum_{\boldsymbol{z} \in \mathcal{Z}} P_n^{(k-2)}(\boldsymbol{x}, \boldsymbol{z}) = p_\star P_{n,X}^{(k-2)}(\boldsymbol{x}) = p_\star P_X(\boldsymbol{x}) \geq p_\star^2. \end{aligned}$$

The result for k even can be proven similarly. We now proceed to prove the inequalities given in the statement, which will rely on the lower bound above.

Proving the first inequality. Then, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\left| \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| = \max \left\{ \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} - 1, 1 - \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} \right\} \leq \begin{cases} \max\{n-1, 1\} & \text{if } k = 1 \\ \max\{1/p_\star^2 - 1, 1\} & \text{if } k > 1 \end{cases},$$

which is the desired result for the first inequality.

Proving the second and third inequalities. Consider an odd $k \geq 1$. By the definition of total variation distance, it holds that

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left| P_X(\boldsymbol{x}) - P_{n,X}^{(k-1)}(\boldsymbol{x}) \right| \leq \mathrm{TV}(P_{n,X}^{(k-1)}, P_X).$$

According to Pinsker's inequality, we have that $\mathrm{TV}(P_{n,X}^{(k-1)}, P_X) \leq \sqrt{\frac{1}{2} \mathrm{KL}(P_{n,X}^{(k-1)} \| P_X)}$, and so we have that

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left| P_X(\boldsymbol{x}) - P_{n,X}^{(k-1)}(\boldsymbol{x}) \right| \leq \sqrt{\frac{1}{2} \mathrm{KL}(P_{n,X}^{(k-1)} \| P_X)} \leq \sqrt{\frac{1}{2} \mathrm{KL}(P_{n,X}^{(0)} \| P_X)},$$

where the last inequality follows from the monotonicity of Sinkhorn iterations given in Proposition B.2.4. We apply the lower bounds to write

$$\max_{\boldsymbol{x} \in \mathcal{X}} \left| \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq \begin{cases} n\sqrt{\frac{1}{2} \mathrm{KL}(P_{n,X} \| P_X)} & \text{if } k = 1 \\ \frac{1}{p_\star^2}\sqrt{\frac{1}{2} \mathrm{KL}(P_{n,X} \| P_X)} & \text{if } k > 1 \end{cases}.$$

Finally, when $\sqrt{\frac{1}{2} \mathrm{KL}(P_{n,X} \| P_X)} \leq p_\star/2$, we have that $\max_{\boldsymbol{x} \in \mathcal{X}} \left| P_X(\boldsymbol{x}) - P_{n,X}^{(k-1)}(\boldsymbol{x}) \right| \leq p_\star/2$ and thus

$$\min_{\boldsymbol{x} \in \mathcal{X}} P_{n,X}^{(k-1)}(\boldsymbol{x}) \geq \min_{\boldsymbol{x} \in \mathcal{X}} P_X(\boldsymbol{x}) - \max_{\boldsymbol{x} \in \mathcal{X}} \left| P_{n,X}^{(k-1)}(\boldsymbol{x}) - P_X(\boldsymbol{x}) \right| \geq \frac{p_\star}{2}.$$

Hence,

$$\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{P_X(\boldsymbol{x})}{P_{n,X}^{(k-1)}(\boldsymbol{x})}-1\right|\leq\frac{\max_{\boldsymbol{x}\in\mathcal{X}}\left|P_{n,X}^{(k-1)}(\boldsymbol{x})-P_X(\boldsymbol{x})\right|}{\min_{\boldsymbol{x}\in\mathcal{X}}P_{n,X}^{(k-1)}(\boldsymbol{x})}\leq\frac{2}{p_\star}\sqrt{\frac{1}{2}\mathrm{KL}(P_{n,X}\|P_X)}.$$

Now, for k even, set $k=2t$ for $t\geq 0$. We have that

$$\max_{\boldsymbol{z}\in\mathcal{Z}}\left|P_{n,Z}^{(2t-1)}(\boldsymbol{z})-P_Z(\boldsymbol{z})\right|\leq\mathrm{TV}(P_{n,Z}^{(2t-1)},P_Z)\leq\sqrt{\frac{1}{2}\mathrm{KL}(P_Z\|P_{n,Z}^{(2t-1)})}.$$

Invoke Proposition B.2.4 once again to achieve

$$\sqrt{\frac{1}{2}\mathrm{KL}(P_Z\|P_{n,Z}^{(2t-1)})}\leq\sqrt{\frac{1}{2}\mathrm{KL}(P_{n,X}\|P_X)},$$

which completes the proof. □

B.3 Statistical Analysis of Balancing Estimators

This section contains the proof of the main result, namely Theorem 4.3.1. We first consolidate notation and then give a broad outline of the proof for readability. Let the expectation of a function h under a probability measure Q on $\mathcal{X}\times\mathcal{Z}$ by denoted by

$$Q(h)=\sum_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}\in\mathcal{Z}}h(\boldsymbol{x},\boldsymbol{z})Q(\boldsymbol{x},\boldsymbol{z})$$

so that

$$\boldsymbol{\theta}_n^{(k)}=P_n^{(k)}(h),\quad\boldsymbol{\theta}=P(h),$$

and

$$\mathbb{G}_n^{(k)}(h)=\sqrt{n}[P_n^{(k)}-P](h)=\sqrt{n}(P_n^{(k)}(h)-P(h)).\tag{B.18}$$

Recalling in addition that $\mathcal{C}_k=\mathcal{C}_X$ for k odd and $\mathcal{C}_k=\mathcal{C}_Z$ for k even. The event

$$\mathcal{S}:=\left\{\mathrm{Supp}(P_{n,X})=\mathrm{Supp}(P_X)\text{ and }\mathrm{Supp}(P_{n,Z})=\mathrm{Supp}(P_Z)\right\},\tag{B.19}$$

is used for purely technical reasons in many results.

Proof Outline We first establish that the recursion formula

$$[P_n^{(k)} - P](h) = [P_n^{(k-1)} - P](\mathcal{C}_k h) + V_n^{(k-1)}(\mathcal{C}_k h)$$

holds in Proposition B.3.1, where

$$V_n^{(k-1)}(h) = \begin{cases} \sum_{\boldsymbol{x},y} \left(\frac{P_X}{P_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1 \right) h(\boldsymbol{x}, \boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ odd} \\ \sum_{\boldsymbol{x},y} \left(\frac{P_Z}{P_{n,Z}^{(k-1)}}(\boldsymbol{z}) - 1 \right) h(\boldsymbol{x}, \boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ even} \end{cases}. \tag{B.20}$$

The quantity $V_n^{(k-1)}(\mathcal{C}_k h)$ describes an error term that accumulates for each iteration of balancing, which explains why k must be scaled appropriately against n to ensure the error does not accumulate too fast. Applying the recursion repeatedly to the balanced sequence $(P_n^{(k)})_{k \geq 1}$ and unrolling the recursion, we see that when k is odd,

$$\begin{aligned} [P_n^{(k)} - P](h) &= [P_n^{(k-1)} - P](\mathcal{C}_X h) + V_n^{(k-1)}(\mathcal{C}_X h) \\ &= [P_n^{(k-2)} - P](\mathcal{C}_Z \mathcal{C}_X h) + V_n^{(k-2)}(\mathcal{C}_Z \mathcal{C}_X h) + V_n^{(k-1)}(\mathcal{C}_X h) \\ &= \underbrace{[P_n^{(0)} - P](\mathcal{C}_1 \dots \mathcal{C}_k h)}_{\text{first-order term}} + \underbrace{\textstyle\sum_{\ell=1}^{k} V_n^{(\ell-1)}(\mathcal{C}_\ell \dots \mathcal{C}_k h)}_{\text{higher-order term}} \end{aligned} \tag{B.21}$$

Additionally, let $h_{\ell,k} := \mathcal{C}_\ell \dots \mathcal{C}_k h$, so that the first-order term can be written as $P_n^{(0)}(h_{1,k}) - P(h_{1,k})$ higher-order term can also be written as $\sum_{\ell=1}^{k} V_n^{(\ell-1)}(h_{\ell,k})$. Because our original goal is to upper bound the mean squared error, we use the expansion above to write

$$\begin{aligned} \mathbb{E}\,|P_n^{(k)}(h) - P(h)|^2 &\leq \mathbb{E}\,|P_n^{(0)}(h_{1,k}) - P(h_{1,k})|^2 \\ &+ 2\mathbb{E}\,|P_n^{(0)}(h_{1,k}) - P(h_{1,k})| \left| \textstyle\sum_{\ell=1}^{k} V_n^{(\ell-1)}(h_{\ell,k}) \right| + \mathbb{E} \left| \textstyle\sum_{\ell=1}^{k} V_n^{(\ell-1)}(h_{\ell,k}) \right|^2 \end{aligned}$$

Regarding the first term, we have that $\mathbb{E}\,|P_n^{(0)}(h_{1,k}) - P(h_{1,k})|^2 = \sigma_k^2/n$, which is the dominant term in Theorem 4.3.1. Thus, the remaining challenge of the proof will be to upper bound the cross term and squared term and show their dependence on n. The dominant term of these

two will be the cross term, as we will essentially show that $|P_n^{(0)}(h_{1,k}) - P(h_{1,k})|$ is $O(n^{-1/2})$ with high probability and that $|\sum_{\ell=1}^{k} V_n^{(\ell-1)}(h_{\ell,k})|$ is in fact $O(n^{-1})$ with high probability. As stated in Section 4.3, a key intermediate result in controlling the higher-order term is Proposition B.2.5, whose proof is given in Appendix B.2. The remaining subsections walk through these steps in detail.

B.3.1 Recursion of Estimation Error

We first recall that the sequence $(P_n^{(k)})_{k\geq 1}$ can be computed with the following formula:

$$P_n^{(0)}(\boldsymbol{x},\boldsymbol{z}) := P_n(\boldsymbol{x},\boldsymbol{z}) \text{ and } P_n^{(k)}(\boldsymbol{x},\boldsymbol{z}) := \begin{cases} \frac{P_X}{P_{n,X}^{(k-1)}}(\boldsymbol{x}) P_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) & k \text{ odd} \\ \frac{P_Z}{P_{n,Z}^{(k-1)}}(\boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) & k \text{ even} \end{cases}. \tag{B.22}$$

Proposition B.3.1 establishes the conditions under which these steps are well-defined (i.e. $P_{n,X}^{(k-1)}(\boldsymbol{x}) > 0$ and $P_{n,Z}^{(k-1)}(\boldsymbol{z}) > 0$).

Proposition B.3.1. *Let $(P_n^{(k)})_{k\geq 1}$, be a sequence computed according to (4.2). These iterations are well-defined under the event $\mathcal{S}$, and for $\mathbb{G}_n^{(k)}$ defined in (B.18) and $V_n^{(k)}$ defined in (B.20), it holds that*

$$\mathbb{G}_n^{(k)}(h) = \mathbb{G}_n^{(k-1)}(h) + \sqrt{n} V_n^{(k-1)}(h). \tag{B.23}$$

and

$$\mathbb{G}_n^{(k)}(h) = \mathbb{G}_n^{(k-1)}(\mathcal{C}_k h) + \sqrt{n} V_n^{(k-1)}(\mathcal{C}_k h). \tag{B.24}$$

Proof. First, assume that $P_{n,X}^{(k-1)}(\boldsymbol{x}) > 0$ and $P_{n,Z}^{(k-1)}(\boldsymbol{z}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$ so that we may establish the recursion, which we will show by induction toward the end of the proof.

Consider the following steps in the case that k is odd:

$$
\begin{aligned}
&P_n^{(k)}(h) \\
&= \sum_{\boldsymbol{x},y} h(\boldsymbol{x},\boldsymbol{z}) P_n^{(k)}(\boldsymbol{x},\boldsymbol{z}) = \sum_{\boldsymbol{x},y} h(\boldsymbol{x},\boldsymbol{z}) \frac{P_X}{P_{n,X}^{(k-1)}}(\boldsymbol{x}) P_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) && \text{by (B.22) for } k \text{ odd} \\
&= \sum_{\boldsymbol{x},y} 1 \cdot h(\boldsymbol{x},\boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) + \sum_{\boldsymbol{x},y} \left[\frac{P_X}{P_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1 \right] \cdot h(\boldsymbol{x},\boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\
&= P_n^{(k-1)}(h) + V_n^{(k-1)}(h).
\end{aligned}
$$

Arguing analogously for k even and subtracting $P(h)$ on both sides, we have that

$$
[P_n^{(k)} - P](h) = [P_n^{(k-1)} - P](h) + V_n^{(k-1)}(h). \tag{B.25}
$$

We refer to this as the "uncentered" recursion, which proves (B.23).

We can then establish the following "centered" recursion using the following decomposition in the case of k odd.

$$
\begin{aligned}
&[P_n^{(k)} - P](h) \\
&= [P_n^{(k)} - P](\mathcal{C}_X h) + [P_n^{(k)} - P](\mu_X h) && h = \mathcal{C}_X h + \mu_X h \\
&= [P_n^{(k-1)} - P](\mathcal{C}_X h) + V_n^{(k-1)}(\mathcal{C}_X h) + [P_n^{(k)} - P](\mu_X h) && \text{apply (B.25) to } \mathcal{C}_X h \\
&= [P_n^{(k-1)} - P](\mathcal{C}_X h) + V_n^{(k-1)}(\mathcal{C}_X h). && P_n^{(k)}(\mu_X h) = P(\mu_X h)
\end{aligned}
$$

The last line follows because $\mu_X h$ is only a function on $\mathcal{X}$, and due to the definition of the marginal rebalancing iterations, $P_{n,X}^{(k)} = P_X$. This gives the desired formula by substituting (B.18).

We proceed to show that the iterations are well-defined. We will in fact show that $P_{n,X}^{(k-1)}(\boldsymbol{x}) > 0$ and $P_{n,Z}^{(k-1)}(\boldsymbol{z}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$. For $k = 1$, $P_{n,X}^{(0)}(\boldsymbol{x}) = P_{n,X}(\boldsymbol{x}) > 0$ and $P_{n,Z}^{(0)}(\boldsymbol{z}) = P_{n,Z}(\boldsymbol{z}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$ this holds under the event $\mathcal{S}$ by assumption. We argue by induction that this holds for all $k > 1$. Assume that the claim is

true for $\{1, \ldots, k-1\}$, and that k is even. Then,

$$\begin{aligned} P_{n,X}^{(k-1)}(\boldsymbol{x}) &= P_X(\boldsymbol{x}) > 0, \\ P_{n,Z}^{(k-1)}(\boldsymbol{z}) &= \sum_{\boldsymbol{x} \in \mathcal{X}} P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \frac{P_X}{P_{n,X}^{(k-2)}}(\boldsymbol{x}) P_n^{(k-2)}(\boldsymbol{x}, \boldsymbol{z}) \\ &\geq \min_{\boldsymbol{x} \in \mathcal{X}} \frac{P_X}{P_{n,X}^{(k-2)}}(\boldsymbol{x}) \cdot P_{n,Z}^{(k-2)}(\boldsymbol{z}) > 0 \end{aligned}$$

as $P_{n,X}^{(k-2)}(\boldsymbol{x}) > 0$ and $P_{n,Z}^{(k-2)}(\boldsymbol{z}) > 0$ by the inductive hypothesis. Arguing analogously for k odd achieves the claim. □

B.3.2 Technical Tools & Intermediate Results

Having established the backbone of the argument, we collect in this subsection some useful tools that are used in the remainder of the proofs.

The following result follows from the method of types in information theory and will be helpful in deriving the dependence of the higher-order term on n.

Theorem B.3.1. *[Cover, 1999, Theorem 11.2.1] Let ν be a discrete probability measure supported on m atoms. Let $U_1, \ldots, U_n \overset{i.i.d}{\sim} \nu$ and ν_n be the associated empirical measure. Then, we have for any $\varepsilon > 0$ that*

$$\mathbb{P}\left(\mathrm{KL}(\nu_n \| \nu) \geq \varepsilon\right) \leq 2^{-n\left(\varepsilon - m\frac{\log(n+1)}{n}\right)}.$$

We then provide a result that counts the number of terms that appear when repeatedly centering via the operators $\mathcal{C}_1, \ldots, \mathcal{C}_k$. This formalizes the pattern

$$\begin{aligned} \mathcal{C}_X &= I - \mu_X \\ \mathcal{C}_Z \mathcal{C}_X &= I - \mu_X - \mu_Z + \mu_Z \mu_X \\ \mathcal{C}_X \mathcal{C}_Z \mathcal{C}_X &= I - \mu_X - \mu_Z + \mu_Z \mu_X + \mu_X \mu_Z - \mu_X \mu_Z \mu_X, \end{aligned}$$

and so on. This will be useful when bounding $h_{\ell,k}$ uniformly.

Lemma B.3.1. *For any $k \geq 1$ and $\ell \in \{1, \ldots, k\}$,*

$$\mathcal{C}_\ell \ldots \mathcal{C}_k = I - \sum_{\tau=0}^{(k-\ell-1)/2} (\mu_X \mu_Z)^\tau \mu_X - \sum_{\tau=0}^{(k-\ell-1)/2} (\mu_Z \mu_X)^\tau \mu_Z + \sum_{\tau=1}^{(k-\ell)/2} (\mu_X \mu_Z)^\tau + \sum_{\tau=1}^{(k-\ell)/2} (\mu_Z \mu_X)^\tau + (-1)^{k-\ell+1} \mu_\ell \ldots \mu_k,$$

where the sum $\sum_{\tau=i}^{j}$ is 0 when $i > j$ and is $\sum_{\tau=i}^{\lfloor j \rfloor}$ when j is not an integer by convention.

Proof. We prove the claim by backward induction on ℓ, for the case that k is odd. In the case $\ell = k$, the claim holds because $\mathcal{C}_k = I - \mu_k$. Next, for any $\ell < k$, assume that the stated result holds for $\{\ell + 1, \ldots, k\}$. Then, if ℓ is also odd (so that $\mu_\ell = \mu_X$),

$$\begin{aligned}
\mathcal{C}_\ell \ldots \mathcal{C}_k &= \mathcal{C}_\ell \mathcal{C}_{\ell+1} \ldots \mathcal{C}_k \\
&= I - \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_X \mu_Z)^\tau \mu_X - \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_Z \mu_X)^\tau \mu_Z \\
&\quad + \sum_{\tau=1}^{(k-\ell-1)/2} (\mu_X \mu_Z)^\tau + \sum_{\tau=1}^{(k-\ell-1)/2} (\mu_Z \mu_X)^\tau + \mu_Z \underbrace{\cdots}_{k-\ell \text{ terms}} \mu_X \\
&\quad - \mu_X + \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_X \mu_Z)^\tau \mu_X + \sum_{\tau=0}^{(k-\ell-2)/2} \mu_X (\mu_Z \mu_X)^\tau \mu_Z \\
&\quad - \sum_{\tau=1}^{(k-\ell-1)/2} (\mu_X \mu_Z)^\tau - \sum_{\tau=1}^{(k-\ell-1)/2} \mu_X (\mu_Z \mu_X)^\tau - (\mu_X \mu_Z)^{(k-\ell)/2} \mu_X
\end{aligned}$$

The red terms and blue terms cancel out to zero. This leaves

$$\begin{aligned}
\mathcal{C}_\ell \ldots \mathcal{C}_k = I &- \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_X \mu_Z)^\tau \mu_X - \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_Z \mu_X)^\tau \mu_Z \\
&+ \sum_{\tau=1}^{(k-\ell-1)/2} (\mu_Z \mu_X)^\tau + (\mu_Z \mu_X)^{(k-\ell)/2} \\
&+ \sum_{\tau=0}^{(k-\ell-2)/2} \mu_X (\mu_Z \mu_X)^\tau \mu_Z + (-1)^{k-\ell+1} \mu_\ell \ldots \mu_k
\end{aligned}$$

wherein we combine the red terms and re-index the blue terms to get

$$\mathcal{C}_\ell \dots \mathcal{C}_k = I - \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_X \mu_Z)^\tau \mu_X - \sum_{\tau=0}^{(k-\ell-2)/2} (\mu_Z \mu_X)^\tau \mu_Z + \sum_{\tau=1}^{(k-\ell)/2} (\mu_Z \mu_X)^\tau + \sum_{\tau=1}^{(k-\ell)/2} (\mu_X \mu_Z)^\tau + (-1)^{k-\ell+1} \mu_\ell \dots \mu_k.$$

Finally, because $k - \ell$ is even when k is odd and ℓ is odd, we can set the upper bound of the first two sums to $(k - \ell - 1)/2$ without changing the number of terms. This proves the desired result. The result can be proved similarly when ℓ is even. As a result, we have proved the claim for any odd k and $\ell \leq k$. Similar arguments can be used for the case of k even and $\ell \leq k$. □

B.3.3 Analysis of Higher-Order Term

Returning to the outline at the start of this section, we may now bound the higher-order remainder term in (B.21), namely

$$\sum_{\ell=1}^{k} V_n^{(\ell-1)}(h_{\ell,k}) = \sum_{\ell=1}^{k} V_n^{(\ell-1)}(\mathcal{C}_\ell \dots \mathcal{C}_k h),$$

depends on controlling the quantity $V_n^{(k-1)}$ in the summation, which we recall for convenience:

$$V_n^{(k-1)}(h) = \begin{cases} \sum_{\boldsymbol{x},y} \left(\frac{P_X}{P_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1 \right) h(\boldsymbol{x}, \boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ odd} \\ \sum_{\boldsymbol{x},y} \left(\frac{P_Z}{P_{n,Z}^{(k-1)}}(\boldsymbol{z}) - 1 \right) h(\boldsymbol{x}, \boldsymbol{z}) P_n^{(k-1)}(\boldsymbol{x}, \boldsymbol{z}) & k \text{ even} \end{cases}. \tag{B.26}$$

Because we have established uniform control over the functions $P_X / P_{n,X}^{(k-1)} - 1$ and $P_Z / P_{n,Z}^{(k-1)} - 1$, via Proposition B.2.5 in Appendix B.2 we can now bound the full remainder in Proposition B.3.2.

We also make use of the following intermediate result, which controls how large the ℓ_∞-norm of the function h can grow after centering.

Lemma B.3.2. $\|h_{\ell,k}\|_\infty \leq 2(k - \ell + 1) \|h\|_\infty$.

Proof. Apply Lemma B.3.1 and the triangle inequality, so that we only need to count the number of terms that appear in the sums, adding 2 for the first and last term in the expression. We subtract 1 from the total, as one of either $(k-\ell)/2$ or $(k-\ell+1)/2$ will be a fraction. This yields $2(k-\ell+1)$ terms total, the desired result. □

We upper bound the sum in Proposition B.3.2. To do so, we introduce some notation. Consider B_1 and B_2 defined by

$$B_1 := M_1 \quad \text{and} \quad B_2 := \max_{2\le\ell\le k} M_\ell \quad \text{for} \quad M_\ell := \begin{cases} \max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{P_X(\boldsymbol{x})}{P_{n,X}^{(\ell-1)}(\boldsymbol{x})} - 1 \right| & \ell \text{ odd} \\ \max_{\boldsymbol{z}\in\mathcal{Z}} \left| \frac{P_Z(\boldsymbol{z})}{P_{n,Z}^{(\ell-1)}(\boldsymbol{z})} - 1 \right| & \ell \text{ even} \end{cases}$$

for $k \ge 1$. We also enumerate the sample spaces as $\mathcal{X} = \{\boldsymbol{x}_1, \dots, x_m\}$ and $\mathcal{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\}$, and define the function

$$\mathbf{1}_{jk}(\boldsymbol{x}, \boldsymbol{z}) := \begin{cases} \mathbb{1}\{\boldsymbol{x} = \boldsymbol{x}_j\} & k \text{ odd} \\ \mathbb{1}\{\boldsymbol{z} = \boldsymbol{z}_j\} & k \text{ even} \end{cases}.$$

This is an indicator function on the j-th element of either $\mathcal{X}$ or $\mathcal{Z}$, depending on whether k is odd or even. Finally, for any function h, use (under the event $\mathcal{S}$) recall the empirical process notation

$$\mathbb{G}_n^{(k)}(h) := \sqrt{n}\left(P_n^{(k)}(h) - P(h)\right). \tag{B.27}$$

Using this notation, we can rewrite the recursion in terms of the quantity $\mathbb{G}_n^{(k)}(h)$ itself. This is established in the following lemma.

Lemma B.3.3. *For k odd, it holds that*

$$\mathbb{G}_n^{(k)}(h) = \mathbb{G}_n^{(k-1)}(\mathcal{C}_X h) + \sum_{j=1}^{m} \left[\frac{P_X(\boldsymbol{x}_j)}{P_{n,X}^{(k-1)}(\boldsymbol{x}_j)} - 1 \right] \mathbb{G}_n^{(k-1)}(\mathcal{C}_X h \mathbf{1}_{jk}),$$

whereas for k even, it holds that

$$\mathbb{G}_n^{(k)}(h) = \mathbb{G}_n^{(k-1)}(\mathcal{C}_Z h) + \sum_{j=1}^{m} \left[\frac{P_Z(\boldsymbol{z}_j)}{P_{n,Z}^{(k-1)}(\boldsymbol{z}_j)} - 1 \right] \mathbb{G}_n^{(k-1)}(\mathcal{C}_Z h \mathbf{1}_{jk}),$$

Proof. We give the proof for k odd. By (B.24) from Proposition B.3.1 and by the definition of $\mathbb{G}_n^{(k)}(h)$, we need only show that $P(\mathcal{C}_X h \mathbf{1}_{jk}) = 0$. Indeed,

$$\mathbb{E}\left[\mathcal{C}_X h \mathbf{1}_{jk} | X\right](\boldsymbol{x}) = \begin{cases} \mathbb{E}\left[\mathcal{C}_X h | X\right](\boldsymbol{x}_j) & \text{if } x = \boldsymbol{x}_j \\ 0 & \text{if } x \neq \boldsymbol{x}_j \end{cases}.$$

But $\mathbb{E}\left[\mathcal{C}_X h | X\right](\boldsymbol{x}_j) = 0$ by definition of $\mathcal{C}_X$. Taking an expectation over P_X gives that $P(\mathcal{C}_X h \mathbf{1}_{jk}) = 0$, which implies the desired result. The proof for k even follows symmetrically. □

The higher-order term in (B.21), can be bounded using Proposition B.3.2.

Proposition B.3.2. *For any $k \geq 1$, the following holds under the event $\mathcal{S}$:*

$$\sqrt{n} \sum_{\ell=1}^{k} |V_n^{(\ell-1)}(\mathcal{C}_\ell \dots \mathcal{C}_k h)| \leq \sum_{j=1}^{m} \left(B_1 \left| \mathbb{G}_n^{(0)}(h_{1,k} \mathbf{1}_{j\ell}) \right| + B_2 \sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k} \mathbf{1}_{j\ell})| \right) + m B_2 \|h\|_\infty \sqrt{n} k(k-1)[B_1 + B_2(k+1)/3].$$

Proof. First, for any $\ell \in \{1, \dots, k\}$, recall the notation $h_{\ell,k} := \mathcal{C}_\ell \dots \mathcal{C}_k h$. By (B.23) from Proposition B.3.1 and by Lemma B.3.3, we have that for ℓ odd,

$$\sqrt{n} V_n^{(\ell-1)}(h_{\ell,k}) = \sum_{j=1}^{m} \left[\frac{P_X}{P_{n,X}^{(\ell-1)}}(\boldsymbol{x}_j) - 1 \right] \mathbb{G}_n^{(\ell-1)}(h_{\ell,k} \mathbf{1}_{j\ell}). \tag{B.28}$$

Using the statement above, we have that

$$\sqrt{n} \left| V_n^{(\ell-1)}(h_{\ell,k}) \right| \leq M_\ell \sum_{j=1}^{m} \left| \mathbb{G}_n^{(\ell-1)}(h_{\ell,k} \mathbf{1}_{j\ell}) \right|.$$

The bound above holds for ℓ even as well. Then, using the (B.23) from Proposition B.3.1

again, we have that for $\ell \geq 2$,

$$[P_n^{(\ell-1)} - P](h_{\ell,k}\mathbf{1}_{j\ell}) = [P_n^{(\ell-2)} - P](h_{\ell,k}\mathbf{1}_{j\ell}) + V_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell})$$

which implies that

$$\begin{aligned}
|\mathbb{G}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell})| &\leq |\mathbb{G}_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell})| + \sqrt{n}\,|V_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell})| \\
&\leq |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| + \sqrt{n}\,|V_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| + \ldots + \sqrt{n}\,|V_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell})| \\
&\leq |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| + M_1\sqrt{n}P_n^{(0)}(|h_{\ell,k}|\,\mathbf{1}_{j\ell}) + \ldots + M_\ell\sqrt{n}P_n^{(\ell-2)}(|h_{\ell,k}|\,\mathbf{1}_{j\ell}) \\
&\leq |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| + 2\,\|h\|_\infty\sqrt{n}\,[B_1 + B_2(\ell-1)]\,(k-\ell+1), \qquad \text{(B.29)}
\end{aligned}$$

by Lemma B.3.2 and $M_1 \leq B_1$ and $M_\ell \leq B_2$ for $\ell \geq 2$. Summing these bounds, we have that

$$\begin{aligned}
&\sqrt{n}\sum_{\ell=1}^{k}|V_n^{(\ell-1)}(h_{\ell,k})| \\
&\leq M_1\sum_{j=1}^{m}|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + \sum_{\ell=2}^{k}M_\ell\sum_{j=1}^{m}|\mathbb{G}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell})| \\
&\leq B_1\sum_{j=1}^{m}|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + B_2\sum_{\ell=2}^{k}\sum_{j=1}^{m}|\mathbb{G}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell})| \\
&\leq B_1\sum_{j=1}^{m}|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| \; + \\
&\quad B_2\sum_{\ell=2}^{k}\sum_{j=1}^{m}\left(|\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| + 2\,\|h\|_\infty\sqrt{n}\,[B_1 + B_2(\ell-1)]\,(k-\ell+1)\right) \quad \text{apply (B.29)} \\
&= \sum_{j=1}^{m}\left(B_1\,|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + B_2\sum_{\ell=2}^{k}|\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})|\right) \; + \\
&\quad 2mB_2\,\|h\|_\infty\sqrt{n}\sum_{\ell=2}^{k}[B_1 + B_2(\ell-1)]\,(k-\ell+1),
\end{aligned}$$

because $|\mathcal{X}| = m$. We sum up the last term:

$$\sum_{\ell=2}^{k} \left[B_1 + B_2(\ell-1)\right](k-\ell+1) = B_1 \sum_{\ell=1}^{k-1}(k-\ell) + B_2 \sum_{\ell=1}^{k-1} \ell(k-\ell)$$
$$= \frac{k(k-1)}{2}\left[B_1 + B_2(k+1)/3\right].$$

completing the proof. □

B.3.4 Proof of Main Results

We can now show the main result of this section: the bound on the mean squared error of the rebalanced estimator. Recall the event

$$\mathcal{S} := \{\text{Supp}(P_{n,X}) = \text{Supp}(P_X) \text{ and } \text{Supp}(P_{n,Z}) = \text{Supp}(P_Z)\} \tag{B.30}$$

as introduced in (B.19). To remind the reader of the high-level steps of the proof, we may decompose the error on the event $\mathcal{S}$ we used the estimator

$$\tilde{\boldsymbol{\theta}}_n^{(k)} := \boldsymbol{\theta}_n^{(k)} \mathbb{1}_{\mathcal{S}} + \boldsymbol{\theta}_n^{(0)} \mathbb{1}_{\mathcal{S}^c}$$

so we decompose on the event $\mathcal{S}$ to write

$$\mathbb{E}_P\left[\left(\tilde{P}_n^{(k)}(h) - P(h)\right)^2\right] = \mathbb{E}_P\left[(P_n(h) - P(h))^2 \mathbf{1}_{\mathcal{S}^c}\right] + \mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2 \mathbb{1}_{\mathcal{S}}\right]. \tag{B.31}$$

Then, we use the upcoming Proposition B.3.3 to bound the first term, which will in turn require showing that $\mathcal{S}$ occurs with high probability. As for the second term, we will apply Proposition B.3.1 and the derivation (B.21) to write

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2 \mathbb{1}_{\mathcal{S}}\right] = \mathbb{E}_P\left[T_1^2 \mathbb{1}_{\mathcal{S}}\right] + 2\mathbb{E}_P\left[T_1 T_2 \mathbb{1}_{\mathcal{S}}\right] + \mathbb{E}_P\left[T_2^2 \mathbb{1}_{\mathcal{S}}\right] \tag{B.32}$$

for

$$T_1 := [P_n^{(0)} - P](\mathcal{C}_1 \dots \mathcal{C}_k h) \text{ and } T_2 := \sum_{\ell=1}^{k} V_n^{(\ell-1)}(\mathcal{C}_\ell \dots \mathcal{C}_k h). \tag{B.33}$$

By definition, we have that $\mathbb{E}_P[T_1^2 \mathbb{1}_{\mathcal{S}}] \leq \mathbb{E}_P[T_1^2] = \sigma_k^2/n$. It then remains to bound the cross term $\mathbb{E}_P[T_1 T_2 \mathbb{1}_{\mathcal{S}}]$ and squared term $\mathbb{E}_P[T_2^2 \mathbb{1}_{\mathcal{S}}]$. This is accomplished by Lemma B.3.5 and Lemma B.3.4, respectively.

Proposition B.3.3. *It holds that $P(\mathcal{S}^c) \leq 2m(1-p_\star)^n$. Moreover, for any $\delta \in (0,1)$, we have*

$$\mathbb{E}_P\left[(P_n(h) - P(h))^2 \mathbf{1}_{\mathcal{S}^c}\right] \leq 4\|h\|_\infty^2 \min\{2m(1-p_\star)^n, \delta\} + \frac{2\log(2/\delta)}{n}\|h\|_\infty^2 2m(1-p_\star)^n.$$

Proof. Define $\mathcal{F}_X := \{\mathrm{Supp}(P_{n,X}) \neq \mathrm{Supp}(P_X)\}$ and $\mathcal{F}_Z := \{\mathrm{Supp}(P_{n,Z}) \neq \mathrm{Supp}(P_Z)\}$, so that $\mathcal{S}^c = \mathcal{F}_X \cup \mathcal{F}_Z$. We first control the probability of $\mathcal{F}_X$. Let $F_j := \{P_{n,X}(\boldsymbol{x}_j) = 0\}$ for $j \in [m]$. We then obtain $\mathcal{F}_X = \cup_{j=1}^m F_j$, which implies by the union bound that

$$P(\mathcal{F}_X) \leq \sum_{j=1}^m P(F_j) = \sum_{j=1}^m (1 - P_X(\boldsymbol{x}_j))^n \leq m(1-p_\star)^n.$$

Similarly, we have that $P(\mathcal{F}_Z) \leq m(1-p_\star)^n$ and thus $P(\mathcal{S}^c) \leq 2m(1-p_\star)^n$, which gives the first claim.

To control the expectation, consider any $\delta > 0$, and define the event

$$\mathcal{E}_\delta := \left\{ \left|P_n^{(0)}(h) - P(h)\right| \leq \sqrt{\frac{2\log(2/\delta)}{n}}\|h\|_\infty \right\}.$$

By Hoeffding's inequality, it holds that $P(\mathcal{E}_\delta) \geq 1 - \delta$. Furthermore, we get

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{\mathcal{S}^c}(P_n^{(0)}(h) - P(h))^2] &= \mathbb{E}[\mathbb{1}_{\mathcal{S}^c}\mathbb{1}_{\mathcal{E}_\delta^c}(P_n^{(0)}(h) - P(h))^2] + \mathbb{E}[\mathbb{1}_{\mathcal{S}^c}\mathbb{1}_{\mathcal{E}_\delta}(P_n^{(0)}(h) - P(h))^2] \\
&\leq 4\|h\|_\infty^2 \mathbb{E}[\mathbb{1}_{\mathcal{S}^c}\mathbb{1}_{\mathcal{E}_\delta^c}] + \frac{2\log(2/\delta)}{n}\|h\|_\infty^2 \mathbb{E}[\mathbb{1}_{\mathcal{S}^c}\mathbb{1}_{\mathcal{E}_\delta}] \\
&\leq 4\|h\|_\infty^2 \min\{P(\mathcal{S}^c), P(\mathcal{E}_\delta^c)\} + \frac{2\log(2/\delta)}{n}\|h\|_\infty^2 P(\mathcal{S}^c) \\
&\leq 4\|h\|_\infty^2 \min\{2m(1-p_\star)^n, \delta\} + \frac{2\log(2/\delta)}{n}\|h\|_\infty^2 2m(1-p_\star)^n.
\end{aligned}$$

□

In order to bound the terms appearing in (B.32), we introduce the events $\mathcal{E}_1^\delta$, $\mathcal{E}_2^\delta$, and $\mathcal{E}_3^\delta$,

defined by

$$\mathcal{E}_1^\delta := \left\{ \max\left\{ \mathrm{KL}(P_{n,X} \| P_X), \mathrm{KL}(P_{n,Z} \| P_Z) \right\} \leq \frac{1}{n} \log_2 \frac{2}{\delta} + m \frac{\log(n+1)}{n} \right\}$$

$$\mathcal{F}_\ell^\delta := \left\{ |\mathbb{G}_n^{(0)}(h_{\ell,k} \mathbf{1}_{j\ell})| \leq \sqrt{2 \log(2mk/\delta)} 2(k-\ell+1) \|h\|_\infty \right\}, \quad \ell = 1, \ldots, k, \ j = 1, \ldots, m$$

$$\mathcal{E}_2^\delta := \bigcap_{\ell=1}^{k} \mathcal{F}_\ell^\delta$$

$$\mathcal{E}_3^\delta := \left\{ |\mathbb{G}_n^{(0)}(h_{1,k})| \leq \sqrt{2 \log(2/\delta)} 2k \|h\|_\infty \right\}.$$

The events are constructed such that $\mathbb{P}(\mathcal{E}_1^\delta) \geq 1-\delta$, $\mathbb{P}(\mathcal{E}_2^\delta) \geq 1-\delta$, and $\mathbb{P}(\mathcal{E}_3^\delta) \geq 1-\delta$, as we used in the upcoming proofs of Lemma B.3.5, Lemma B.3.4, and Theorem B.3.2.

Lemma B.3.4 (Squared term bound)**.** *Let T_2 be defined as in (B.33). For any $\delta > 0$, assuming that $n \geq 2[\log_2(2/\delta) + m \log(n+1)]/p_\star^2$, we have that*

$$\mathbb{E}_P \left[T_2^2 \mathbb{1}_{\mathcal{S}} \right] \leq \frac{2 \|h\|_\infty^2 m^2 k^2}{p_\star^2} \left[\log_2(2/\delta) + m \log(n+1) \right]^{2 - \mathbb{1}\{k=1\}} \times$$
$$\left[\left(4n + \frac{k-1}{p_\star^2} \left(n + 2 + \frac{k+1}{p_\star^2} \right) \right)^2 \delta + \frac{8}{n^2} \left(\sqrt{2 \log \frac{2mk}{\delta}} (k+1) + \frac{(k-1)(k+4)}{p_\star^2} \right)^2 \right].$$

Proof. The following computations are done under the event $\mathcal{S}$. First, apply Proposition B.3.2 to write

$$|T_2| \leq \frac{1}{\sqrt{n}} \sum_{j=1}^{m} \left(B_1 |\mathbb{G}_n^{(0)}(h_{1,k} \mathbf{1}_{j\ell})| + B_2 \sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k} \mathbf{1}_{j\ell})| \right) +$$
$$m B_2 \|h\|_\infty k(k-1)[B_1 + B_2(k+1)/3]. \tag{B.34}$$

We decompose on the event $\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta$. Note that by Theorem B.3.1, we have that $\mathbb{P}(\mathcal{E}_1^\delta) \geq 1-\delta$. It follows from Hoeffding's inequality, the union bound, and boundedness of $\|h_{\ell,k} \mathbf{1}_{j\ell}\|$ by Lemma B.3.2 that $\mathbb{P}(\mathcal{E}_2^\delta) \geq 1-\delta$ As a result, $\mathbb{P}(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta) \geq 1-2\delta$.

Bound $|T_2|$ under the event $\mathcal{S}\backslash(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta)$. In this case, we apply (B.16) from Proposition B.2.5 to get $B_1 \le n$ and $B_2 \le 1/p_\star^2$, along with the universal bounds from Lemma B.3.2:

$$\frac{1}{\sqrt{n}} |\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| \le 2 \|h_{1,k}\|_\infty \le 4k \|h\|_\infty$$
$$\frac{1}{\sqrt{n}} \sum_{\ell=2}^k |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| \le 2 \sum_{\ell=2}^k \|h_{\ell,k}\|_\infty \le \sum_{\ell=2}^k 4(k-\ell+1) \|h\|_\infty = 2k(k-1) \|h\|_\infty$$

so that by plugging into (B.34),

$$|T_2| \le \|h\|_\infty mk \left[4n + \frac{k-1}{p_\star^2}\left(n + 2 + \frac{k+1}{3p_\star^2}\right)\right],$$

and in turn,

$$\mathbb{E}_P\left[T_2^2 \mathbb{1}_{\mathcal{S}\backslash(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta)}\right] \le 2 \|h\|_\infty^2 m^2 k^2 \left[4n + \frac{k-1}{p_\star^2}\left(n + 2 + \frac{k+1}{3p_\star^2}\right)\right]^2 \delta. \tag{B.35}$$

Bound $|T_2|$ under the event $\mathcal{S} \cap \mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta$. In this case, we may use that $n \ge 2[\log_2(2/\delta) + m\log(n+1)]/p_\star^2$ apply (B.17) from Proposition B.2.5 to get

$$\max\{B_1, B_2\} \le \frac{2}{p_\star}\sqrt{\frac{1}{2}\mathrm{KL}(P_{n,X}\|P_X)} \le \frac{1}{p_\star\sqrt{n}}\sqrt{2\log_2(2/\delta) + 2m\log(n+1)}$$

and the bounds based on $\mathcal{E}_2^\delta$, which give

$$|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| \le \sqrt{2\log\frac{2mk}{\delta}} 2k \|h\|_\infty$$
$$\sum_{\ell=2}^k |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| \le \sum_{\ell=2}^k \sqrt{2\log\frac{2mk}{\delta}} 2(k-\ell+1) \|h\|_\infty \le \sqrt{2\log\frac{2mk}{\delta}} k(k-1) \|h\|_\infty,$$

By plugging into (B.34),

$$|T_2| \leq \frac{2m \|h\|_\infty \sqrt{2\log(2mk/\delta)\,[2\log_2(2/\delta) + 2m\log(n+1)]}}{np_\star} k(k+1) + \tag{B.36}$$

$$\frac{m \|h\|_\infty [2\log_2(2/\delta) + 2m\log(n+1)]}{3np_\star^2} k(k-1)(k+4) \tag{B.37}$$

$$\leq \frac{4mk \|h\|_\infty [\log_2(2/\delta) + 2m\log(n+1)]^{1-\mathbb{1}\{k=1\}/2}}{np_\star^2} \times \tag{B.38}$$

$$\left[p_\star \sqrt{2\log(2mk/\delta)}(k+1) + (k-1)(k+4)\right]. \tag{B.39}$$

In turn,

$$\mathbb{E}_P\left[T_2^2 \mathbb{1}_{\mathcal{S}\setminus(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta)}\right] \leq \frac{16 \|h\|_\infty^2 m^2k^2 [\log_2(2/\delta) + m\log(n+1)]^{2-\mathbb{1}\{k=1\}}}{n^2p_\star^4} \times \left[p_\star \sqrt{2\log(2mk/\delta)}(k+1) + (k-1)(k+4)\right]^2. \tag{B.40}$$

Combining together both (B.40) and (B.35) and using that $[\log_2(2/\delta) + 2m\log(n+1)] \geq 1$, we have that

$$\mathbb{E}_P\left[T_2^2 \mathbb{1}_{\mathcal{S}}\right] \leq \frac{2\|h\|_\infty^2 m^2k^2}{p_\star^2} [\log_2(2/\delta) + m\log(n+1)]^{2-\mathbb{1}\{k=1\}} \times \left[\left(4n + \frac{k-1}{p_\star^2}\left(n+2+\frac{k+1}{p_\star^2}\right)\right)^2 \delta + \frac{8}{n^2}\left(\sqrt{2\log(2mk/\delta)}(k+1) + \frac{(k-1)(k+4)}{p_\star^2}\right)^2\right],$$

the result as desired. □

Lemma B.3.5 (Cross term bound)**.** *Let T_1 and T_2 be defined as in (B.33). For any $\delta > 0$, assuming that $n \geq 2[\log_2(2/\delta) + m\log(n+1)]/p_\star^2$, we have that*

$$\mathbb{E}_P[T_1T_2\mathbb{1}_{\mathcal{S}}] \leq \frac{2mk^2 \|h\|_\infty^2 \sqrt{2\log(2/\delta)}\,[\log_2(2/\delta) + 2m\log(n+1)]^{1-\mathbb{1}\{k=1\}/2}}{p_\star^2} \times \left[\frac{p_\star\sqrt{2\log(2mk/\delta)}(k+1) + (k-1)(k+4)}{n^{3/2}} + 6\left(4np_\star^2 + (k-1)\left(n+2+\frac{k+1}{p_\star^2}\right)\right)\delta\right],$$

Proof. The following computations are done under the event $\mathcal{S}$. First, apply Proposi-

tion B.3.2 to write

$$|T_1T_2| \leq \frac{1}{\sqrt{n}} |\mathbb{G}_n^{(0)}(h_{1,k})| \left[\frac{1}{\sqrt{n}} \sum_{j=1}^{m} \left(B_1 |\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + B_2 \sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| \right) + mB_2 \|h\|_\infty k(k-1)[B_1 + B_2(k+1)/3] \right]. \tag{B.41}$$

We decompose on the event $\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta$. Note that by Theorem B.3.1 and that $n \geq \log_2(2/\delta) + m\log(n+1)$, we have that $\mathbb{P}(\mathcal{E}_1^\delta) \geq 1-\delta$. It follows from Hoeffding's inequality and the union bound that $\mathbb{P}(\mathcal{E}_2^\delta) \geq 1-\delta$. Similarly, we also have by Hoeffding's inequality that $\mathbb{P}(\mathcal{E}_3^\delta) \geq 1-\delta$. As a result, $\mathbb{P}(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta) \geq 1-3\delta$.

Bound $|T_1T_2|$ under the event $\mathcal{S}\backslash(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta)$. In this case, we apply (B.16) from Proposition B.2.5 to get $B_1 \leq n$ and $B_2 \leq 1/p_\star^2$, along with the universal bounds from Lemma B.3.2:

$$\begin{aligned}
\frac{1}{\sqrt{n}} |\mathbb{G}_n^{(0)}(h_{1,k})| &\leq 2\|h_{1,k}\|_\infty \leq 4k\|h\|_\infty \\
\frac{1}{\sqrt{n}} |\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| &\leq 2\|h_{1,k}\|_\infty \leq 4k\|h\|_\infty \\
\frac{1}{\sqrt{n}} \sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| &\leq 2\sum_{\ell=2}^{k} \|h_{\ell,k}\|_\infty \leq \sum_{\ell=2}^{k} 4(k-\ell+1)\|h\|_\infty = 2k(k-1)\|h\|_\infty,
\end{aligned}$$

so that by plugging into (B.41),

$$|T_1T_2| \leq 4k^2 \|h\|_\infty^2 m \left[4n + \frac{k-1}{p_\star^2} \left(n + 2 + \frac{k+1}{3p_\star^2} \right) \right],$$

and in turn,

$$\mathbb{E}_P \left[T_1T_2 \mathbb{1}_{\mathcal{S}\backslash(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta)} \right] \leq \frac{12k^2 \|h\|_\infty^2 m}{p_\star^2} \left[4np_\star^2 + (k-1)\left(n + 2 + \frac{k+1}{3p_\star^2} \right) \right] \delta. \tag{B.42}$$

Bound $|T_1T_2|$ under the event $\mathcal{S} \cap \mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta$. In this case, we may use that $n \geq 2[\log_2(2/\delta) + m\log(n+1)]/p_\star^2$ apply (B.17) from Proposition B.2.5 to get

$$\max\{B_1, B_2\} \leq \frac{2}{p_\star}\sqrt{\frac{1}{2}\operatorname{KL}(P_{n,X}\|P_X)} \leq \frac{1}{\sqrt{n}}\frac{1}{p_\star}\sqrt{2\log_2(2/\delta) + 2m\log(n+1)}$$

and the bounds based on $\mathcal{E}_2^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta$ which give

$$\begin{aligned}
|\mathbb{G}_n^{(0)}(h_{1,k})| &\leq \sqrt{2\log(2/\delta)}2k\,\|h\|_\infty \\
|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| &\leq \sqrt{2\log(2mk/\delta)}2k\,\|h\|_\infty \\
\sum_{\ell=2}^{k}|\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| &\leq \sum_{\ell=2}^{k}\sqrt{2\log\frac{2mk}{\delta}}2(k-\ell+1)\,\|h\|_\infty \leq \sqrt{2\log\frac{2mk}{\delta}}k(k-1)\,\|h\|_\infty,
\end{aligned}$$

By plugging into (B.41),

$$\begin{aligned}
|T_2| &\leq \frac{m\,\|h\|_\infty\sqrt{2\log(2mk/\delta)\,[2\log_2(2/\delta)+2m\log(n+1)]}}{np_\star}k(k+1) + \\
&\quad \frac{m\,\|h\|_\infty\,[2\log_2(2/\delta)+2m\log(n+1)]}{3np_\star^2}k(k-1)(k+4) \\
&\leq \frac{mk\,\|h\|_\infty\,[\log_2(2/\delta)+2m\log(n+1)]^{1-\mathbb{1}\{k=1\}/2}}{np_\star^2}\times \\
&\quad \left[p_\star\sqrt{2\log(2mk/\delta)}(k+1)+(k-1)(k+4)\right] \\
|T_1T_2| &\leq \frac{2mk^2\,\|h\|_\infty^2\sqrt{2\log(2/\delta)}\,[\log_2(2/\delta)+2m\log(n+1)]^{1-\mathbb{1}\{k=1\}/2}}{n^{3/2}p_\star^2}\times \\
&\quad \left[p_\star\sqrt{2\log(2mk/\delta)}(k+1)+(k-1)(k+4)\right],
\end{aligned}$$

In turn,

$$\begin{aligned}
\mathbb{E}_P\left[T_2^2\mathbb{1}_{\mathcal{S}\setminus(\mathcal{E}_1^\delta\cap\mathcal{E}_2^\delta\cap\mathcal{E}_3^\delta)}\right] &\leq \frac{2mk^2\,\|h\|_\infty^2\sqrt{2\log(2/\delta)}\,[\log_2(2/\delta)+2m\log(n+1)]^{1-\mathbb{1}\{k=1\}/2}}{n^{3/2}p_\star^2}\times \\
&\quad \left[p_\star\sqrt{2\log(2mk/\delta)}(k+1)+(k-1)(k+4)\right], \qquad \text{(B.43)}
\end{aligned}$$

Combining together both (B.43) and (B.42) and using that $[\log_2(2/\delta) + 2m\log(n+1)] \geq 1$, we have that

$$\begin{aligned}
&\mathbb{E}_P\left[T_1 T_2 \mathbb{1}_{\mathcal{S}}\right] \\
&\leq \frac{2mk^2 \left\|h\right\|_\infty^2 \sqrt{2\log(2/\delta)}\left[\log_2(2/\delta) + 2m\log(n+1)\right]^{1-\mathbb{1}\{k=1\}/2}}{p_\star^2} \times \\
&\quad \left[\frac{p_\star\sqrt{2\log(2mk/\delta)}(k+1) + (k-1)(k+4)}{n^{3/2}} + 6\left(4np_\star^2 + (k-1)\left(n+2+\frac{k+1}{p_\star^2}\right)\right)\delta\right],
\end{aligned}$$

the result as desired. □

We now combine the previous results to prove Theorem B.3.2.

Theorem B.3.2. *For a sequence of rebalanced distributions* $(\tilde{P}_n^{(k)})_{k\geq 1}$*, there exists an absolute constant* $C > 0$ *such that when* $n \geq C[\log_2(2n/p_\star) + m\log(n+1)]/p_\star^2$*,*

$$\mathbb{E}_P[(\tilde{P}_n^{(k)}(h) - P(h))^2] \leq \frac{\sigma_k^2}{n} + \frac{CB}{n^{3/2}}, \tag{B.44}$$

where

$$B = \frac{\sqrt{\log(2n/p_\star)}m^2k^4\left\|h\right\|_\infty^2}{p_\star^2}\left(\log_2\frac{2n}{p_\star} + m\log(n+1)\right)^{2-\mathbb{1}\{k\}}\left(\log\frac{2mkn}{p_\star} + \frac{(k-1)^2}{p_\star^2}\right).$$

Proof. We apply the decomposition (B.31), and subsequently handle the second term using bounds on the terms in (B.32). Set $\delta = p_\star^4/n^4$. We apply Lemma B.3.4 and Lemma B.3.5 with this choice of δ, so that there exists an absolute constants $\tilde{C}$, C_1, and C_2 such that

$$\begin{aligned}
\mathbb{E}_P\left[T_1T_2\mathbb{1}_{\mathcal{S}}\right] &\leq C_1\frac{\left\|h\right\|_\infty^2 m^2k^3\sqrt{\log(2n/p_\star)}}{n^{3/2}p_\star^2}\left[\log_2(2n/p_\star) + m\log(n+1)\right]^{1-\mathbb{1}\{k=1\}/2} \times \\
&\qquad \left(\log\frac{2mnk}{p_\star} + \frac{k-1}{p_\star^2}\right) \\
\mathbb{E}_P\left[T_2^2\mathbb{1}_{\mathcal{S}}\right] &\leq C_2\frac{\left\|h\right\|_\infty^2 m^2k^4}{n^2p_\star^2}\left[\log_2(2n/p_\star) + m\log(n+1)\right]^{2-\mathbb{1}\{k=1\}} \times \\
&\qquad \left(\log\frac{2mnk}{p_\star} + \frac{(k-1)^2}{p_\star^2}\right),
\end{aligned}$$

when $n \geq \tilde{C}[\log_2(2n/p_\star) + m\log(n+1)]/p_\star^2$. This then implies that there is an absolute

constant C_3 such that

$$\begin{aligned}
&\mathbb{E}_P\left[\left(\tilde{P}_n^{(k)}(h) - P(h)\right)^2\right] \\
&\leq \mathbb{E}_P\left[\left(P_n^{(0)}(h) - P(h)\right)^2 \mathbb{1}_{\mathcal{S}^c}\right] + \frac{\sigma_k^2}{n} + \\
&\quad \frac{C_3 \left\|h\right\|_\infty^2 m^2 k^4 \sqrt{\log(2n/p_\star)}}{n^{3/2} p_\star^2} \left[\log_2 \frac{2n}{p_\star} + m \log(n+1)\right]^{2-\mathbb{1}\{k=1\}} \left(\log \frac{2mnk}{p_\star} + \frac{(k-1)^2}{p_\star^2}\right).
\end{aligned}$$

Next, we apply Proposition B.3.3 with the same choice of δ. Because $2[\log_2(2/\delta) + m\log(n+1)] \geq \log(m/\delta)$ and $-\log(1-p_\star) \geq p_\star \geq p_\star^2$, we have that $n \geq \log(\delta/m)/\log(1-p_\star)$, which implies that $m(1-p_\star)^n \leq \delta$. Combining with the display above, we have that there exists an absolute constant $C > 0$ such that

$$\begin{aligned}
\mathbb{E}_P\left[\left(\tilde{P}_n^{(k)}(h) - P(h)\right)^2\right] \leq \frac{\sigma_k^2}{n} + \frac{C \left\|h\right\|_\infty^2 m^2 k^4 \sqrt{\log(2n/p_\star)}}{n^{3/2} p_\star^2} \\
\times \left[\log_2(2/\delta) + m\log(n+1)\right]^{2-\mathbb{1}\{k=1\}} \left(\log \frac{2mnk}{p_\star} + \frac{(k-1)^2}{p_\star^2}\right),
\end{aligned}$$

which is the claimed result. □

While not shown in the main text, similar techniques to those used above can also control the bias of $\tilde{P}_n^{(k)}(h)$ as in Theorem B.3.3. Interestingly, this bias is of order $O(n^{-2})$, which confirms the intuition that even though $\tilde{P}_n^{(k)}(h)$ may be biased, the dominant term is the variance.

Theorem B.3.3. *For a sequence of rebalanced distributions $(P^{(k)})_{k\geq 1}$, there exists an absolute constant $C > 0$ such that when $n \geq C[\log_2(2n/p_\star) + m\log(n+1)]/p_\star^2$,*

$$\left|\mathbb{E}_P[\tilde{P}_n^{(k)}(h) - P(h)]\right|^2 \leq \frac{CB}{n^2}, \tag{B.45}$$

where B is as defined in Theorem B.3.2.

Proof. First, apply the decomposition (B.31) so that

$$\left|\mathbb{E}_P\left[\tilde{P}_n^{(k)}(h) - P(h)\right]\right| \leq \left|\mathbb{E}_P\left[\left(P_n(h) - P(h)\right)\mathbf{1}_{\mathcal{S}^c}\right]\right| + \left|\mathbb{E}_P\left[\left(P_n^{(k)}(h) - P(h)\right)\mathbb{1}_{\mathcal{S}}\right]\right|.$$

By using the argument of Proposition B.3.3, we have that

$$|\mathbb{E}_P\left[P_n(h) - P(h)\right]\mathbf{1}_{\mathcal{S}^c}| \leq 2\left\|h\right\|_\infty \min\left\{2m(1-p_\star)^n, \delta\right\} + \sqrt{\frac{2\log(2/\delta)}{n}}\left\|h\right\|_\infty 2m(1-p_\star)^n.$$

Then, by the recursion formula Equation (B.21), we have that

$$\begin{aligned}
&\sqrt{n}\left|\mathbb{E}_P\left[\left(P_n^{(k)}(h) - P(h)\right)\mathbb{1}_{\mathcal{S}}\right]\right| \\
&= \left|\mathbb{E}_P\left[\mathbb{G}_n^{(k)}(h)\mathbb{1}_{\mathcal{S}}\right]\right| = \left|\mathbb{E}_P\left[(1-\mathbb{1}_{\mathcal{S}^c})\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h) + \sqrt{n}\mathbb{1}_{\mathcal{S}}\sum_{\ell=1}^{k} V_n^{(\ell-1)}(\mathcal{C}_\ell\ldots\mathcal{C}_k h)\right]\right|.
\end{aligned}$$

Because $\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h)$ has zero mean, it follows that

$$\sqrt{n}\left|\mathbb{E}_P\left[\left(P_n^{(k)}(h) - P(h)\right)\mathbb{1}_{\mathcal{S}}\right]\right| \leq \left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}^c}\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h)\right]\right| + \sqrt{n}\left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}}T_2\right]\right|$$

We have by Hoeffding's inequality that $\mathbb{P}(\mathcal{E}_3^\delta) \geq 1-\delta$, and that by Lemma B.3.2 that $\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h) \leq 4k\sqrt{n}\left\|h\right\|_\infty$ universally. As a result, applying Proposition B.3.3 once again,

$$\begin{aligned}
&\left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}^c}\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h)\right]\right| \\
&\leq \left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}^c}\mathbb{1}_{\mathcal{E}_3^\delta}\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h)\right]\right| + \left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}^c}\mathbb{1}_{\mathcal{E}_3^\delta}\mathbb{G}_n^{(0)}(\mathcal{C}_1\ldots\mathcal{C}_k h)\right]\right| \\
&\leq 4k\sqrt{n}\left\|h\right\|_\infty \min\left\{2m(1-p_\star)^n, \delta\right\} + \sqrt{2\log(2/\delta)}2k\left\|h\right\|_\infty 2m(1-p_\star)^n.
\end{aligned}$$

Using a similar argument to Lemma B.3.4, we have that under $\mathcal{S}\backslash(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta)$ (which occurs with probability no more than 2δ),

$$|T_2| \leq \left\|h\right\|_\infty mk\left[4n + \frac{k-1}{p_\star^2}\left(n+2+\frac{k+1}{3p_\star^2}\right)\right],$$

and that under $\mathcal{S}\cap\mathcal{E}_1^\delta\cap\mathcal{E}_2^\delta$ (which occurs with probability at least $1-2\delta$),

$$\begin{aligned}
|T_2| \leq \frac{4mk\left\|h\right\|_\infty\left[\log_2(2/\delta) + 2m\log(n+1)\right]^{1-\mathbb{1}\{k=1\}/2}}{np_\star^2} \\
\left[p_\star\sqrt{2\log(2mk/\delta)}(k+1) + (k-1)(k+4)\right].
\end{aligned}$$

Applying the decomposition $|\mathbb{E}_P[\mathbb{1}_{\mathcal{S}} T_2]| \leq \left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}\setminus(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta)} T_2\right]\right| + \left|\mathbb{E}_P\left[\mathbb{1}_{\mathcal{S}\cap\mathcal{E}_1^\delta\cap\mathcal{E}_2^\delta} T_2\right]\right|$ and setting $\delta = \frac{p_\star^2}{n^2}$ achieves the desired result. □

B.3.5 Misspecified Marginal Distributions

As described in Section 4.6.1, we now adapt the main results to cases in which the marginal distributions (P_X, P_Y) are misspecified, in that the user is provided marginal distributions $(\hat{P}_{X,\varepsilon}, \hat{P}_{Z,\varepsilon})$ which satisfy Assumption 4.6.1. The sequence $(\hat{P}_n^{(k)})_{k\geq 1}$ is generated via (4.26).

We start by deriving a result similar to Proposition B.3.1. Since $\varepsilon < 1$, the (possibly misspecified) target marginals $\hat{P}_{X,\varepsilon}(\boldsymbol{x}) > 0$ and $\hat{P}_{Z,\varepsilon}(\boldsymbol{z}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$. Define the error term

$$\hat{V}_n^{(k-1)}(h) := \begin{cases} \sum_{\boldsymbol{x},\boldsymbol{z}} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1 \right) h(\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) & k \text{ odd} \\ \sum_{\boldsymbol{x},\boldsymbol{z}} \left(\frac{\hat{P}_{Z,\varepsilon}}{\hat{P}_{n,Z}^{(k-1)}}(\boldsymbol{z}) - 1 \right) h(\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) & k \text{ even} \end{cases} \tag{B.46}$$

as well as the empirical process-style notation

$$\hat{\mathbb{G}}_n^{(k)}(h) := \sqrt{n}\left(\hat{P}_n^{(k)}(h) - P(h)\right).$$

The format of this section will be to derive results analogous to the building blocks of the previous section. From that point, the computations from Appendix B.3.4 will achieve the desired result. For the sake of comparison to Theorem 4.3.1 we consider error terms containing ε only by their dependence on $(\varepsilon, k, n, \hat{p}_{\star,\varepsilon})$.

Intermediate Results The following result provides the necessary recursion formula, although as an inequality rather than an equality.

Proposition B.3.4. *Let $(\hat{P}_n^{(k)})_{k\geq 1}$ be a sequence computed according to (4.26). Define*

$$c^2 = \max\left\{\chi^2(\hat{P}_X \| P_X), \chi^2(\hat{P}_Z \| P_Y)\right\}.$$

These iterations are well-defined under the event $\mathcal{S}$, and for $\mathbb{G}_n^{(k)}$ defined in (B.27), it holds

that

$$\hat{\mathbb{G}}_n^{(k)}(h) = \hat{\mathbb{G}}_n^{(k-1)}(h) + \sqrt{n}\hat{V}_n^{(k-1)}(h) \tag{B.47}$$

and

$$\hat{\mathbb{G}}_n^{(k)}(h) = \hat{\mathbb{G}}_n^{(k-1)}(\mathcal{C}_k h) + \sqrt{n}\hat{V}_n^{(k-1)}(\mathcal{C}_k h) + \begin{cases} \sqrt{n}[\hat{P}_{X,\varepsilon} - P_X](\mu_X h) & \text{if } k \text{ odd} \\ \sqrt{n}[\hat{P}_{Z,\varepsilon} - P_Y](\mu_Z h) & \text{if } k \text{ even} \end{cases}. \tag{B.48}$$

Furthermore,

$$\begin{aligned} \left|\hat{\mathbb{G}}_n^{(k)}(h)\right| &\leq \left|\hat{\mathbb{G}}_n^{(k-1)}(\mathcal{C}_k h)\right| + \sqrt{n}\left|\hat{V}_n^{(k-1)}(\mathcal{C}_k h)\right| + c\,\|h\|_{\mathbf{L}^2(P)}\sqrt{n\varepsilon} \\ &= \left|\hat{\mathbb{G}}_n^{(k-1)}(\mathcal{C}_k h)\right| + \sqrt{n}\left|\hat{V}_n^{(k-1)}(\mathcal{C}_k h)\right| + O\left(\sqrt{n\varepsilon}\right). \end{aligned} \tag{B.49}$$

Proof. The proof that $\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x}) > 0$ and $\hat{P}_{n,Z}^{(k-1)}(\boldsymbol{z}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$ follows the exact same steps as in the proof of Proposition B.3.1. We take this for granted and establish the recursion.

Consider the following steps in the case that k is odd:

$$\begin{aligned} \hat{P}_n^{(k)}(h) &= \sum_{\boldsymbol{x},\boldsymbol{z}} h(\boldsymbol{x},\boldsymbol{z})\hat{P}_n^{(k)}(\boldsymbol{x},\boldsymbol{z}) = \sum_{\boldsymbol{x},\boldsymbol{z}} h(\boldsymbol{x},\boldsymbol{z})\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x})\hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\ &= \sum_{\boldsymbol{x},\boldsymbol{z}} 1 \cdot h(\boldsymbol{x},\boldsymbol{z})\hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) + \sum_{\boldsymbol{x},\boldsymbol{z}} \left[\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1\right] \cdot h(\boldsymbol{x},\boldsymbol{z})\hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\ &= \hat{P}_n^{(k-1)}(h) + \hat{V}_n^{(k-1)}(h). \end{aligned}$$

Subtracting $P(h)$ on both sides, we have that

$$[\hat{P}_n^{(k)} - P](h) = [\hat{P}_n^{(k-1)} - P](h) + \hat{V}_n^{(k-1)}(h). \tag{B.50}$$

This proves the uncentered recursion formula given in (B.47). We then show the centered

version.

$$
\begin{aligned}
&[\hat{P}_n^{(k)} - P](h) \\
&= [\hat{P}_n^{(k)} - P](\mathcal{C}_X h) + [\hat{P}_n^{(k)} - P](\mu_X h) \\
&= [\hat{P}_n^{(k)} - P](\mathcal{C}_X h) + [\hat{P}_{X,\varepsilon} - P_X](\mu_X h) \\
&= [\hat{P}_n^{(k-1)} - P](\mathcal{C}_X h) + \hat{V}_n^{(k-1)}(\mathcal{C}_X h) + [\hat{P}_{X,\varepsilon} - P_X](\mu_X h).
\end{aligned}
$$

Next, we bound the additional error term. By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
[\hat{P}_{X,\varepsilon} - P_X](\mu_X h) &\leq \left\| \tfrac{\hat{P}_{X,\varepsilon}}{P_X} - \mathbf{1} \right\|_{\mathbf{L}^2(P_X)} \cdot \|\mu_X h\|_{\mathbf{L}^2(P_X)} \\
&= \sqrt{\chi^2(\hat{P}_{X,\varepsilon} \| P_X)} \cdot \|\mu_X h\|_{\mathbf{L}^2(P_X)} \\
&\leq \sqrt{\chi^2(\hat{P}_{X,\varepsilon} \| P_X)} \cdot \|h\|_{\mathbf{L}^2(P)},
\end{aligned}
$$

as μ_X is an orthogonal projection in $\mathbf{L}^2(P)$. Using convexity of f-divergences, we have that

$$
\chi^2(\hat{P}_{X,\varepsilon} \| P_X) \leq \varepsilon \chi^2(\hat{P}_X \| P_X) + (1-\varepsilon)\chi^2(P_X \| P_X) = \varepsilon \chi^2(\hat{P}_X \| P_X).
$$

This achieves the desired result. □

Using similar ideas, we then prove an analog of Lemma B.3.3.

Lemma B.3.6. *For k odd, it holds that*

$$
\sqrt{n}\hat{V}_n^{(k-1)}(\mathcal{C}_X h) = \sum_{j=1}^{m} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}_j) - 1 \right) \hat{\mathbb{G}}_n^{(k-1)}(\mathcal{C}_X h \mathbf{1}_{jk}),
$$

whereas for k even, it holds that

$$
\sqrt{n}\hat{V}_n^{(k-1)}(\mathcal{C}_Z h) = \sum_{j=1}^{m} \left(\frac{\hat{P}_{Z,\varepsilon}}{\hat{P}_{n,Z}^{(k-1)}}(\boldsymbol{x}_j) - 1 \right) \hat{\mathbb{G}}_n^{(k-1)}(\mathcal{C}_Z h \mathbf{1}_{jk}).
$$

Proof. We give the proof for k odd. We claim that we need only show that $P(\mathcal{C}_X h \mathbf{1}_{jk}) = 0$.

This would show that

$$\begin{aligned}
\sqrt{n}\hat{V}_n^{(k-1)}(\mathcal{C}_X h) &= \sum_{\boldsymbol{x},\boldsymbol{z}} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1 \right) [\mathcal{C}_X h](\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\
&= \sqrt{n} \sum_{j=1}^{m} \sum_{\boldsymbol{x},\boldsymbol{z}} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}) - 1 \right) [\mathcal{C}_X h \mathbf{1}_{jk}](\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\
&= \sqrt{n} \sum_{j=1}^{m} \sum_{\boldsymbol{x},\boldsymbol{z}} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}_j) - 1 \right) [\mathcal{C}_X h \mathbf{1}_{jk}](\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\
&= \sum_{j=1}^{m} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}_j) - 1 \right) \sqrt{n} \sum_{\boldsymbol{x},\boldsymbol{z}} [\mathcal{C}_X h \mathbf{1}_{jk}](\boldsymbol{x},\boldsymbol{z}) \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) \\
&= \sum_{j=1}^{m} \left(\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(k-1)}}(\boldsymbol{x}_j) - 1 \right) \hat{\mathbb{G}}_n^{(k-1)}(\mathcal{C}_X h \mathbf{1}_{jk}),
\end{aligned}$$

where $P(\mathcal{C}_X h \mathbf{1}_{jk}) = 0$ is employed in the last step. Now the result follows from (B.48) in Proposition B.3.4 and the definition of $\hat{\mathbb{G}}_n^{(k)}(h)$. To prove the claim, as in Lemma B.3.3, write

$$\mathbb{E}\left[\mathcal{C}_X h \mathbf{1}_{jk} | X\right](\boldsymbol{x}) = \begin{cases} \mathbb{E}\left[\mathcal{C}_X h | X\right](\boldsymbol{x}_j) & \text{if } x = x_j \\ 0 & \text{if } x \neq x_j \end{cases}.$$

But $\mathbb{E}\left[\mathcal{C}_X h | X\right](\boldsymbol{x}_j) = 0$ by definition of $\mathcal{C}_X$. Taking an expectation over P_X gives that $P(\mathcal{C}_X h \mathbf{1}_{jk}) = 0$, which implies the desired result. The proof for k even follows symmetrically. □

For the remainder of the argument, we see that (B.49) can be unrolled so that

$$\left|\hat{\mathbb{G}}_n^{(k)}(h)\right| \leq \underbrace{\left|\mathbb{G}_n^{(0)}(\mathcal{C}_1 \dots \mathcal{C}_k h)\right|}_{\text{first-order term}} + \underbrace{\sqrt{n} \sum_{\ell=1}^{k} \left|\hat{V}_n^{(\ell-1)}(\mathcal{C}_\ell \dots \mathcal{C}_k h)\right|}_{\text{higher-order term}} + \underbrace{O(k\sqrt{n\varepsilon})}_{\text{misspecification}}, \tag{B.51}$$

where we use that $\mathbb{G}_n^{(0)} = \hat{\mathbb{G}}_n^{(0)}$.

Next, we need to bound $\left|\hat{V}_n^{(\ell-1)}(\mathcal{C}_\ell \dots \mathcal{C}_k h)\right|$, in particular accounting for the marginal violation term. We follow similar steps as in the analysis of the higher-order term in Appendix B.3.3.

Proposition B.3.5. *Assume that $P_{n,X}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$. It holds that*

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq \begin{cases} \max\{n-1, 1\} & \text{if } k = 1 \\ \max\{1/\hat{p}_{\star,\varepsilon}^2 - 1, 1\} & \text{if } k > 1. \end{cases} \tag{B.52}$$

In addition, we have that

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq \begin{cases} O\left(n\sqrt{\log \frac{1}{1-\varepsilon}}\right) + n\sqrt{\frac{1}{2}\operatorname{KL}(P_{n,X}\|P_X)} & \text{if } k = 1 \\ O\left(\frac{1}{\hat{p}_{\star,\varepsilon}^2}\sqrt{\log \frac{1}{1-\varepsilon}}\right) + \frac{1}{\hat{p}_{\star,\varepsilon}^2}\sqrt{\frac{1}{2}\operatorname{KL}(P_{n,X}\|P_X)} & \text{if } k > 1 \end{cases},$$

Moreover, when $\operatorname{KL}(P_{n,X}\|P_X) \leq \frac{\hat{p}_{\star,\varepsilon}^2}{8}$ and $\varepsilon \leq 1 - \exp\left(-\frac{\hat{p}_{\star,\varepsilon}^2}{8}\right)$, we have

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left| \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| \leq O\left(\frac{1}{\hat{p}_{\star,\varepsilon}}\sqrt{\log \frac{1}{1-\varepsilon}}\right) + \frac{2}{\hat{p}_{\star,\varepsilon}}\sqrt{\frac{1}{2}\operatorname{KL}(P_{n,X}\|P_X)}. \tag{B.53}$$

Proof. First, observe that $\hat{P}_{n,X}^{(0)}(\boldsymbol{x}) = P_{n,X}^{(0)}(\boldsymbol{x}) \geq 1/n$ under the event $\mathcal{S}$. For $k > 1$ such that k is odd, we have that for $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned} \hat{P}_{n,X}^{(k-1)}(\boldsymbol{x}) &= \sum_{\boldsymbol{z}\in\mathcal{Z}} \hat{P}_n^{(k-1)}(\boldsymbol{x},\boldsymbol{z}) = \sum_{\boldsymbol{z}\in\mathcal{Z}} \frac{\hat{P}_{Z,\varepsilon}(\boldsymbol{z})}{\hat{P}_{n,Z}^{(k-2)}(\boldsymbol{z})} \hat{P}_n^{(k-2)}(\boldsymbol{x},\boldsymbol{z}) \\ &\geq \hat{p}_{\star,\varepsilon} \sum_{\boldsymbol{z}\in\mathcal{Z}} \hat{P}_n^{(k-2)}(\boldsymbol{x},\boldsymbol{z}) = \hat{p}_{\star,\varepsilon}\hat{P}_{n,X}^{(k-2)}(\boldsymbol{x}) = \hat{p}_{\star,\varepsilon}\hat{P}_{X,\varepsilon}(\boldsymbol{x}) \geq \hat{p}_{\star,\varepsilon}^2. \end{aligned}$$

The result for k even can be proven similarly. We now prove the inequalities listed in the statement using on the lower bounds above.

Proving the first inequality. For any $\boldsymbol{x} \in \mathcal{X}$,

$$\left| \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} - 1 \right| = \max\left\{ \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} - 1, 1 - \frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})} \right\} \leq \begin{cases} \max\{n-1, 1\} & \text{if } k = 1 \\ \max\{1/\hat{p}_{\star,\varepsilon}^2 - 1, 1\} & \text{if } k > 1 \end{cases},$$

which is the desired result.

Proving the second and third inequalities. Consider an odd $k \geq 1$. By the definition

of total variation distance, it holds that

$$\max_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{P}_{X,\varepsilon}(\boldsymbol{x})-\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})\right|\leq \mathrm{TV}(\hat{P}_{n,X}^{(k-1)},\hat{P}_{X,\varepsilon}).$$

According to Pinsker's inequality, we have that $\mathrm{TV}(\hat{P}_{n,X}^{(k-1)},\hat{P}_{X,\varepsilon})\leq\sqrt{\frac{1}{2}\,\mathrm{KL}(\hat{P}_{n,X}^{(k-1)}\|\hat{P}_{X,\varepsilon})}$, and so we have that

$$\max_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{P}_{X,\varepsilon}(\boldsymbol{x})-\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})\right|\leq\sqrt{\frac{1}{2}\,\mathrm{KL}(\hat{P}_{n,X}^{(k-1)}\|\hat{P}_{X,\varepsilon})}\leq\sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X}^{(0)}\|\hat{P}_{X,\varepsilon})},$$

where the last inequality follows by the monotonicity of Sinkhorn iterations given in Proposition B.2.4. Notice that the remaining term is $\mathrm{KL}(P_{n,X}^{(0)}\|\hat{P}_{X,\varepsilon})=\mathrm{KL}(P_{n,X}\|\hat{P}_{X,\varepsilon})$, which may not decay to zero as $n\to\infty$. Because $\varepsilon<1$, write

$$\begin{aligned}
\mathrm{KL}(P_{n,X}\|\hat{P}_{X,\varepsilon}) &= \sum_{\boldsymbol{x}\in\mathcal{X}}P_{n,X}(\boldsymbol{x})\log\frac{P_{n,X}(\boldsymbol{x})}{(1-\varepsilon)P_X(\boldsymbol{x})+\varepsilon\hat{P}_X(\boldsymbol{x})}\\
&\leq \sum_{\boldsymbol{x}\in\mathcal{X}}P_{n,X}(\boldsymbol{x})\log\frac{P_{n,X}(\boldsymbol{x})}{(1-\varepsilon)P_X(\boldsymbol{x})}\\
&= \mathrm{KL}(P_{n,X}\|P_X)+\log\frac{1}{1-\varepsilon}\\
\implies \sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X}\|\hat{P}_{X,\varepsilon})} &\leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X}\|P_X)}+\sqrt{\frac{1}{2}\log\frac{1}{1-\varepsilon}}.
\end{aligned}$$

We can then apply the lower bounds

$$\max_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})}-1\right|\leq\begin{cases}n\left(\sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X}\|P_X)}+\sqrt{\frac{1}{2}\log\frac{1}{1-\varepsilon}}\right) & \text{if } k=1\\ \frac{1}{\hat{p}_{\star,\varepsilon}^2}\left(\sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X}\|P_X)}+\sqrt{\frac{1}{2}\log\frac{1}{1-\varepsilon}}\right) & \text{if } k>1\end{cases}.$$

Finally, combining the arguments above, we have that

$$\begin{aligned}
\max_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{P}_{X,\varepsilon}(\boldsymbol{x})-\hat{P}_{n,X}^{(k-1)}(\boldsymbol{x})\right| &\leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P_{n,X}\|P_X)}+\sqrt{\frac{1}{2}\log\frac{1}{1-\varepsilon}}\\
&\leq \frac{\hat{p}_{\star,\varepsilon}}{4}+\frac{\hat{p}_{\star,\varepsilon}}{4}=\frac{\hat{p}_{\star,\varepsilon}}{2},
\end{aligned}$$

where the last step invoked the assumption that

$$\mathrm{KL}(P_{n,X}\|P_X) \le \frac{\hat{p}^2_{\star,\varepsilon}}{8} \quad \text{and} \quad \varepsilon \le 1 - \exp\left(-\tfrac{\hat{p}^2_{\star,\varepsilon}}{8}\right).$$

This means that

$$\min_{\boldsymbol{x}\in\mathcal{X}} \hat{P}^{(k-1)}_{n,X}(\boldsymbol{x}) \ge \min_{\boldsymbol{x}\in\mathcal{X}} \hat{P}_{X,\varepsilon}(\boldsymbol{x}) - \max_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{P}^{(k-1)}_{n,X}(\boldsymbol{x}) - \hat{P}_{X,\varepsilon}(\boldsymbol{x})\right| \ge \frac{\hat{p}_{\star,\varepsilon}}{2}.$$

Hence,

$$\max_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}^{(k-1)}_{n,X}(\boldsymbol{x})} - 1\right| \le \frac{\max_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{P}^{(k-1)}_{n,X}(\boldsymbol{x}) - \hat{P}_{X,\varepsilon}(\boldsymbol{x})\right|}{\min_{\boldsymbol{x}\in\mathcal{X}} \hat{P}^{(k-1)}_{n,X}(\boldsymbol{x})} \le \frac{2}{\hat{p}_{\star,\varepsilon}} \sqrt{\frac{1}{2}\mathrm{KL}(P_{n,X}\|\hat{P}_{X,\varepsilon})}.$$

Now, for k even, set $k = 2t$ for $t \ge 0$. We have that

$$\max_{\boldsymbol{z}\in\mathcal{Z}} \left|\hat{P}^{(2t-1)}_{n,Z}(\boldsymbol{z}) - \hat{P}_{Z,\varepsilon}(\boldsymbol{z})\right| \le \mathrm{TV}(\hat{P}^{(2t-1)}_{n,Z}, \hat{P}_{Z,\varepsilon}) \le \sqrt{\frac{1}{2}\mathrm{KL}(\hat{P}_{Z,\varepsilon}\|\hat{P}^{(2t-1)}_{n,Z})}.$$

Invoke Proposition B.2.4 once again to achieve

$$\sqrt{\frac{1}{2}\mathrm{KL}(\hat{P}_{Z,\varepsilon}\|\hat{P}^{(2t-1)}_{n,Z})} \le \sqrt{\frac{1}{2}\mathrm{KL}(P_{n,X}\|\hat{P}_{X,\varepsilon})} \le \sqrt{\frac{1}{2}\mathrm{KL}(P_{n,X}\|P_X)} + \sqrt{\frac{1}{2}\log\frac{1}{1-\varepsilon}},$$

which completes the proof. □

Proceeding with similar steps, define the quantities

$$\hat{B}_1 := \hat{M}_1 \quad \text{and} \quad \hat{B}_2 := \max_{2\le\ell\le k} \hat{M}_\ell \quad \text{for} \quad \hat{M}_\ell := \begin{cases} \max_{\boldsymbol{x}\in\mathcal{X}} \left|\frac{\hat{P}_{X,\varepsilon}(\boldsymbol{x})}{\hat{P}^{(\ell-1)}_{n,X}(\boldsymbol{x})} - 1\right| & \ell \text{ odd} \\ \max_{\boldsymbol{z}\in\mathcal{Z}} \left|\frac{\hat{P}_{Z,\varepsilon}(\boldsymbol{z})}{\hat{P}^{(\ell-1)}_{n,Z}(\boldsymbol{z})} - 1\right| & \ell \text{ even} \end{cases}.$$

We must now establish an analog of Proposition B.3.2.

Proposition B.3.6. *For any $k \ge 1$, the following holds under the event $\mathcal{S}$:*

$$\sqrt{n}\sum_{\ell=1}^{k} \left|\hat{V}^{(\ell-1)}_n(\mathcal{C}_\ell \ldots \mathcal{C}_k h)\right| \le \sum_{j=1}^{m} \left(\hat{B}_1 \left|\mathbb{G}^{(0)}_n(h_{1,k}\mathbf{1}_{j\ell})\right| + \hat{B}_2 \sum_{\ell=2}^{k} \left|\mathbb{G}^{(0)}_n(h_{\ell,k}\mathbf{1}_{j\ell})\right|\right)$$
$$+ m\hat{B}_2 \|h\|_\infty \sqrt{n}k(k-1)[\hat{B}_1 + \hat{B}_2(k+1)/3].$$

Proof. This proof largely follows the argument of Proposition B.3.2, while accounting for the misspecified marginal error. Using again the notation $h_{\ell,k} := \mathcal{C}_\ell \dots \mathcal{C}_k h$, it follows from Lemma B.3.6 that, for odd ℓ,

$$\sqrt{n}\hat{V}_n^{(\ell-1)}(h_{\ell,k}) = \sum_{j=1}^{m} \left[\frac{\hat{P}_{X,\varepsilon}}{\hat{P}_{n,X}^{(\ell-1)}}(\boldsymbol{x}_j) - 1 \right] \hat{\mathbb{G}}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell}) \leq \hat{M}_\ell \sum_{j=1}^{m} \left| \hat{\mathbb{G}}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right|.$$

The bound above holds for ℓ even as well. Then, using (B.47) from Proposition B.3.4 along with the triangle inequality, we have that for $\ell \geq 2$,

$$\left| [\hat{P}_n^{(\ell-1)} - P](h_{\ell,k}\mathbf{1}_{j\ell}) \right| \leq \left| \hat{P}_n^{(\ell-2)} - P](h_{\ell,k}\mathbf{1}_{j\ell}) \right| + \left| \hat{V}_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right|$$

which implies that

$$\begin{aligned}
&\left| \hat{\mathbb{G}}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| && \text{(B.54)} \\
&\leq \left| \hat{\mathbb{G}}_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| + \sqrt{n} \left| \hat{V}_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| \\
&\leq \left| \hat{\mathbb{G}}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| + \sqrt{n} \left| \hat{V}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| + \dots + \sqrt{n} \left| \hat{V}_n^{(\ell-2)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| \\
&\leq \left| \hat{\mathbb{G}}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| + \hat{M}_1\sqrt{n}\hat{P}_n^{(0)}(|h_{\ell,k}|\,\mathbf{1}_{j\ell}) + \dots + \hat{M}_\ell\sqrt{n}\hat{P}_n^{(\ell-2)}(|h_{\ell,k}|\,\mathbf{1}_{j\ell}) \\
&\leq \left| \hat{\mathbb{G}}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell}) \right| + 2\,\|h\|_\infty \sqrt{n} \left[\hat{B}_1 + \hat{B}_2(\ell-1) \right] (k-\ell+1), && \text{(B.55)}
\end{aligned}$$

by Lemma B.3.2 and $\hat{M}_1 \leq \hat{B}_1$ and $\hat{M}_\ell \leq \hat{B}_2$ for $\ell \geq 2$. The bound above holds trivially for

$\ell = 1$. Summing these bounds over ℓ and j, we have that

$$
\begin{aligned}
&\sqrt{n}\sum_{\ell=1}^{k}\left|\hat{V}_n^{(\ell-1)}(h_{\ell,k})\right| \\
&\leq \hat{M}_1\sum_{j=1}^{m}|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + \sum_{\ell=2}^{k}\hat{M}_\ell\sum_{j=1}^{m}\left|\hat{\mathbb{G}}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell})\right| \\
&\leq \hat{B}_1\sum_{j=1}^{m}|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + \hat{B}_2\sum_{\ell=2}^{k}\sum_{j=1}^{m}\left|\hat{\mathbb{G}}_n^{(\ell-1)}(h_{\ell,k}\mathbf{1}_{j\ell})\right| \\
&\leq \hat{B}_1\sum_{j=1}^{m}|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| \\
&\quad + \hat{B}_2\sum_{\ell=2}^{k}\sum_{j=1}^{m}\left(\left|\hat{\mathbb{G}}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})\right| + 2\|h\|_\infty\sqrt{n}\left[\hat{B}_1 + \hat{B}_2(\ell-1)\right](k-\ell+1)\right) \quad \text{apply (B.55)} \\
&= \sum_{j=1}^{m}\left(\hat{B}_1|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| + \hat{B}_2\sum_{\ell=2}^{k}|\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})|\right) \\
&\quad + 2m\hat{B}_2\|h\|_\infty\sqrt{n}\sum_{\ell=2}^{k}\left[\hat{B}_1 + \hat{B}_2(\ell-1)\right](k-\ell+1),
\end{aligned}
$$

because $|\mathcal{X}| = m$. We sum up the last term:

$$
\begin{aligned}
\sum_{\ell=2}^{k}\left[\hat{B}_1 + \hat{B}_2(\ell-1)\right](k-\ell+1) &= \hat{B}_1\sum_{\ell=1}^{k-1}(k-\ell) + \hat{B}_2\sum_{\ell=1}^{k-1}\ell(k-\ell) \\
&= \frac{k(k-1)}{2}\left[\hat{B}_1 + \hat{B}_2(k+1)/3\right],
\end{aligned}
$$

which completes the proof. □

Mean Squared Error Bound Ultimately, we wish to construct an upper bound for

$$
\mathbb{E}_P\left[\left(\hat{P}_n^{(k)}(h) - P(h)\right)^2\mathbb{1}_{\mathcal{S}}\right] + \mathbb{E}_P\left[\left(P_n(h) - P(h)\right)^2\mathbb{1}_{\mathcal{S}^c}\right], \tag{B.56}
$$

as the method returns $P_n(h)$ when $\mathcal{S}$ is not satisfied. The first term will be controlled by intermediate tools developed above. The second term that includes $\mathcal{S}^c$ is no different from the one analyzed in Proposition B.3.3. We handle the second term first. Recall from

Proposition B.3.3 that for any $\delta \in (0,1)$,

$$\mathbb{E}_P\left[(P_n(h) - P(h))^2 \mathbf{1}_{\mathcal{S}^c}\right] \leq$$
$$4\|h\|_\infty^2 \min\left\{2m(1-p_\star)^n, \delta\right\} + \frac{2\log(2/\delta)}{n}\|h\|_\infty^2 2m(1-p_\star)^n. \tag{B.57}$$

Repeat the argument from the proof of Theorem B.3.2: because $2[\log_2(2/\delta) + m\log(n+1)] \geq \log(m/\delta)$ and $-\log(1-p_\star) \geq p_\star \geq p_\star^2$, we have that

$$n \geq 2[\log_2(2/\delta) + m\log(n+1)]/p_\star^2 \implies n \geq \log(\delta/m)/\log(1-p_\star). \tag{B.58}$$

This in turn implies that $m(1-p_\star)^n \leq \delta$, and gives as a condition on the sample size n. Further in the analysis, we will set $\delta = (\hat{p}_{\star,\varepsilon}/n)^4$, so right-hand side of (B.57) can then be upper bounded further, resulting in

$$\mathbb{E}_P\left[(P_n(h) - P(h))^2 \mathbf{1}_{\mathcal{S}^c}\right] \leq 4\|h\|_\infty^2 \delta\left(2 + \frac{\log(2/\delta)}{n}\right) = \tilde{O}\left(\tfrac{\hat{p}_{\star,\varepsilon}^4}{n^4}\right),$$

a higher-order term compared to other components of the bound.

Next, we must control the left-hand side of (B.56). We perform the decomposition based on (B.51):

$$\mathbb{E}_P\left[\left(\hat{P}_n^{(k)}(h) - P(h)\right)^2 \mathbb{1}_{\mathcal{S}}\right]$$
$$\leq \mathbb{E}_P\left[T_1^2 \mathbb{1}_{\mathcal{S}}\right] + 2\mathbb{E}_P\left[\left|T_1\hat{T}_2\right| \mathbb{1}_{\mathcal{S}}\right] + \mathbb{E}_P\left[\hat{T}_2^2 \mathbb{1}_{\mathcal{S}}\right] \tag{B.59}$$
$$+ O(k\sqrt{\varepsilon}) \cdot \mathbb{E}_P\left[\left(|T_1| + |\hat{T}_2|\right) \mathbb{1}_{\mathcal{S}}\right] + O(k^2\varepsilon) \tag{B.60}$$

for

$$T_1 := [P_n - P](\mathcal{C}_1 \ldots \mathcal{C}_k h) \text{ and } \hat{T}_2 := \sum_{\ell=1}^{k} \left|\hat{V}_n^{(\ell-1)}(\mathcal{C}_\ell \ldots \mathcal{C}_k h)\right|. \tag{B.61}$$

Recall the events $\mathcal{E}_1^\delta$ and $\mathcal{E}_2^\delta$ and $\mathcal{E}_3^\delta$ from Appendix B.3.4. To perform this computation efficiently, we will split the bounds on each term into two components. In particular, we will show that

- Under the event $\mathcal{S} \cap \mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta : \left|\hat{T}_2\right| \leq \mathcal{T}_2 + E_2$,
- Under the event $\mathcal{S} \backslash (\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta) : \left|\hat{T}_2\right| \leq \mathcal{T}_2^c + E_2^c$,
- Under the event $\mathcal{S} \cap \mathcal{E}_3^\delta : |T_1| \leq \mathcal{T}_1$,
- Under the event $\mathcal{S} \backslash \mathcal{E}_3^\delta : |T_1| \leq \mathcal{T}_1^c$,

where any term denoted with "E" will represent all error terms that include ε and will be written in big-O notation. There are no errors for the bounds on T_1, as this term does not depend on the misspecified marginals. The idea is that for the "$\mathcal{T}_2$" terms we may reuse the bounds derived in Appendix B.3.4 by simply replacing $p_\star$ with $\hat{p}_{\star,\varepsilon}$. This is due to the fact that the dependence of the analogous terms from Appendix B.3.4 depend on $p_\star$ only through Proposition B.2.5; similarly, the corresponding terms in this section depend on $\hat{p}_{\star,\varepsilon}$ through Proposition B.3.5. We return to the terms in (B.59) and (B.60).

Decomposing on $\mathcal{E}_3^\delta$ will result in a bound of the form

$$O(k\sqrt{\varepsilon}) \cdot \mathbb{E}_P\left[|T_1| \mathbb{1}_{\mathcal{S}}\right] \leq O(k\sqrt{\varepsilon}) \cdot \left(\delta \mathcal{T}_1^c + \mathcal{T}_1\right).$$

Decomposing on $\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta$ will result in a bounds of the form

$$\begin{aligned} \mathbb{E}_P\left[\hat{T}_2^2 \mathbb{1}_{\mathcal{S}}\right] &\leq 2\delta(\mathcal{T}_2^c)^2 + \mathcal{T}_2^2 + \tilde{O}\left(\delta\left((E_2^c)^2 + E_2^c \mathcal{T}_2^c\right) + \left(E_2^2 + E_2 \mathcal{T}_2\right)\right) \\ O(k\sqrt{\varepsilon}) \cdot \mathbb{E}_P\left[|\hat{T}_2| \mathbb{1}_{\mathcal{S}}\right] &\leq O(k\sqrt{\varepsilon}) \cdot \left(\delta\left(\mathcal{T}_2^c + E_2^c\right) + \mathcal{T}_2 + E_2\right). \end{aligned}$$

Finally, decomposing on $\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta \cap \mathcal{E}_3^\delta$ will result in a bound of the form

$$\mathbb{E}_P\left[\left|T_1 \hat{T}_2\right| \mathbb{1}_{\mathcal{S}}\right] \leq 3\delta \mathcal{T}_1^c \mathcal{T}_2^c + \mathcal{T}_1 \mathcal{T}_2 + \tilde{O}\left(\delta \mathcal{T}_1^c E_2^c + \mathcal{T}_1 E_2\right).$$

The leading terms $2\delta(\mathcal{T}_2^c)^2 + \mathcal{T}_2^2$ and $3\delta \mathcal{T}_1^c \mathcal{T}_2^c + \mathcal{T}_1 \mathcal{T}_2$ from both bounds should have the exact same form as the terms in Lemma B.3.4 and Lemma B.3.5, with $p_\star$ replaced by $\hat{p}_{\star,\varepsilon}$, thus retaining the same dependence on (n, k). By setting $\delta = \hat{p}_{\star,\varepsilon}^4 / n^4$, we will achieve a similar

result to Theorem B.3.2, i.e., that

$$\begin{aligned}
&\mathbb{E}_P\left[\left(\hat{P}_n^{(k)}(h) - P(h)\right)^2 \mathbb{1}_{\mathcal{S}}\right] \\
&\leq \frac{\sigma_k^2}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right) \\
&+ \tilde{O}\Big((\hat{p}_{\star,\varepsilon}/n)^4 \left(E_2^c(E_2^c + \mathcal{T}_2^c)\right) + E_2\left(E_2 + \mathcal{T}_2\right) + (\hat{p}_{\star,\varepsilon}/n)^4 \mathcal{T}_1^c E_2^c + \mathcal{T}_1 E_2\Big). \qquad \text{(B.62)} \\
&+ \tilde{O}\left(k\sqrt{\varepsilon}\left((\hat{p}_{\star,\varepsilon}/n)^4 \mathcal{T}_1^c + \mathcal{T}_1 + (\hat{p}_{\star,\varepsilon}/n)^4\left(\mathcal{T}_2^c + E_2^c\right) + \mathcal{T}_2 + E_2\right) + k^2\varepsilon\right). \qquad \text{(B.63)}
\end{aligned}$$

It remains to quantify the $\tilde{O}$ terms by computing the order of the 6 constants $(\mathcal{T}_2, E_2, \mathcal{T}_2^c, E_2^c, \mathcal{T}_1, \mathcal{T}_1^c)$. We follow similar steps to Lemma B.3.4 and Lemma B.3.5 to achieve this.

Lemma B.3.7. *For $\delta = (\hat{p}_{\star,\varepsilon}/n)^4$, assume that $n \geq 8[\log_2(2/\delta) + m\log(n+1)]/\hat{p}_{\star,\varepsilon}^2$ and $\varepsilon \leq 1 - \exp\left(-\frac{\hat{p}_{\star,\varepsilon}^2}{8}\right)$. Then, it holds that*

$$\begin{aligned}
\mathcal{T}_2^c &= \tilde{O}\left(\tfrac{k^2}{\hat{p}_{\star,\varepsilon}^2}\left(n + \tfrac{k}{\hat{p}_{\star,\varepsilon}^2}\right)\right), \quad E_2^c = 0 \\
\mathcal{T}_2 &= \tilde{O}\left(\frac{k^3}{n\hat{p}_{\star,\varepsilon}^2}\right), \quad E_2 = \tilde{O}\left(\tfrac{k^3}{\hat{p}_{\star,\varepsilon}^2}\left(\sqrt{\tfrac{1}{n}\log\tfrac{1}{1-\varepsilon}} + \log\tfrac{1}{1-\varepsilon}\right)\right).
\end{aligned}$$

Proof. The following computations are done under the event $\mathcal{S}$. First, apply Proposition B.3.6 to write

$$\begin{aligned}
\sqrt{n}\left|\hat{T}_2\right| \leq \sum_{j=1}^{m}\left(\hat{B}_1\left|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})\right| + \hat{B}_2\sum_{\ell=2}^{k}\left|\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})\right|\right) \\
+ m\hat{B}_2\left\|h\right\|_\infty k(k-1)[\hat{B}_1 + \hat{B}_2(k+1)/3]. \qquad \text{(B.64)}
\end{aligned}$$

We decompose on the event $\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta$.

Bound $|T_2|$ under the event $\mathcal{S}\backslash(\mathcal{E}_1^\delta \cap \mathcal{E}_2^\delta)$. In this case, we apply (B.52) from Proposition B.3.5 to get $\hat{B}_1 \le n$ and $\hat{B}_2 \le 1/\hat{p}_{\star,\varepsilon}^2$, along with the universal bounds from Lemma B.3.2:

$$\frac{1}{\sqrt{n}} |\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| \le 2 \|h_{1,k}\|_\infty \le 4k \|h\|_\infty$$
$$\frac{1}{\sqrt{n}} \sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| \le 2 \sum_{\ell=2}^{k} \|h_{\ell,k}\|_\infty \le \sum_{\ell=2}^{k} 4(k-\ell+1) \|h\|_\infty = 2k(k-1) \|h\|_\infty$$

so that by plugging into (B.64),

$$\left|\hat{T}_2\right| \le \underbrace{\|h\|_\infty mk \left[4n + \frac{k-1}{\hat{p}_{\star,\varepsilon}^2}\left(n+2+\frac{k+1}{3\hat{p}_{\star,\varepsilon}^2}\right)\right]}_{\mathcal{T}_2^c} + \underbrace{0}_{E_2^c}.$$

Bound $|T_2|$ under the event $\mathcal{S}\cap\mathcal{E}_1^\delta\cap\mathcal{E}_2^\delta$. In this case, we may use that $n \ge 8/\hat{p}_{\star,\varepsilon}^2$ (because $[\log_2(2/\delta) + m\log(n+1)] \ge 1$ for $\delta \in (0,1)$) and apply (B.53) from Proposition B.3.5 to get

$$\max\left\{\hat{B}_1, \hat{B}_2\right\} \le O\left(\tfrac{1}{\hat{p}_{\star,\varepsilon}}\sqrt{\log\tfrac{1}{1-\varepsilon}}\right) + \frac{2}{\hat{p}_{\star,\varepsilon}}\sqrt{\frac{2\log_2(2/\delta) + 2m\log(n+1)}{2n}}$$

The bounds based on $\mathcal{E}_2^\delta$ give

$$|\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| \le \sqrt{2\log\frac{2mk}{\delta}} 2k \|h\|_\infty$$
$$\sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| \le \sum_{\ell=2}^{k} \sqrt{2\log\frac{2mk}{\delta}} 2(k-\ell+1) \|h\|_\infty \le \sqrt{2\log\frac{2mk}{\delta}} k(k-1) \|h\|_\infty .$$

By plugging into (B.64), we can reuse the steps in the bound from (B.39) (for all terms without ε) to write

$$\left|\hat{T}_2\right| \le \frac{4mk \|h\|_\infty [\log_2(2/\delta) + 2m\log(n+1)]^{1-\mathbb{1}\{k=1\}/2}}{n\hat{p}_{\star,\varepsilon}^2} \times$$
$$\left[\hat{p}_{\star,\varepsilon}\sqrt{2\log(2mk/\delta)}(k+1) + (k-1)(k+4)\right] + E_2,$$

so that

$$\mathcal{T}_2 = \frac{4mk\,\|h\|_\infty \left[\log_2(2/\delta) + 2m\log(n+1)\right]^{1-\mathbb{1}\{k=1\}/2}}{n\hat{p}_{\star,\varepsilon}^2} \times \left[\hat{p}_{\star,\varepsilon}\sqrt{2\log\left(2mk/\delta\right)}(k+1) + (k-1)(k+4)\right].$$

We compute E_2 by using that

$$\begin{aligned} \max\left\{\hat{B}_1, \hat{B}_2\right\} &\leq O\left(\tfrac{1}{\hat{p}_{\star,\varepsilon}}\sqrt{\log\tfrac{1}{1-\varepsilon}}\right) + \tilde{O}\left(\tfrac{1}{\hat{p}_{\star,\varepsilon}\sqrt{n}}\right) \\ |\mathbb{G}_n^{(0)}(h_{1,k}\mathbf{1}_{j\ell})| &\leq \tilde{O}\left(k\right) \\ \sum_{\ell=2}^{k} |\mathbb{G}_n^{(0)}(h_{\ell,k}\mathbf{1}_{j\ell})| &\leq \tilde{O}\left(k^2\right), \end{aligned}$$

which gives

$$E_2 = \tilde{O}\left(\tfrac{k^3}{\hat{p}_{\star,\varepsilon}^2}\left(\sqrt{\tfrac{1}{n}\log\tfrac{1}{1-\varepsilon}} + \log\tfrac{1}{1-\varepsilon}\right)\right).$$

□

We now make the corresponding argument for the term T_1.

Lemma B.3.8. *For $\delta = (\hat{p}_{\star,\varepsilon}/n)^4$, it holds that*

$$\mathcal{T}_1^c = \tilde{O}(k), \quad \mathcal{T}_1 = \tilde{O}\left(\frac{k}{\sqrt{n}}\right).$$

Proof. The following computations are done under the event $\mathcal{S}$.

Bound $|T_1|$ under the event $\mathcal{S}\backslash\mathcal{E}_3^\delta$. Here we simply apply a universal bound on the empirical process term:

$$\frac{1}{\sqrt{n}}\,|\mathbb{G}_n^{(0)}(h_{1,k})| \leq 2\,\|h_{1,k}\|_\infty \leq 4k\,\|h\|_\infty,$$

so that $\mathcal{T}_1^c = 4k\,\|h\|_\infty$

Bound $|T_1|$ under the event $\mathcal{S} \cap \mathcal{E}_3^\delta$. Now, we may use the definition of the event $\mathcal{E}_3^\delta$ to achieve

$$\frac{1}{\sqrt{n}} |\mathbb{G}_n^{(0)}(h_{1,k})| \leq \sqrt{\frac{2\log(2/\delta)}{n}} 2k \|h\|_\infty = \mathcal{T}_1.$$

□

Knowing that $E_2^c = 0$, we simplify (B.62) and (B.62) to read

$$\tilde{O}\Big(E_2 \left(E_2 + \mathcal{T}_2 + \mathcal{T}_1\right)\Big)$$
$$\tilde{O}\left(k\sqrt{\varepsilon}\left((\hat{p}_{\star,\varepsilon}/n)^4 \mathcal{T}_1^c + \mathcal{T}_1 + (\hat{p}_{\star,\varepsilon}/n)^4 \mathcal{T}_2^c + \mathcal{T}_2 + E_2\right) + k^2\varepsilon\right).$$

We now combine the bounds from the previous two lemmas to compute (B.62) and (B.63) to achieve the main result, Theorem 4.6.1.

B.4 Zero-Shot Prediction

This appendix section contains basic definitions and identities used in Section 4.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space. The following definition allows for the discussion of conditional dependence measures based on regular conditional distributions.

Definition B.4.1. Consider random variables $(U, V) : \Omega \to \mathcal{U} \times \mathcal{V}$. Let $\mathcal{B}(\mathcal{U})$ denote the Borel σ-algebra on $\mathcal{U}$. A map: $\mu : \mathcal{V} \times \mathcal{B}(\mathcal{U}) :\to [0,1]$ is called a *regular conditional distribution (r.c.d.)* if the following two properties hold:

1. For each $A \in \mathcal{B}(\mathcal{U})$ and $\boldsymbol{v} \in \mathcal{V}$, it holds that

$$\mu(\boldsymbol{v}, A) = \mathbb{E}_{P_{U,V}}\left[\mathbb{1}_A(U)|V\right](\boldsymbol{v}).$$

2. For P_V-almost every $\boldsymbol{v} \in \mathcal{V}$, $\mu(\boldsymbol{v}, \cdot)$ is a probability measure on $\mathcal{B}(\mathcal{U})$.

For a joint probability measure $P_{X,Z}$, the Radon-Nikodym derivative $\mathsf{R} = \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}(P_X \otimes P_Z)}$ is useful for converting conditional expectation computations into marginal expectation com-

putations. The following identity is referenced by Buja [1990, Section 3] and Dytso et al. [2023, Lemma 1, Eq. (14)]. We provide a self-contained proof below.

Lemma B.4.1. *Assume that $P_{X,Z} \ll P_X \otimes P_Z$, in which case there exists a Radon-Nikodym derivative $\mathsf{R} = \frac{dP_{X,Z}}{d(P_X \otimes P_Z)}$. Then, for all $g \in \mathbf{L}^2(P_Z)$ and $h \in \mathbf{L}^2(P_X)$, it holds that*

$$\mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) = \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x}, Z)\right] \text{ for } P_X\text{-almost all } \boldsymbol{x} \in \mathcal{X},$$
$$\mathbb{E}_{P_{X,Z}}\left[h(X)|Z\right](\boldsymbol{z}) = \mathbb{E}_{P_X}\left[h(X)\mathsf{R}(X, \boldsymbol{z})\right] \text{ for } P_Z\text{-almost all } \boldsymbol{z} \in \mathcal{Z}.$$

Proof. We prove the first identity, whereas the second follows by a symmetric argument. To confirm that the two functions are equal almost surely, it is sufficient to prove that for any measurable set $A \in \sigma(X)$ (the σ-algebra generated by X) the relation

$$\int_A \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x})\, \mathrm{d}P_X(\boldsymbol{x}) = \int_A \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x}, Z)\right]\, \mathrm{d}P_X(\boldsymbol{x}). \tag{B.65}$$

By the definition of conditional expectation, we have that

$$\begin{aligned}\int_A \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x})\, \mathrm{d}P_X(\boldsymbol{x}) &= \int_{\mathcal{X}} \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) \mathbb{1}_A(\boldsymbol{x})\, \mathrm{d}P_X(\boldsymbol{x}) \\ &= \mathbb{E}_{P_{X,Z}}\left[g(Z)\mathbb{1}_A(X)\right] \\ &= \mathbb{E}_{P_X \otimes P_Z}\left[g(Z)\mathbb{1}_A(X)\mathsf{R}(X, Z)\right],\end{aligned}$$

where the last step follows from the Radon-Nikodym theorem [Schilling, 2017, Theorem 20.2]. Next, we compute the expectation, taken under the product measure, using Fubini's theorem [Schilling, 2017, Corollary 14.9]. That is,

$$\begin{aligned}\int_A \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x})\, \mathrm{d}P_X(\boldsymbol{x}) &= \mathbb{E}_{P_X \otimes P_Z}\left[g(Z)\mathbb{1}_A(X)\mathsf{R}(X, Z)\right] \\ &= \int_A \left(\int_{\mathcal{Z}} g(\boldsymbol{z})\mathsf{R}(\boldsymbol{x}, \boldsymbol{z})\, \mathrm{d}P_Z(\boldsymbol{z})\right) \mathrm{d}P_X(\boldsymbol{x}) \\ &= \int_A \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x}, Z)\right]\, \mathrm{d}P_X(\boldsymbol{x}).\end{aligned}$$

This achieves (B.65) and completes the proof. □

B.5 Experimental Details

We provide the full details of the experimental results from Section 4.5.

B.5.1 Datasets

Pre-Training Data For the balancing-based experiments, the pre-training data was taken from the public ImageNet-Captions dataset [Fang et al., 2023]. We subset the dataset by selecting the 250 classes that were most frequent in the dataset, resulting in 174,594 images and associated Flickr captions. The exact images used and their associated captions are given in the GitHub repo https://github.com/ronakdm/balancing.

Evaluation Data We perform zero-shot classification with various image classification and image-caption datasets. For the balanced pre-training experiments, we used the default prompt templates for classification from the CLIP Benchmark repo. Other, customized prompting strategies used in our experiments are described at the end of this appendix. The datasets (test splits) used were:

- **CIFAR-10:** 10,000 colored natural images labeled with one of 10 classes.
- **CIFAR-100:** 10,000 colored natural images labeled with one of 100 classes.
- **STL-10:** 80,000 colored natural images labeled with one of 10 classes.
- **MS-COCO:** 41,000 colored natural images with associated captions.
- **Flickr8k:** 8,000 colored natural images with associated captions.
- **Rendered SST2:** 1,821 images of typed natural language with sentiment label (2 classes).
- **VOC2007:** 4,952 colored natural images labeled with one of 20 classes.

- **FGVC Aircraft:** 34,000 colored natural images labeled with one of 102 classes.

Evaluation scripts using the various embedding models (described below) are provided.

B.5.2 Model Specification and Hyperparameters

CLIP Architectures First, we specify which OpenCLIP models and pre-training sets were used. These models were chosen due to their range of top-1 zero-shot accuracies on the ImageNet-1k benchmark (as shown below). As opposed to already highly performant models ($\geq$50% on ImageNet-1k), these models benefitted more from optimized prompting techniques in our initial experiments.

Model	OpenCLIP Model Tag	Pre-Training Set Tag	ImageNet-1k Top-1 Acc.
ResNet-50	`RN50`	`yfcc15m`	28.11%
NLLB-CLIP	`nllb-clip-base`	`v1`	33.51%
ViT-B/32	`ViT-B-32`	`datacomp_m_s128m_b4k`	32.81%

Optimizer For optimization, models were trained with stochastic gradient descent (SGD) with the learning rate tuned along the grid $\{1^{-3}, 3^{-3}, 1^{-2}, 3^{-2}, 1^{-1}\}$ and a fixed weight decay parameter of 0.01. Momentum-variants such as Adam [Kingma and Ba, 2015] were not used to isolate the effect of variance reduction as described in the balancing example of Section 4.5.

B.5.3 Compute Environment

Experiments were run on a CPU/GPU workstation with 12 virtual cores, 126G of memory, and four NVIDIA TITAN Xp GPUs with 12G memory each. The code was written in Python 3 and we use PyTorch for automatic differentiation. The OpenCLIP and CLIP Benchmark repositories were used for zero-shot evaluation.

Prompt-Generating Model We employed the `meta-llama/Llama-3.2-1B-Instruct` model publicly available on HuggingFace. For the purpose of generation, we used a **top-p** hyperparameter of **0.9** and **temperature** hyperparameter of **0.99** for more diverse responses. Meta-prompting was based on the following instructions per dataset, which are slight variations of those used in Pratt et al. [2023]:

- **Describable Textures Dataset (DTD):**
  - *"What does __ material look like?"*,
  - *"What does a __ surface look like?"*,
  - *"What does a __ texture look like?"*,
  - *"What does a __ object look like?"*,
  - *"What does a __ pattern look like?"*
- **Flowers 102:**
  - *"Describe how to identify a(n) __, a type of flower."*,
  - *"What does a(n) __ flower looks like?"*
- **FGVC Aircraft:**
  - *"Describe a(n) __ aircraft."*,
  - *"Describe the __ aircraft."*
- **SUN397:**
  - *"Describe what a(n) __ looks like."*,
  - *"How can you identify a(n) __?"*,
  - *"Describe a photo of a(n) __."*,

  - *"Describe the scene of a(n) ___."*

- **ImageNet-1k:**

  - *"Describe what a(n) ___ looks like."*,
  - *"How can you identify a(n) ___?"*,
  - *"What does a(n) ___ look like?"*,
  - *"Describe an image from the Internet of a(n) ___."*,
  - *"Write a caption of an image of a(n) ___."*

The following additional instruction was appended for better-formatted responses: *"Please format your response as one that contains only lower case letters and no special characters (including new lines, bold, and any markdown artifacts) other than a period ('.') or commas (','). The response should be a single sentence ending in a period that is directed toward the final instruction in this message. Your sentence should be a minimum of three words and a maximum of thirty."*.

Our reproducibility effort includes not only the full list of all 164,400 prompts generated from LlaMA 3, but the subset of prompts used for each class and each seed used to generate the figures in Section 4.5.