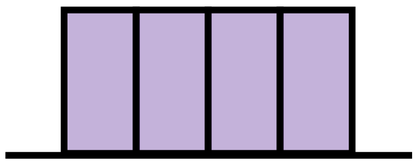
Statistical Foundations of Foundation Modeling

Final Examination
June 02, 2025

Ronak Mehta

Data Generating Distribution

P



Empirical distribution of training examples.

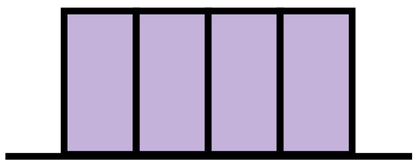
Data Generating Distribution

P

Distribution Shift

Evaluation Distribution

Q

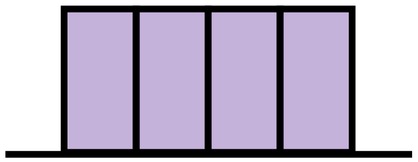


Empirical distribution of training examples.

Data Generating Distribution P

Distribution Shift

Evaluation Distribution Q



Empirical distribution of training examples.

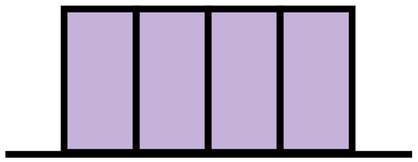
INFO

Estimate a parameter with good performance under Q, given data from P.

Data Generating Distribution P

Distribution Shift

Evaluation Distribution Q



Empirical distribution of training examples.

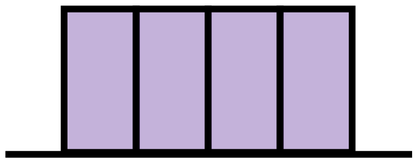
INFO

Estimate a parameter with good performance under Q, given data from P.

Data Generating Distribution P

Distribution Shift

Evaluation Distribution Q



Empirical distribution of training examples.

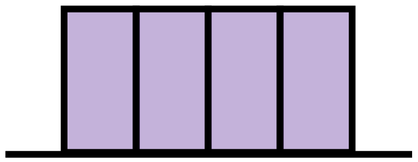
INFO

Estimate a parameter with good performance under Q, given data from P.

Data Generating Distribution P

Distribution Shift

Evaluation Distribution Q



Empirical distribution of training examples.

INFO

Estimate a parameter with good performance under Q, given data from P.

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \underbrace{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle} - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

General Exam: Distributionally Robust Optimization

$\mathcal{X} = \mathbb{R}^d$	Model parameters
$\mathcal{Y} = \Delta^{n-1}$	Weights on training examples
$f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x}))$	Losses on training examples

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \underbrace{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})}$$

General Exam: Distributionally Robust Optimization

$\mathcal{X} = \mathbb{R}^d$	Model parameters
$\mathcal{Y} = \Delta^{n-1}$	Weights on training examples
$f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x}))$	Losses on training examples
$\phi(\boldsymbol{x}) = \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$	Regularizer
$\psi(\boldsymbol{y}) = D_{\mathrm{KL}}(\boldsymbol{y}\|\mathbf{1}/n)$	Shift penalty

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

Example: Minimization with Functional Constraints

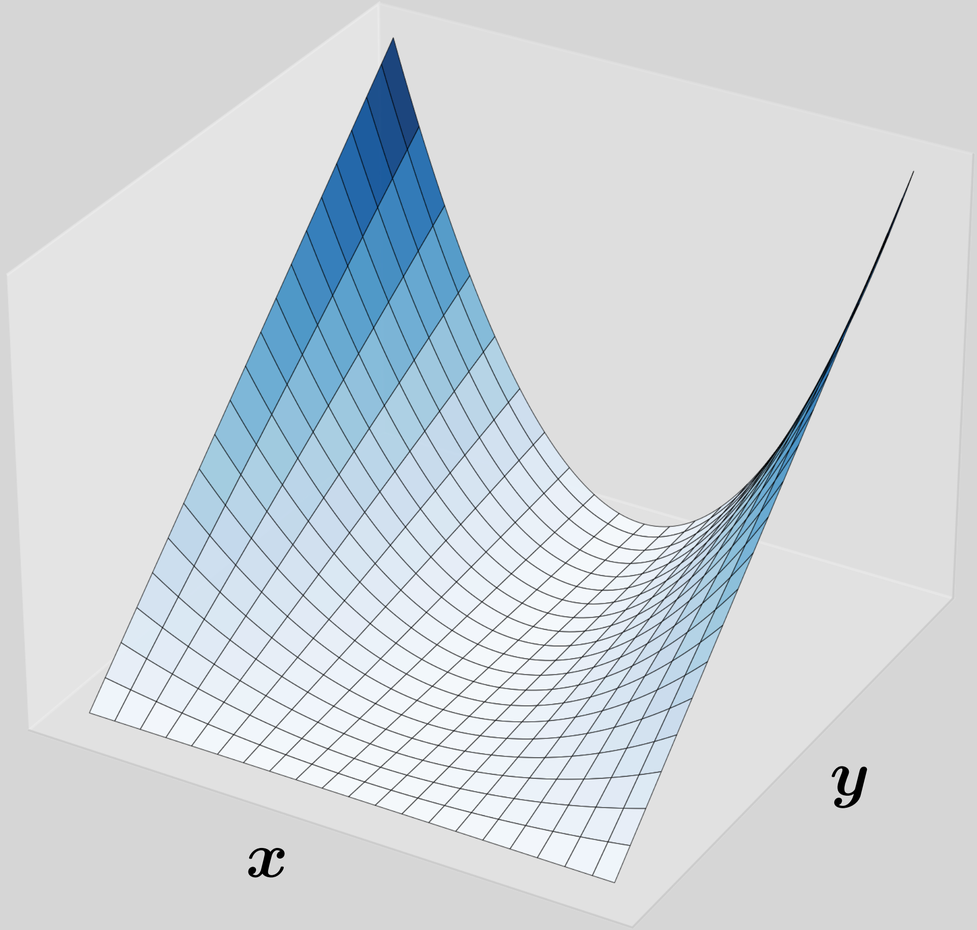
$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

Example: Minimization with Functional Constraints

$$\min_{\boldsymbol{x}\in\mathcal{X}} \phi(\boldsymbol{x})$$

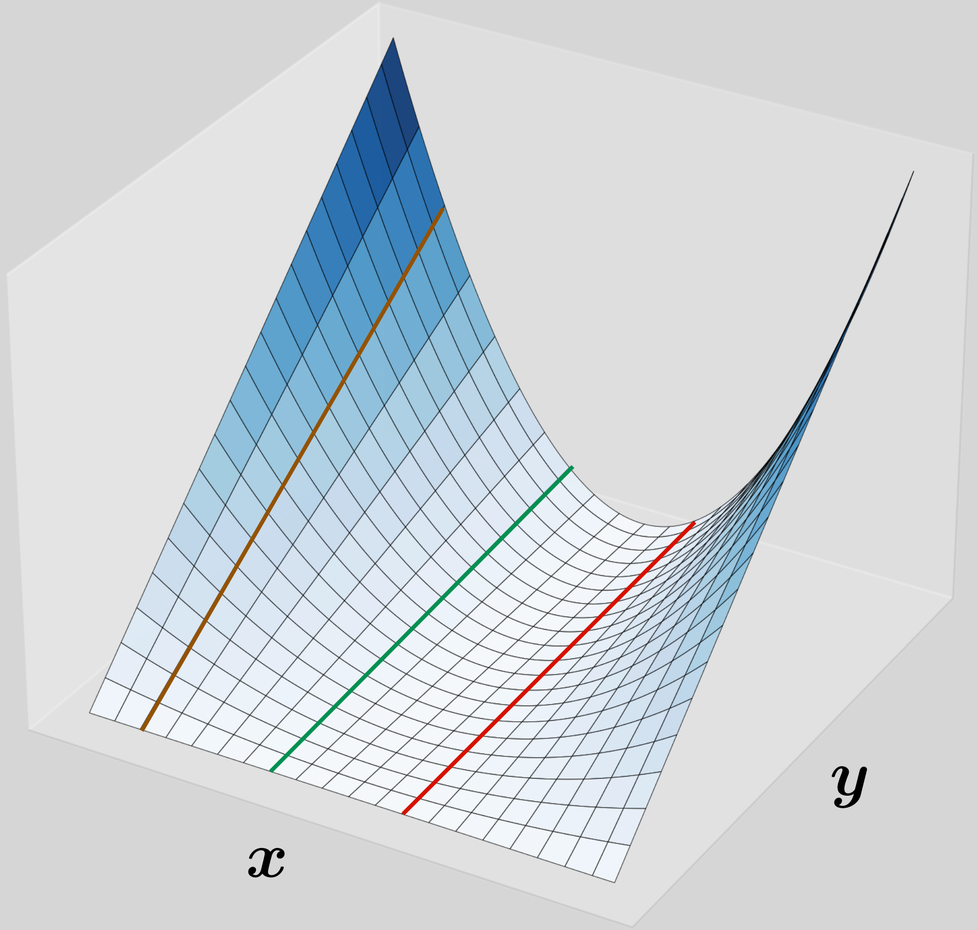
$$\text{s. t. } f_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, n$$

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \underbrace{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle} - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$



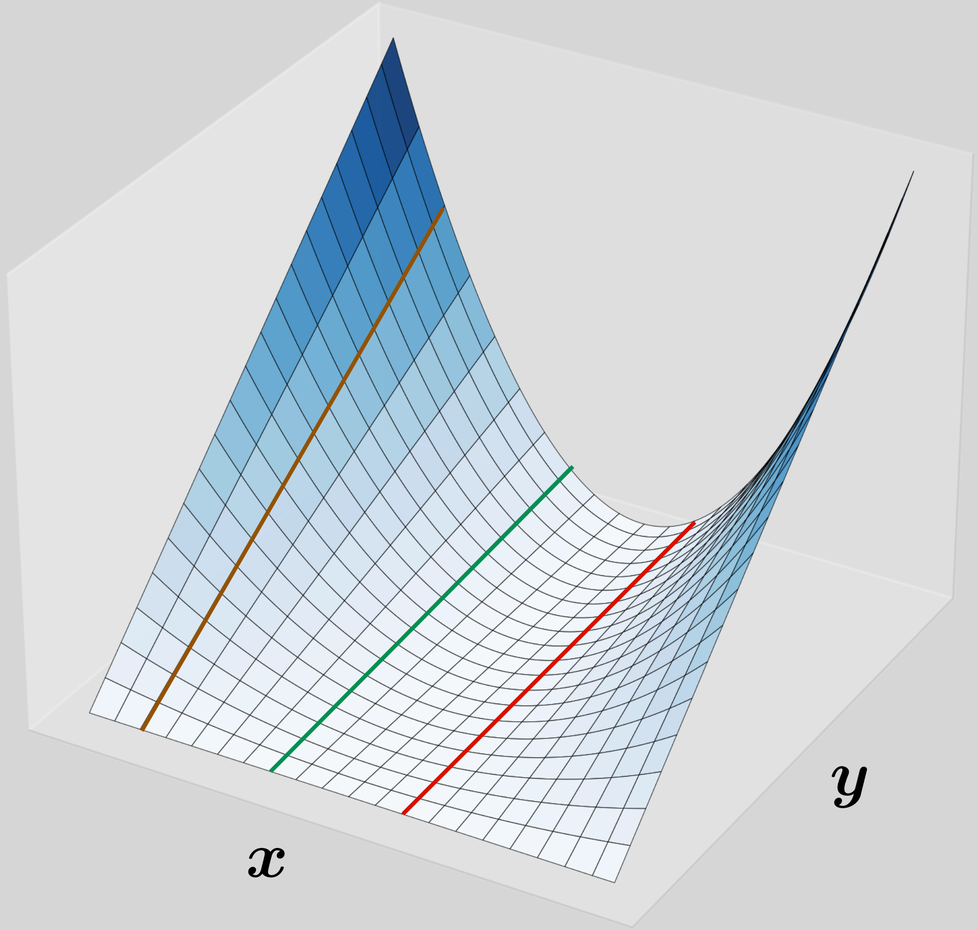
Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \underbrace{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle} - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$



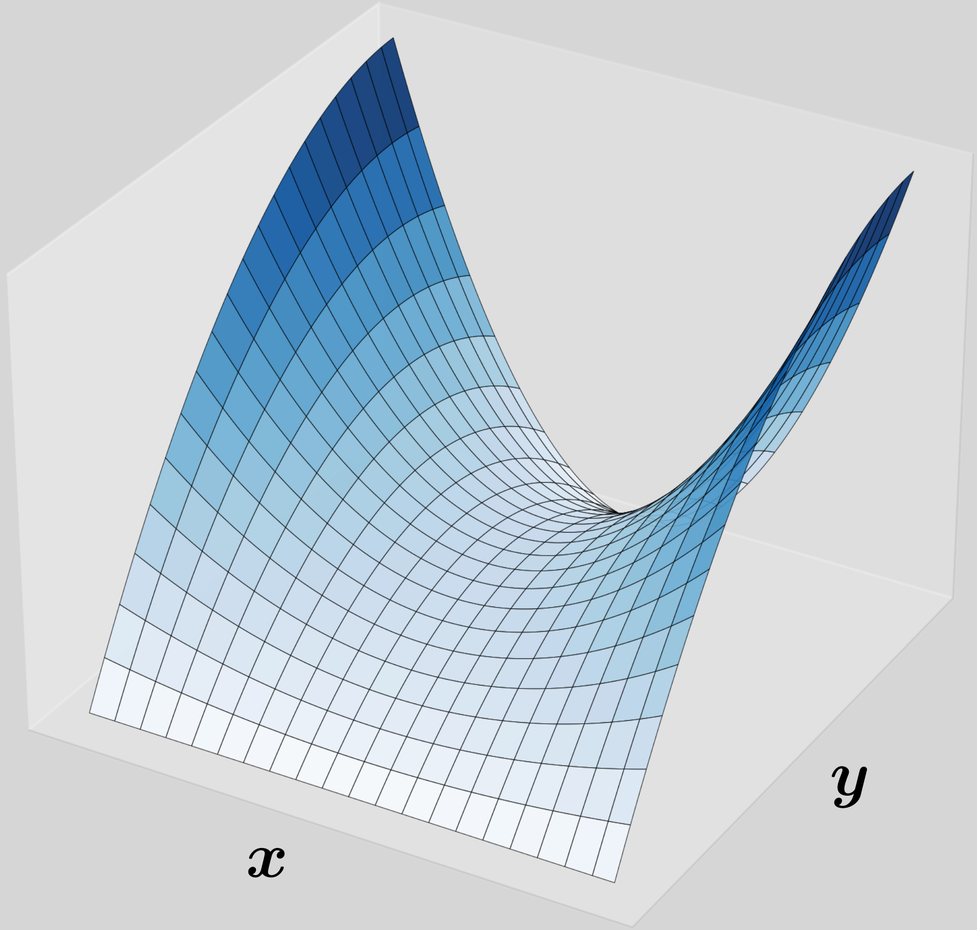
Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \underbrace{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle}_{\subseteq \langle \boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}\rangle \text{ Bilinear}} - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x}) \underbrace{\subseteq g(\boldsymbol{x}, \boldsymbol{y})}_{\text{Nonbilinear}}$$



Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \underbrace{\langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})}$$



Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

What do we hope to achieve?

Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

What do we hope to achieve?

Functions	Lipschitz	Smooth
f_1		
$\vdots$		
f_n		

Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

What do we hope to achieve?

Functions	Lipschitz	Smooth
f_1	G_1	
$\vdots$	$\vdots$	
f_n	G_n	

$$|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}')| \leq G_i \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}$$

Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

What do we hope to achieve?

Functions	Lipschitz	Smooth
f_1	G_1	L_1
$\vdots$	$\vdots$	$\vdots$
f_n	G_n	L_n

$$|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}')| \leq G_i \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}$$

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}')\|_{\mathcal{X}^*} \leq L_i \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}$$

Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

What do we hope to achieve?

Functions	Lipschitz	Smooth
f_1	G_1	L_1
$\vdots$	$\vdots$	$\vdots$
f_n	G_n	L_n

$$|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}')| \le G_i \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}$$

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}')\|_{\mathcal{X}^*} \le L_i \|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}}$$

1. Can we use **arbitrary norms** (not only the Euclidean) to define (G_i, L_i)?
2. Can we achieve the best **aggregation** of constants in the final complexity?

$$nG_{\max} \ge \sqrt{n\sum_{i=1}^n G_i^2} \ge \sum_{i=1}^n G_i$$

Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

lower is better

Global Complexity Guarantee (Convex/Concave):

$$O\left(\frac{d\sqrt{(\sum_{i=1}^{n}\lambda_i)\cdot(\sum_{i=1}^{n}G_i+\sum_{i=1}^{n}L_i)}}{\epsilon}\right)$$

$$\lambda_i = \sqrt{G_i^2 + L_i^2}$$

Extragradient (Korpelevich '76)/ Mirror Prox (Nemirovski, '04) Baseline:

$$O\left(\frac{nd\sqrt{G^2+L^2}}{\epsilon}\right)$$

(Lipschitz/Smoothness constants of the vector-valued function *f*).

Jelena Diakonikolas
University of Wisconsin-Madison

Zaid Harchaoui
University of Washington

Semilinear Min-Max Problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \langle \boldsymbol{y}, f(\boldsymbol{x})\rangle - \psi(\boldsymbol{y}) + \phi(\boldsymbol{x})$$

Dual Linear Min-Max Optimization

Ronak Mehta[1] Jelena Diakonikolas[2] Zaid Harchaoui[1]

[1]University of Washington, Seattle [2]University of Wisconsin, Madison

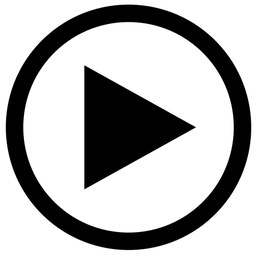
June 2, 2025

Abstract

We study a class of convex-concave min-max problems in which the coupled component of the objective is linear in at least one of the two decision vectors. This problem structure interpolates between the well-studied bilinearly coupled and relatively less-studied nonbilinearly coupled min-max problems Relevant applications include distributionally robust optimization, fully composite optimization, and convex minimization with functional constraints. We prove the first complexity upper bounds for this class using an algorithm that combines randomized and cyclic coordinate-wise updates, which have previously been employed to achieve optimal complexities for bilinearly-coupled min-max problems. The analysis handles strongly convex and non-strongly convex components in a unified manner. We also provide lower bounds for particular parameter regimes of practical interest.

Jelena Diakonikolas
University of Wisconsin-Madison

Zaid Harchaoui
University of Washington

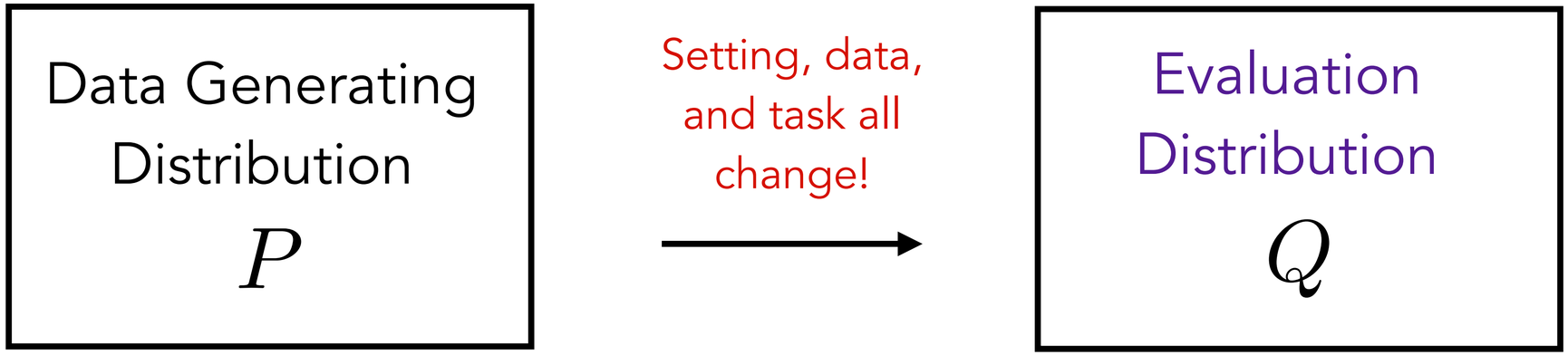


Data Generating Distribution

P

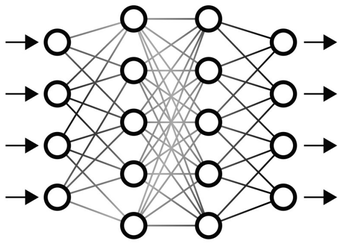
Evaluation Distribution

Q


Data Generating Distribution
P
Setting, data, and task all change!
Evaluation Distribution
Q

Data Generating Distribution

P



Foundation Model

Setting, data, and task all change!

Evaluation Distribution

Q

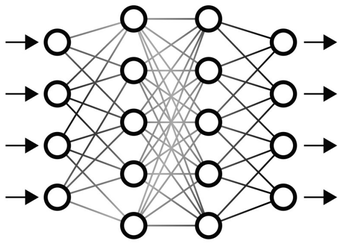
Data Generating Distribution

P

Setting, data, and task all change!

Evaluation Distribution

Q



Foundation Model

PROMPT

Generalize to unseen tasks, *without* additional training data.

The Mystery of Foundation Models

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture

Mahmoud Assran[1,2,3*] Quentin Duval[1] Ishan Misra[1] Piotr Bojanowski[1]
Pascal Vincent[1] Michael Rabbat[1,3] Yann LeCun[1,4] Nicolas Ballas[1]

[1]Meta AI (FAIR) [2]McGill University [3] Mila, Quebec AI Institute [4]New York University

GPT-4 Technical Report

GPT-4.

deepseek

DeepSeek-R1: Incentivizing Reasoning Capabilit Reinforcement Learning

DeepSeek-AI

research@deepseek.com

DINOv2: Learning Robust Visual Features without Supervision

laxime Oquab**, Timothée Darcet**, Théo Moutakanni**,
Marc Szafraniec*, Vasil Khalidov*, Pierre Fernandez, Daniel Haziza,
Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba,
wes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat,
arma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal[1],
Patrick Labatut*, Armand Joulin*, Piotr Bojanowski*

Meta AI Research [1]*Inria*

*core team **equal contribution

Meta

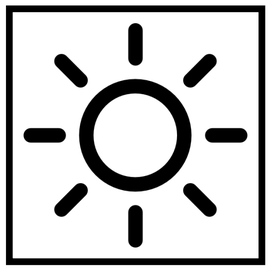
The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]
[1]A detailed contributor list can be found in the appendix of this paper.

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech capabilities into Llama 3 via a compositional approach. We observe this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.

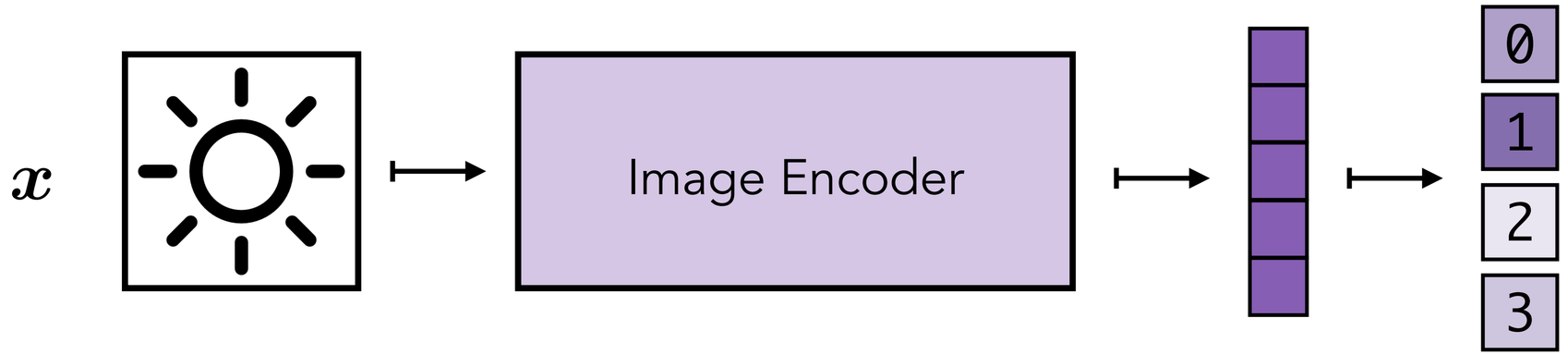
Date: July 23, 2024
Website: https://llama.meta.com/

Supervised Learning

x 

0	moon
1	sun
2	star
3	planet

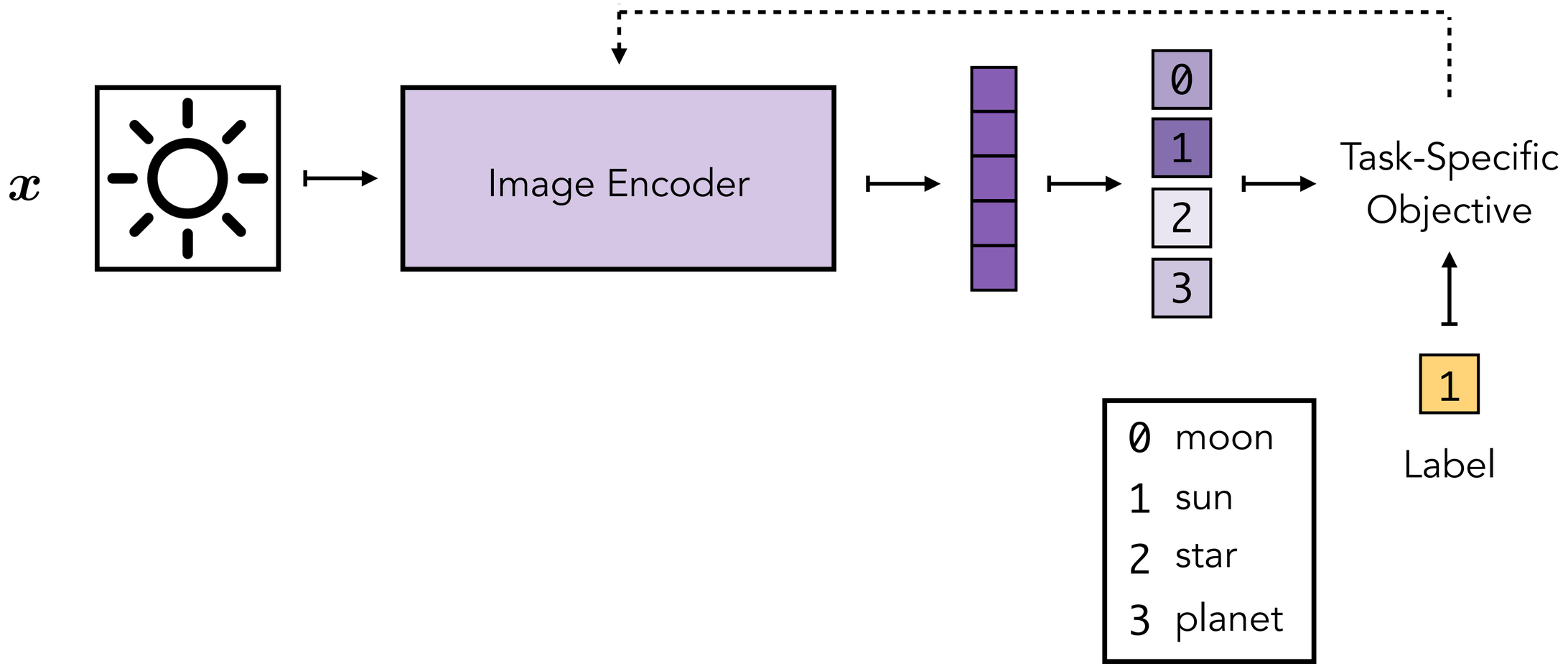
Supervised Learning



0	moon
1	sun
2	star
3	planet

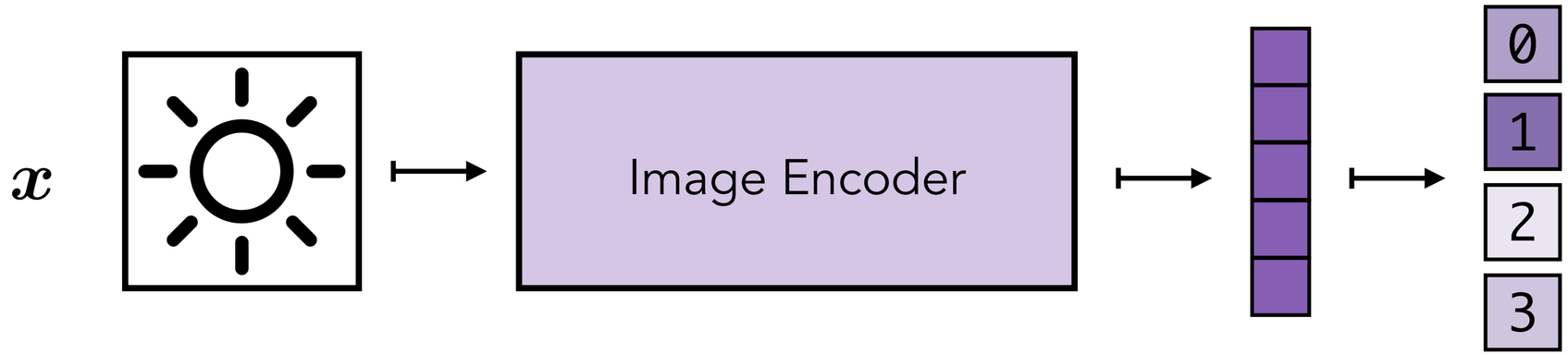
Training

Supervised Learning



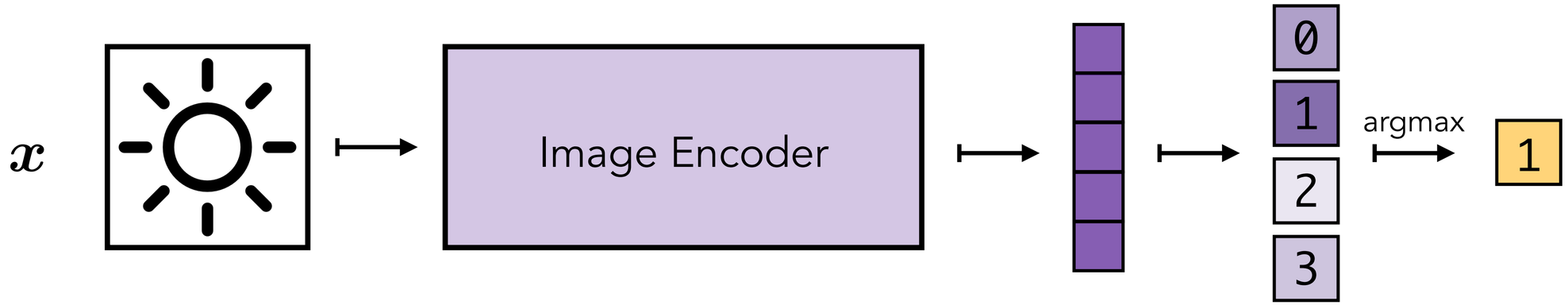
Training

Supervised Learning

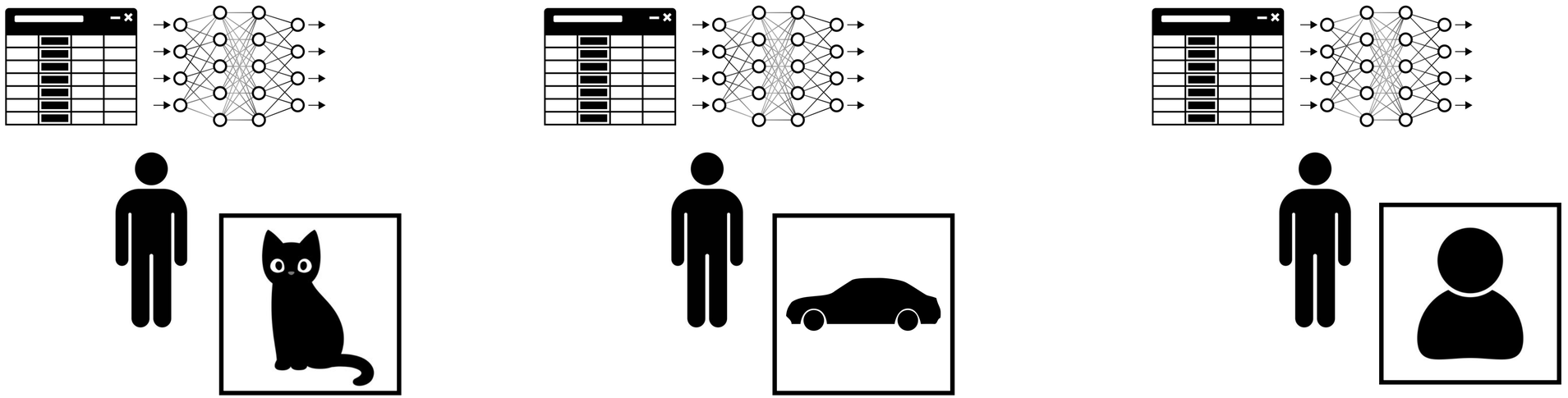


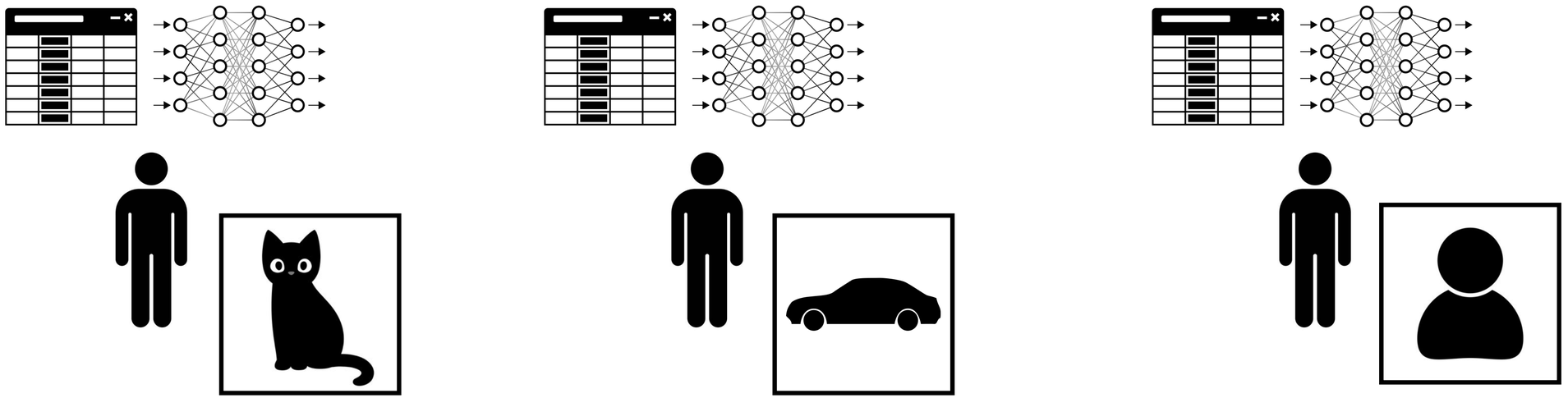
Evaluation

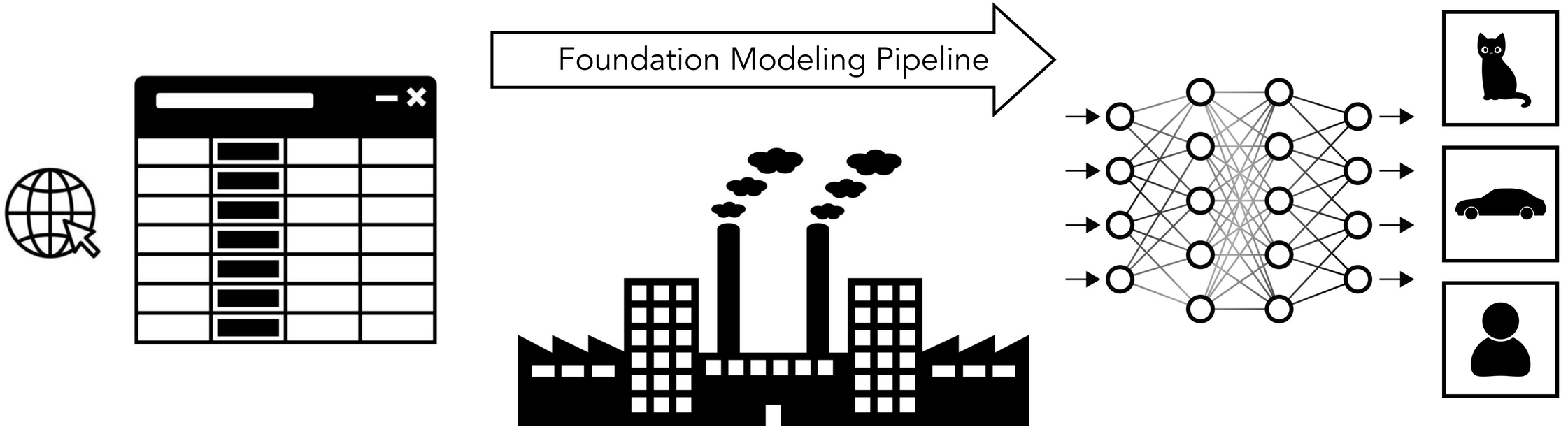
Supervised Learning



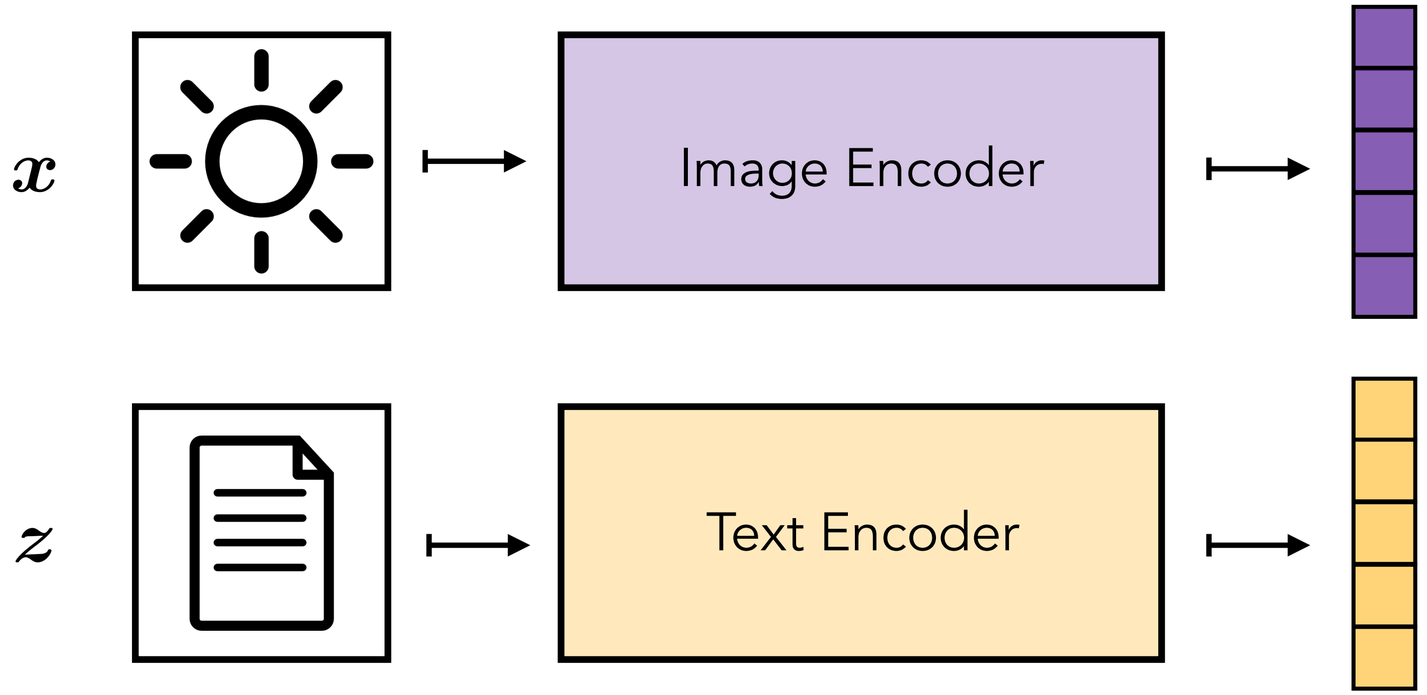
Evaluation



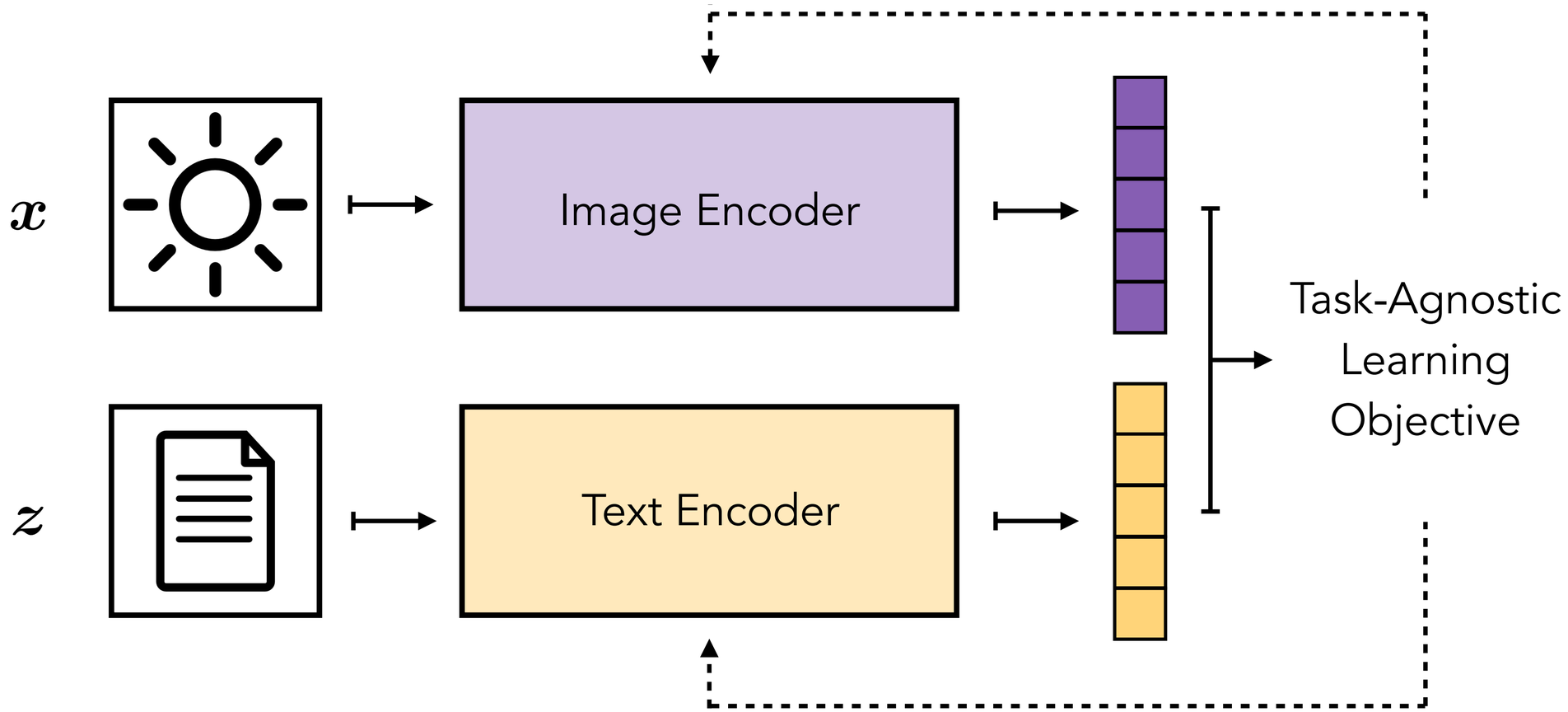



Foundation Modeling Pipeline

Foundation Modeling

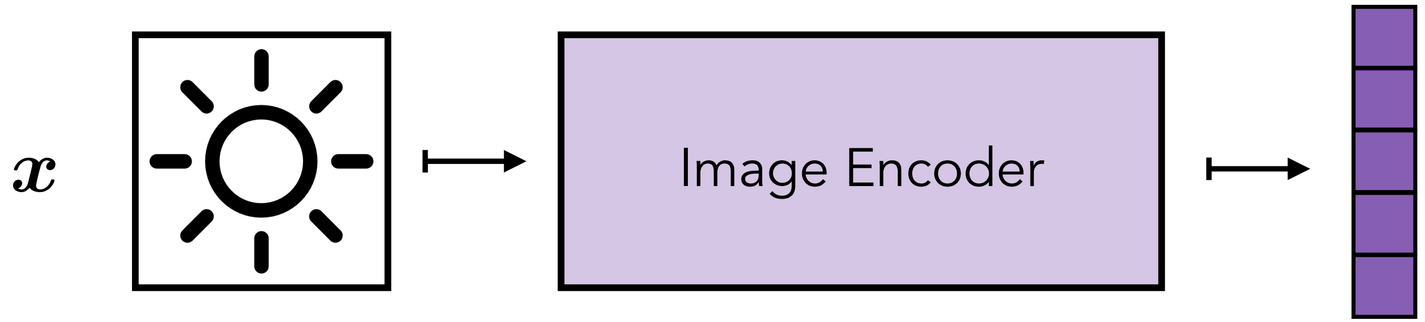


Foundation Modeling



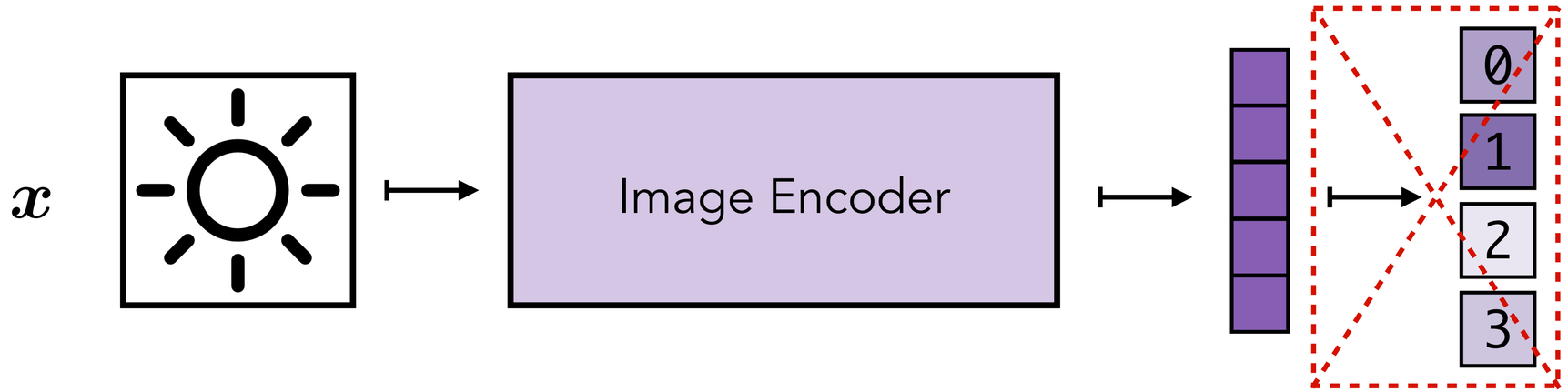
Pre-Training

Foundation Modeling



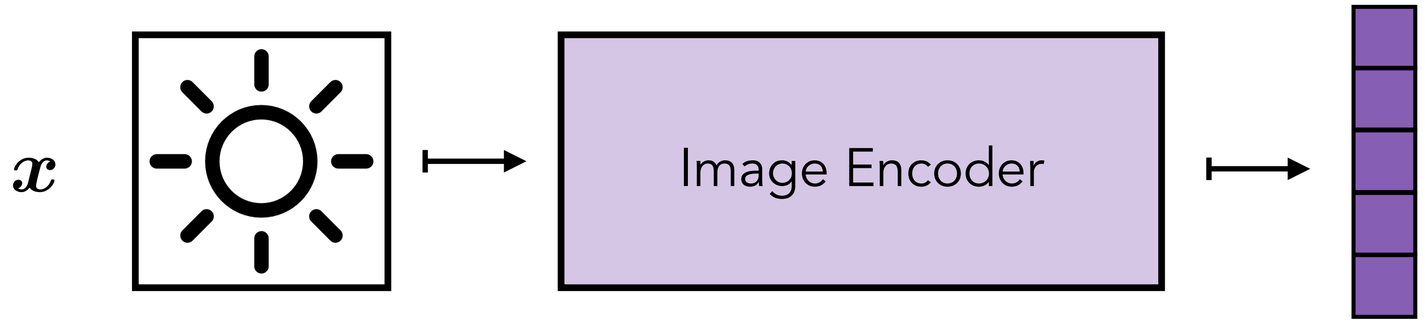
Evaluation

Foundation Modeling



Evaluation

Foundation Modeling



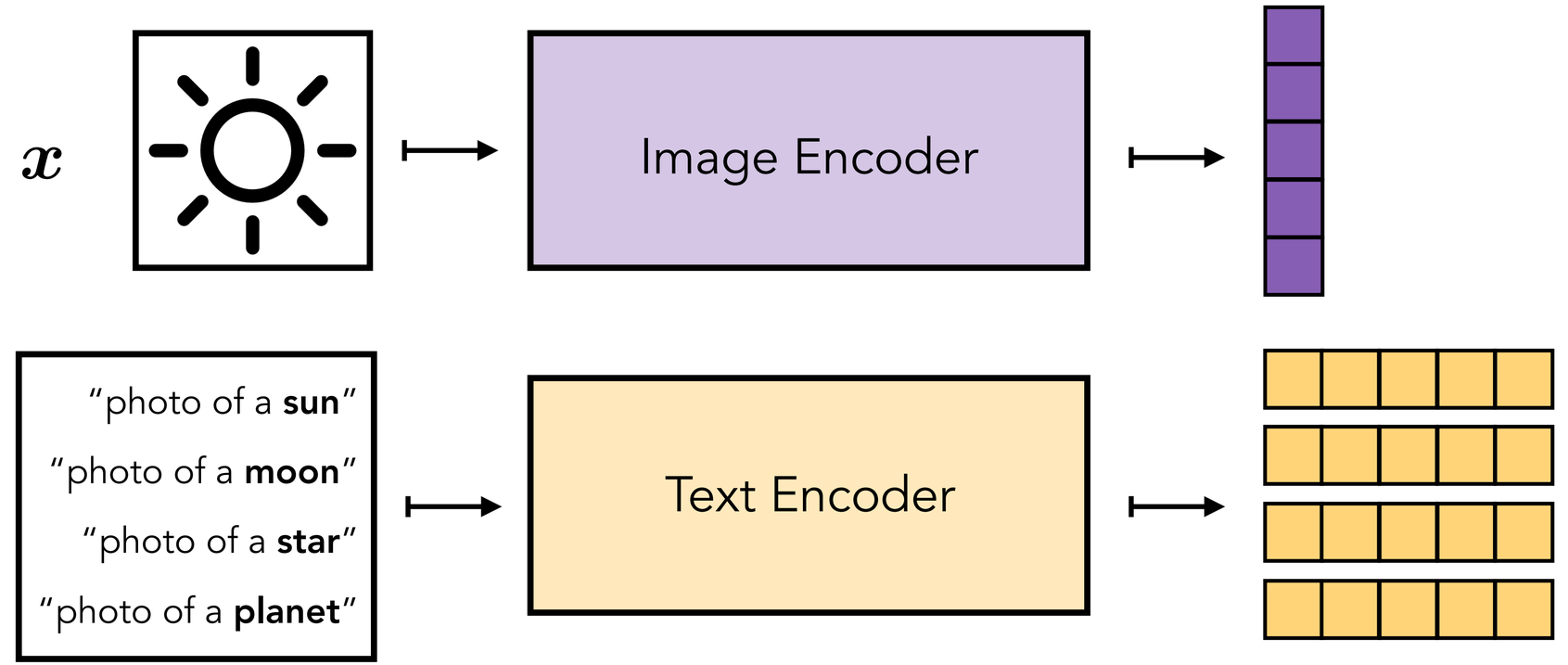
$(z^y)_{y \in \mathcal{Y}}$

"photo of a **sun**"
"photo of a **moon**"
"photo of a **star**"
"photo of a **planet**"

Idea: Convert labels into pseudo-captions (prompts)

Evaluation

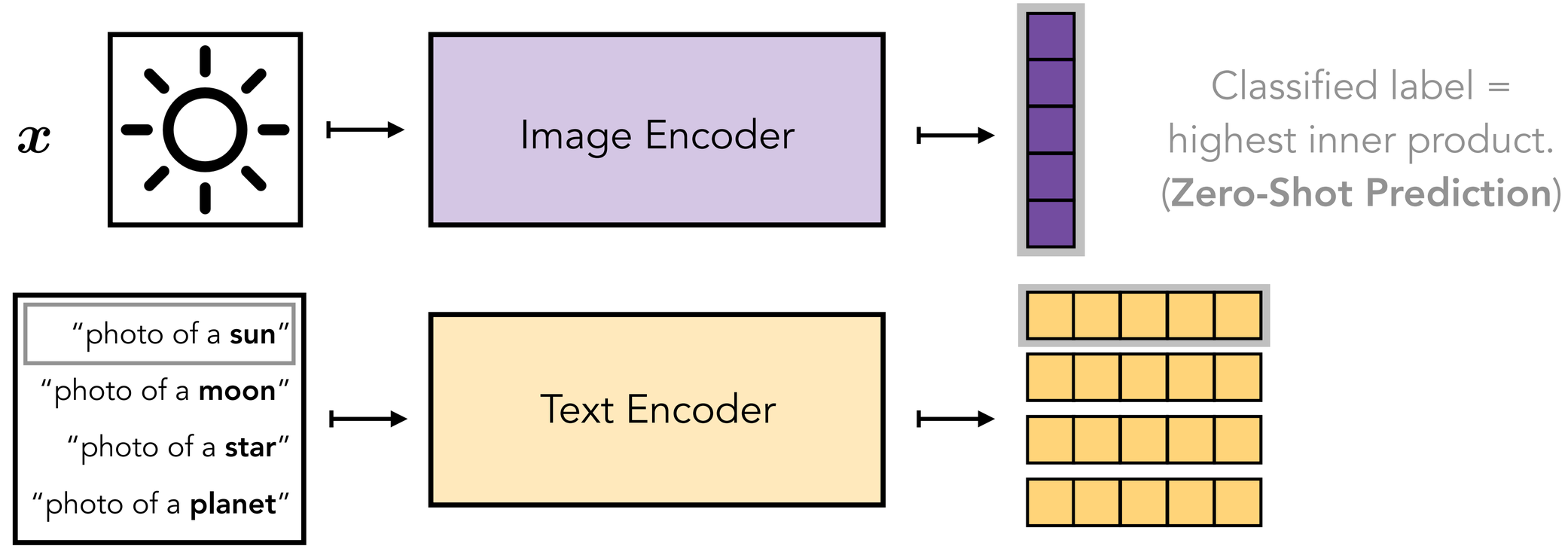
Foundation Modeling



Idea: Convert labels into pseudo-captions (prompts)

Evaluation

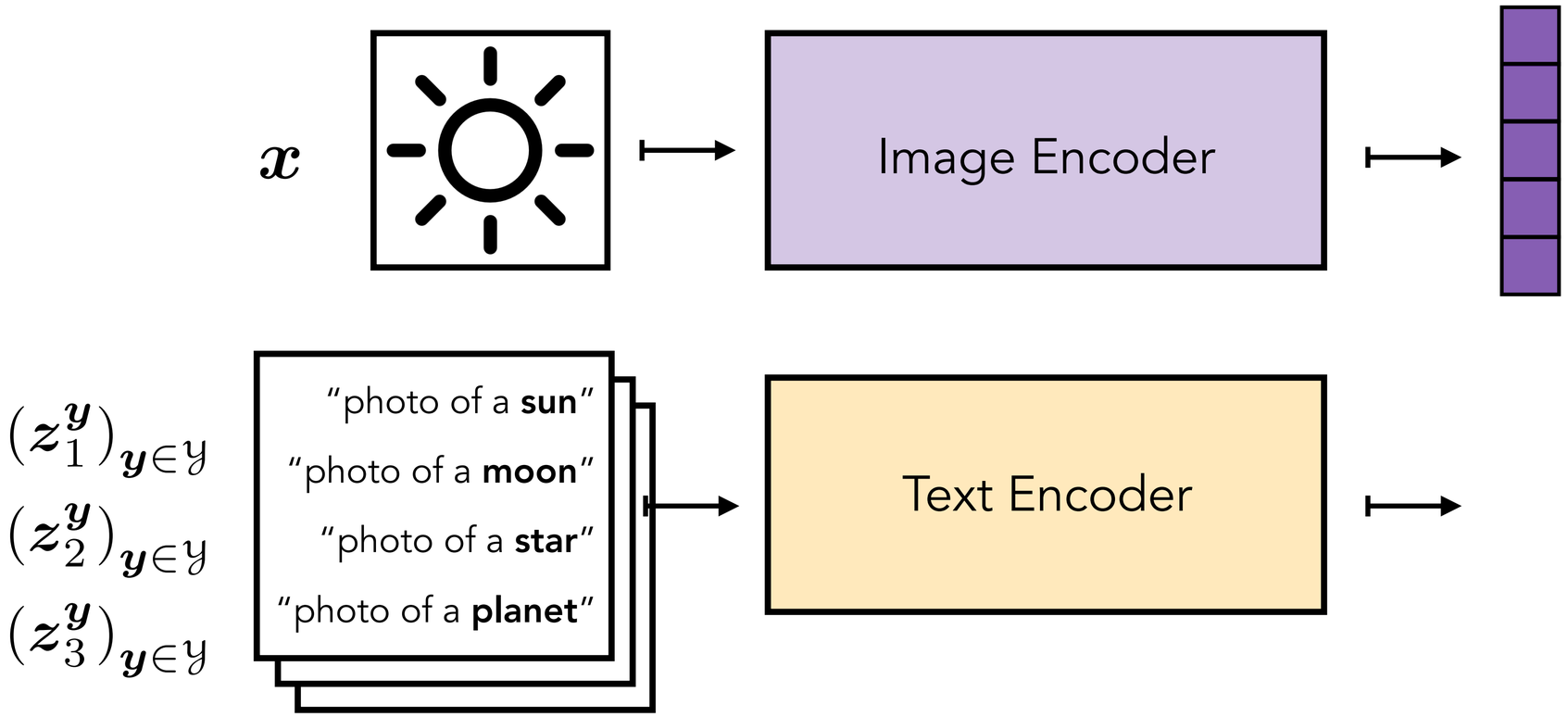
Foundation Modeling



Idea: Convert labels into pseudo-captions (prompts)

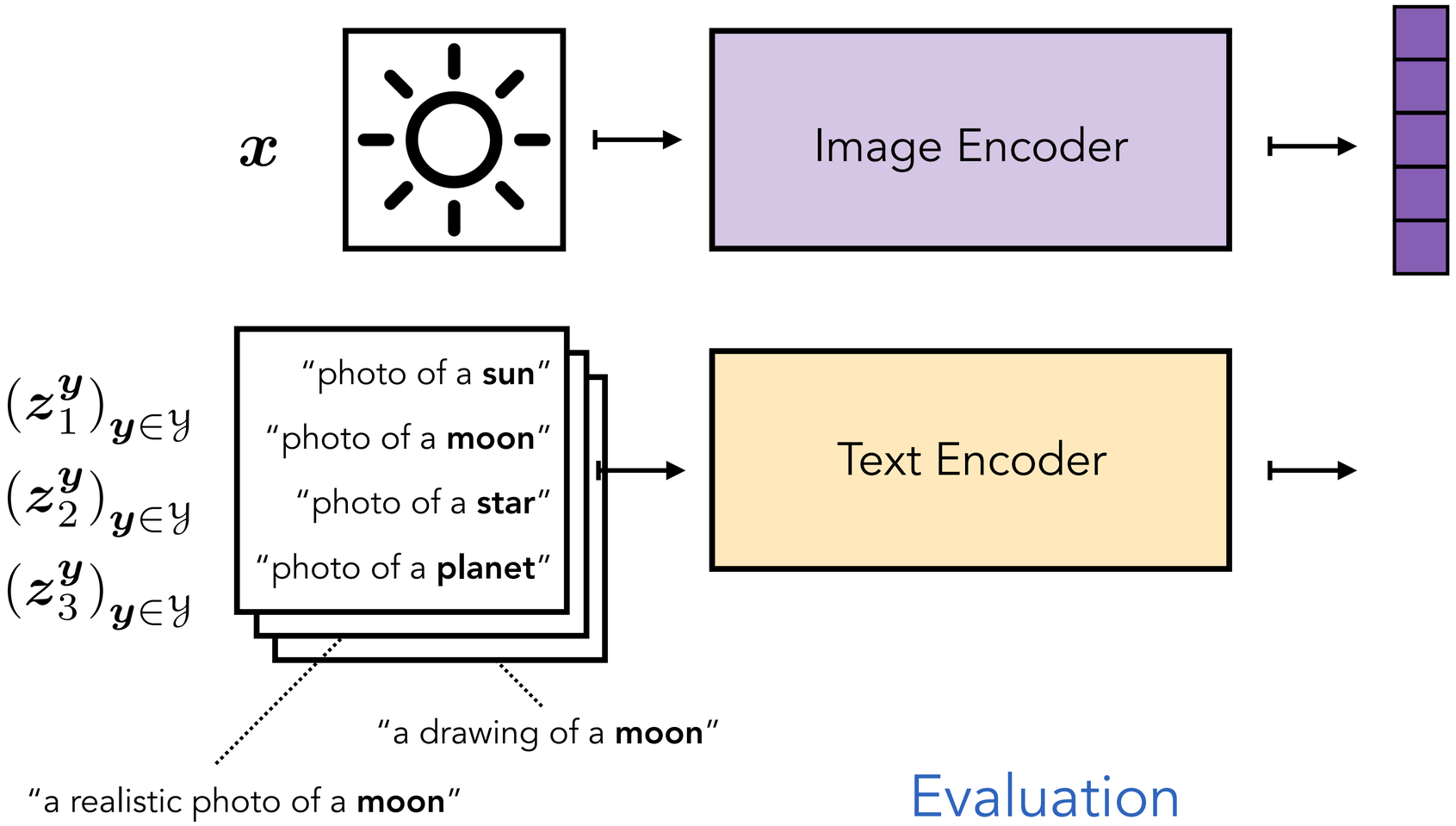
Evaluation

Foundation Modeling



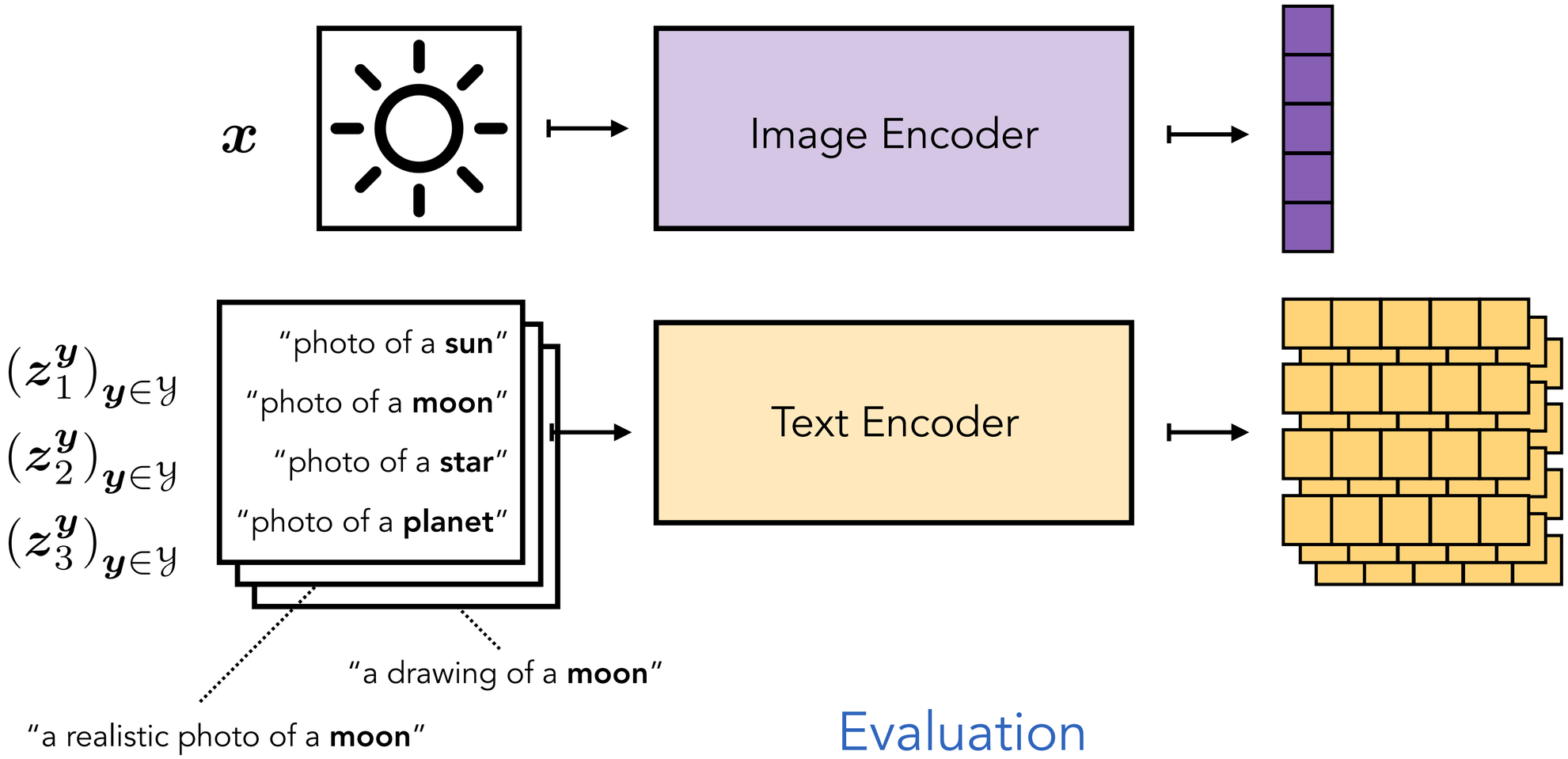
Evaluation

Foundation Modeling



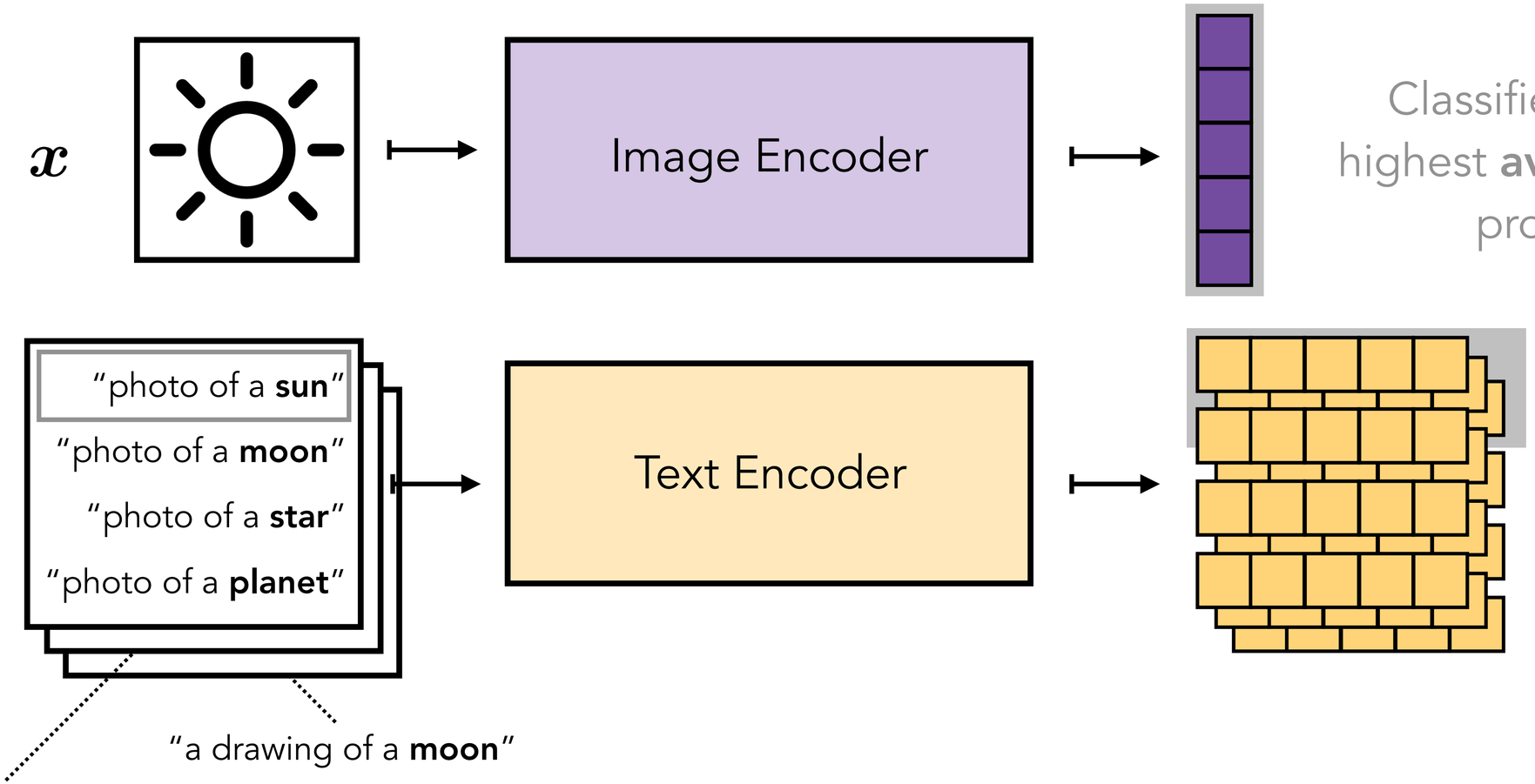
Evaluation

Foundation Modeling



Evaluation

Foundation Modeling



Evaluation

Many empirical studies have confirmed that selecting pre-training data is arguably the most important part of foundation modeling. Specifically, **balancing** paired data to achieve certain marginal constraints is an effective recipe.

DATACOMP:
In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2], Gabriel Ilharco*[1], Alex Fang*[1], Jonathan Hayase[1], Georgios Smyrnis[5], Thao Nguyen[1], Ryan Marten[7,9], Mitchell Wortsman[1], Dhruba Ghosh[1], Jieyu Zhang[1], Eyal Orgad[3], Rahim Entezari[10], Giannis Daras[5], Sarah Pratt[1], Vivek Ramanujan[1], Yonatan Bitton[11], Kalyani Marathe[1], Stephen Mussmann[1], Richard Vencu[6], Mehdi Cherti[6,8], Ranjay Krishna[1], Song[2],

DEMYSTIFYING CLIP DATA

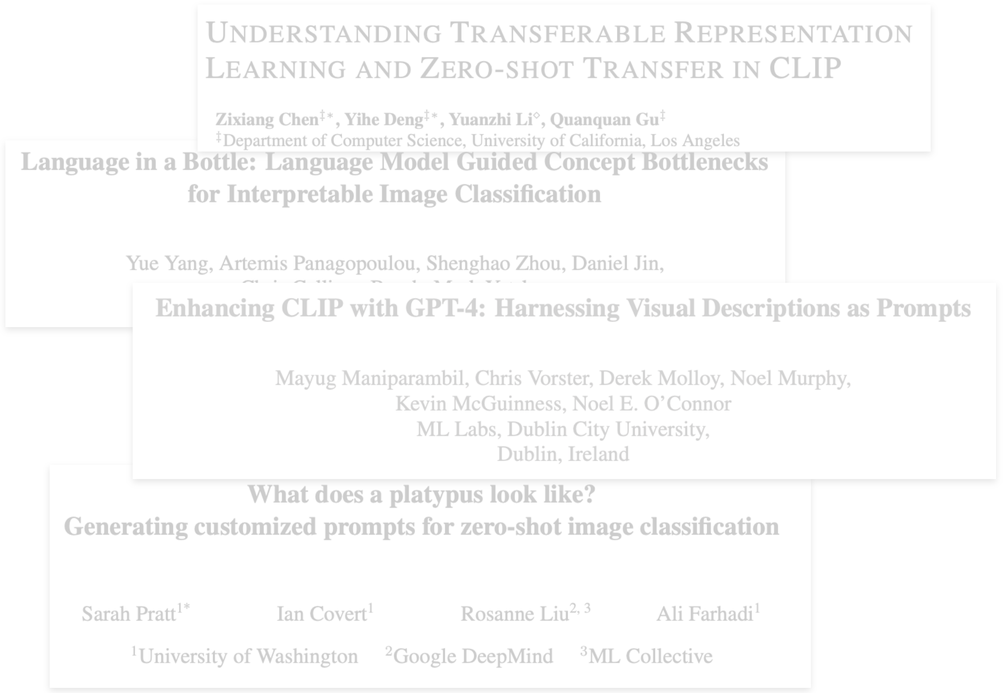
Hu Xu[1] Saining Xie[2] Xiaoqing Ellen Tan[1] Po-Yao Huang[1] Russell Howes[1] Vasu Sharma[1]
Shang-Wen Li[1] Gargi Ghosh[1] Luke Zettlemoyer[1,3] Christoph Feichtenhofer[1]
[1]FAIR, Meta AI

DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

Sang Michael Xie[1,2], Hieu Pham[1], Xuanyi Dong[1], Nan Du[1], Hanxiao Liu[1], Yifeng Lu[1], Percy Liang[2], Quoc V. Le[1], Tengyu Ma[2], and Adams Wei Yu[1]

[1]Google DeepMind
[2]Stanford University

Techniques for pre-training, prompting, and overall model reuse have advanced significantly in applications. Virtually no **generalization guarantees** (from the perspective of statistical learning theory) exist for this pipeline.



UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen[‡*], Yihe Deng[‡*], Yuanzhi Li[◇], Quanquan Gu[‡]
[‡]Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
Kevin McGuinness, Noel E. O'Connor
ML Labs, Dublin City University,
Dublin, Ireland

What does a platypus look like?
Generating customized prompts for zero-shot image classification

Sarah Pratt[1*] Ian Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML Collective

Many empirical studies have confirmed that selecting pre-training data is arguably the most important part of foundation modeling. Specifically, **balancing** paired data to achieve certain marginal constraints is an effective recipe.


DATACOMP:
In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2], Gabriel Ilharco*[1], Alex Fang*[1], Jonathan Hayase[1], Georgios Smyrnis[5], Thao Nguyen[1], Ryan Marten[7,9], Mitchell Wortsman[1], Dhruba Ghosh[1], Jieyu Zhang[1], Eyal Orgad[3], Rahim Entezari[10], Giannis Daras[5], Sarah Pratt[1], Vivek Ramanujan[1], Yonatan Bitton[11], Kalyani Marathe[1], Stephen Mussmann[1], Richard Vencu[6], Mehdi Cherti[6,8], Ranjay Krishna[1], Song[2],

DEMYSTIFYING CLIP DATA

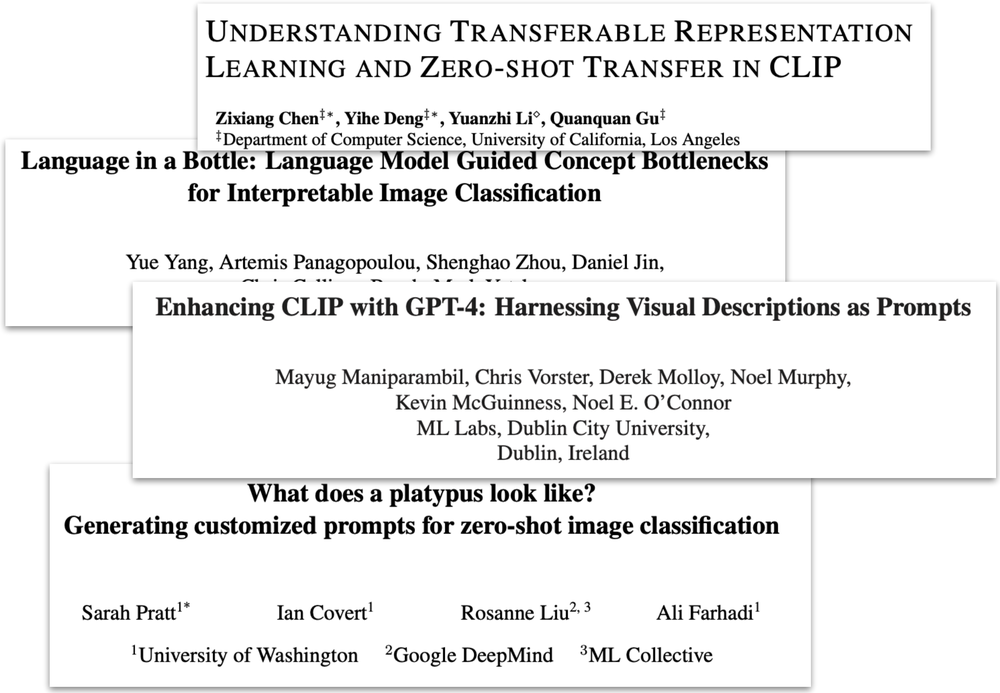
Hu Xu[1] Saining Xie[2] Xiaoqing Ellen Tan[1] Po-Yao Huang[1] Russell Howes[1] Vasu Sharma[1] Shang-Wen Li[1] Gargi Ghosh[1] Luke Zettlemoyer[1,3] Christoph Feichtenhofer[1]
[1]FAIR, Meta AI

DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

Sang Michael Xie[1,2], Hieu Pham[1], Xuanyi Dong[1], Nan Du[1], Hanxiao Liu[1], Yifeng Lu[1], Percy Liang[2], Quoc V. Le[1], Tengyu Ma[2], and Adams Wei Yu[1]

[1]Google DeepMind
[2]Stanford University

Techniques for pre-training, prompting, and overall model reuse have advanced significantly in applications. Virtually no **generalization guarantees** (from the perspective of statistical learning theory) exist for this pipeline.


UNDERSTANDING TRANSFERABLE REPRESENTATION LEARNING AND ZERO-SHOT TRANSFER IN CLIP

Zixiang Chen‡*, Yihe Deng‡*, Yuanzhi Li◇, Quanquan Gu‡
‡Department of Computer Science, University of California, Los Angeles

Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,

Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy,
Kevin McGuinness, Noel E. O'Connor
ML Labs, Dublin City University,
Dublin, Ireland

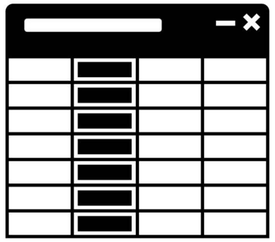
What does a platypus look like?
Generating customized prompts for zero-shot image classification

Sarah Pratt[1*] Ian Covert[1] Rosanne Liu[2, 3] Ali Farhadi[1]
[1]University of Washington [2]Google DeepMind [3]ML Collective

Motivating Questions

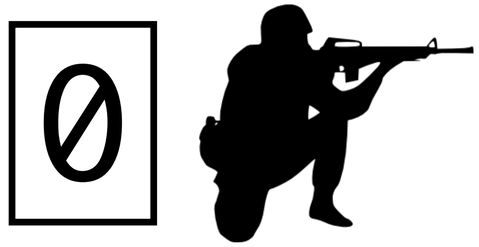
Foundation Modeling

What is the statistical effect of **balancing** methods on the pre-training and the resulting foundation model?



Zero-Shot Prediction

How do we analyze prompt-based **zero-shot prediction** as a statistical estimator, such as its comparison to direct supervision?



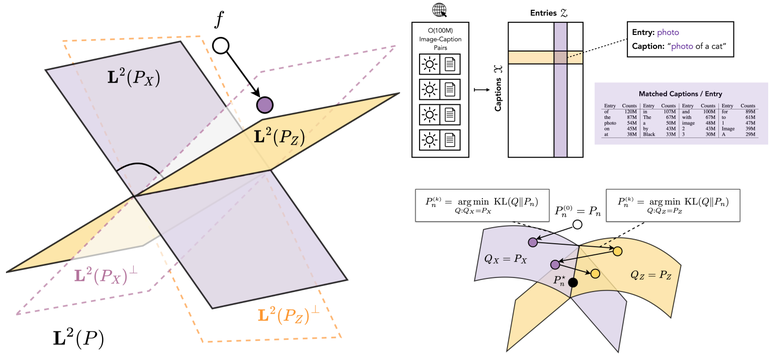
Research Contributions

Exact variance reduction analysis of balancing-based mean estimation!

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

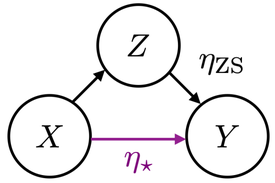
(k times)

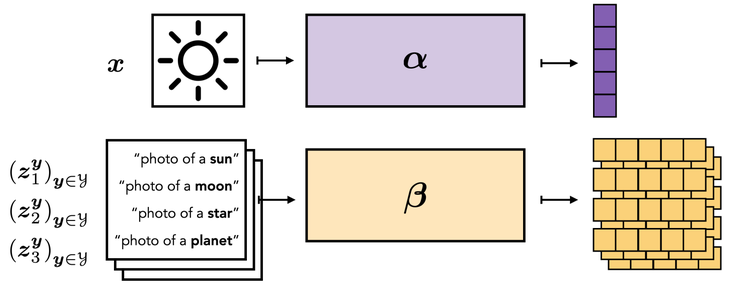


Theoretical framework for obtaining generalization guarantees for two broad classes of zero-shot prediction models!

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$





Preliminaries

Balanced Pre-Training

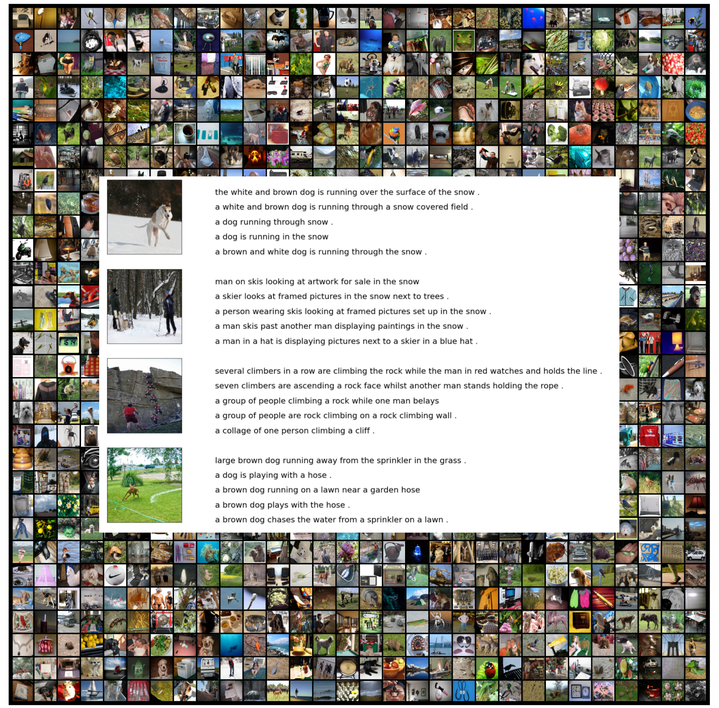
Zero-Shot Prediction

Conclusion

$\mathcal{X}$ Sample Space (Images)

$\mathcal{Z}$ Sample Space (Text)

$P \equiv P_{X,Z}$ Pre-Training Distribution



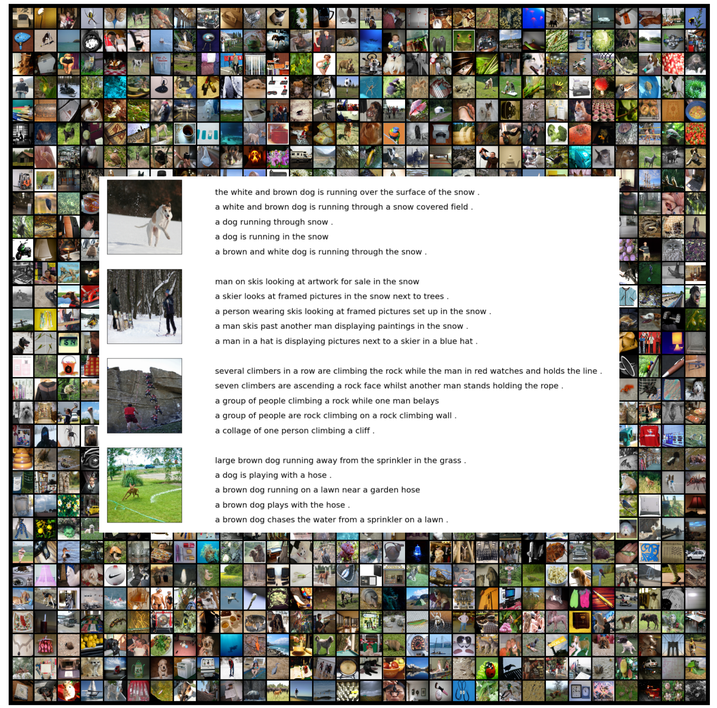
$\mathcal{X}$ Sample Space (Images)

$\mathcal{Z}$ Sample Space (Text)

$P \equiv P_{X,Z}$ Pre-Training Distribution

$$\mathbf{L}^2(P) = \left\{ f : \|f\|^2_{\mathbf{L}^2(P)} := \int_{\mathcal{X}\times\mathcal{Z}} f^2(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}P(\boldsymbol{x}, \boldsymbol{z}) < \infty \right\}$$

$\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$ Defined Similarly



$\mathcal{X}$ Sample Space (Images)

$\mathcal{Z}$ Sample Space (Text)

$P \equiv P_{X,Z}$ Pre-Training Distribution

$$\mathbf{L}^2(P) = \left\{ f : \|f\|^2_{\mathbf{L}^2(P)} := \int_{\mathcal{X}\times\mathcal{Z}} f^2(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}P(\boldsymbol{x}, \boldsymbol{z}) < \infty \right\}$$

$\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$ Defined Similarly

Conditional Mean

Projecting onto $\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} \mathbb{E}_P \left[f(X,Z) | Z \right] (\cdot) \in \mathbf{L}^2(P_Z) \\ \mathbb{E}_P \left[f(X,Z) | X \right] (\cdot) \in \mathbf{L}^2(P_X) \end{cases}$$

$\mathcal{X}$ Sample Space (Images)

$\mathcal{Z}$ Sample Space (Text)

$P \equiv P_{X,Z}$ Pre-Training Distribution

$$\mathbf{L}^2(P) = \left\{ f : \|f\|^2_{\mathbf{L}^2(P)} := \int_{\mathcal{X}\times\mathcal{Z}} f^2(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}P(\boldsymbol{x}, \boldsymbol{z}) < \infty \right\}$$

$\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$ Defined Similarly

Conditional Mean

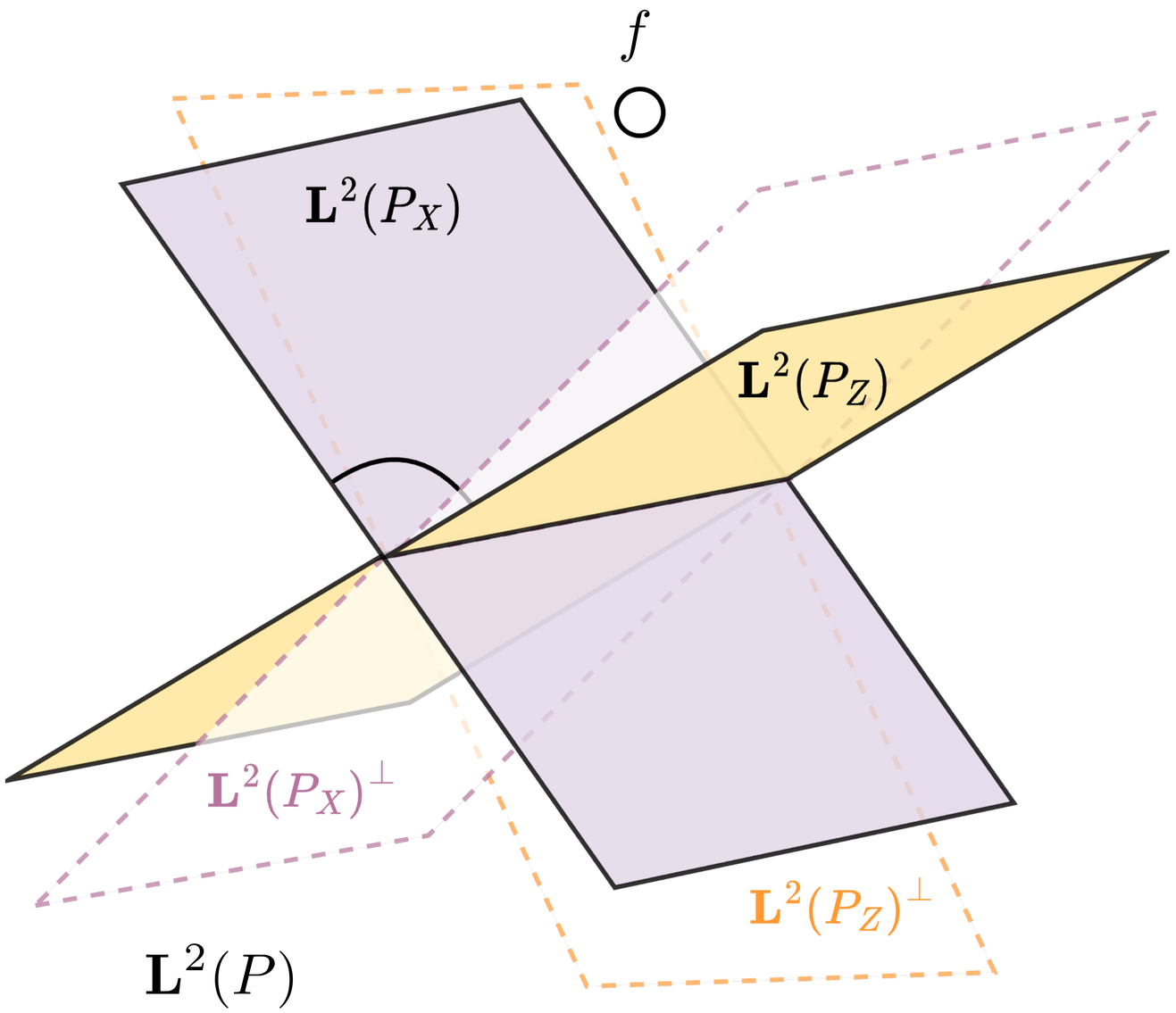
Projecting onto $\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} \mathbb{E}_P\left[f(X,Z)|Z\right](\cdot) \in \mathbf{L}^2(P_Z) \\ \mathbb{E}_P\left[f(X,Z)|X\right](\cdot) \in \mathbf{L}^2(P_X) \end{cases}$$

Conditional Centering

Projecting onto $\mathbf{L}^2(P_X)^\perp, \mathbf{L}^2(P_Z)^\perp$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} f - \mathbb{E}_P\left[f(X,Z)|Z\right](\cdot) \in \mathbf{L}^2(P) \\ f - \mathbb{E}_P\left[f(X,Z)|X\right](\cdot) \in \mathbf{L}^2(P) \end{cases}$$



Conditional Mean

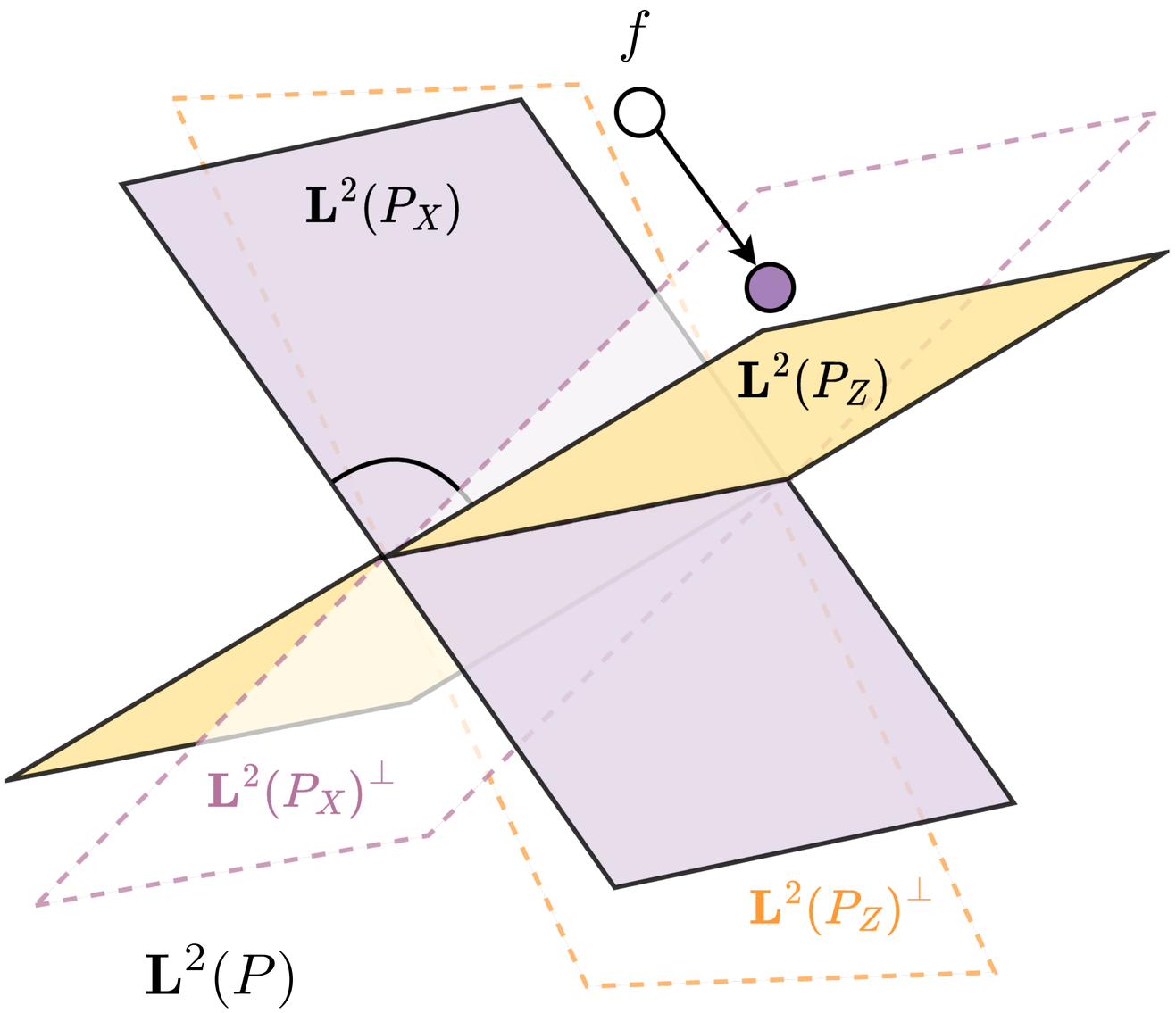
Projecting onto $\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} \mathbb{E}_P\left[f(X,Z)|Z\right](\cdot) \in \mathbf{L}^2(P_Z) \\ \mathbb{E}_P\left[f(X,Z)|X\right](\cdot) \in \mathbf{L}^2(P_X) \end{cases}$$

Conditional Centering

Projecting onto $\mathbf{L}^2(P_X)^\perp, \mathbf{L}^2(P_Z)^\perp$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} f - \mathbb{E}_P\left[f(X,Z)|Z\right](\cdot) \in \mathbf{L}^2(P) \\ f - \mathbb{E}_P\left[f(X,Z)|X\right](\cdot) \in \mathbf{L}^2(P) \end{cases}$$



Conditional Mean

Projecting onto $\mathbf{L}^2(P_X), \mathbf{L}^2(P_Z)$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} \mathbb{E}_P\left[f(X,Z)|Z\right](\cdot) \in \mathbf{L}^2(P_Z) \\ \mathbb{E}_P\left[f(X,Z)|X\right](\cdot) \in \mathbf{L}^2(P_X) \end{cases}$$

Conditional Centering

Projecting onto $\mathbf{L}^2(P_X)^\perp, \mathbf{L}^2(P_Z)^\perp$

$$f \in \mathbf{L}^2(P) \mapsto \begin{cases} f - \mathbb{E}_P\left[f(X,Z)|Z\right](\cdot) \in \mathbf{L}^2(P) \\ f - \mathbb{E}_P\left[f(X,Z)|X\right](\cdot) \in \mathbf{L}^2(P) \end{cases}$$

Conditional Mean Operator: Singular Value Decomposition

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P) \to \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} f](\boldsymbol{x}) = \mathbb{E}_P \left[f(X, Z) | X \right](\boldsymbol{x})$$

Conditional Mean Operator: Singular Value Decomposition

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P_Z) \to \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_P \left[g(Z) | X \right](\boldsymbol{x})$$

Conditional Mean Operator: Singular Value Decomposition

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P_Z) \rightarrow \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_P \left[g(Z) | X \right](\boldsymbol{x})$$

$$\langle h, \mu_{X \leftarrow Z} g \rangle_{\mathbf{L}^2(P_X)} = \mathbb{E}_{P_X} \left[h(X) \mathbb{E}_P \left[g(Z) | X \right] \right] = \mathbb{E}_P \left[h(X) g(Z) \right]$$

Conditional Mean Operator: Singular Value Decomposition

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P_Z) \to \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_P \left[g(Z) | X \right](\boldsymbol{x})$$

$$\langle h, \mu_{X \leftarrow Z} g \rangle_{\mathbf{L}^2(P_X)} = \mathbb{E}_{P_X} \left[h(X) \mathbb{E}_P \left[g(Z) | X \right] \right] = \mathbb{E}_P \left[h(X) g(Z) \right] = \mathbb{E}_{P_Z} \left[\mathbb{E}_P \left[h(X) | Z \right] g(Z) \right] = \langle \mu_{Z \leftarrow X} h, g \rangle_{\mathbf{L}^2(P_Z)}$$

The conditional mean of Z given X and vice versa are adjoint!

Conditional Mean Operator: Singular Value Decomposition

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P_Z) \to \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_P \left[g(Z) | X \right](\boldsymbol{x})$$

$$\langle h, \mu_{X \leftarrow Z} g \rangle_{\mathbf{L}^2(P_X)} = \mathbb{E}_{P_X} \left[h(X) \mathbb{E}_P \left[g(Z) | X \right] \right] = \mathbb{E}_P \left[h(X) g(Z) \right] = \mathbb{E}_{P_Z} \left[\mathbb{E}_P \left[h(X) | Z \right] g(Z) \right] = \langle \mu_{Z \leftarrow X} h, g \rangle_{\mathbf{L}^2(P_Z)}$$

The conditional mean of Z given X and vice versa are adjoint!

Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

Conditional Mean Operator: Singular Value Decomposition

$$\mu_{X\leftarrow Z} = \begin{bmatrix} | & | & \\ \alpha_1 & \alpha_2 & \cdots \\ | & | & \end{bmatrix} \begin{bmatrix} s_1 & & \\ & s_2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} - & \beta_1 & - \\ - & \beta_2 & - \\ & \vdots & \end{bmatrix}$$

$$\mu_{Z\leftarrow X} = \begin{bmatrix} | & | & \\ \beta_1 & \beta_2 & \cdots \\ | & | & \end{bmatrix} \begin{bmatrix} s_1 & & \\ & s_2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} - & \alpha_1 & - \\ - & \alpha_2 & - \\ & \vdots & \end{bmatrix}$$

The conditional mean of Z given X and vice versa are adjoint!

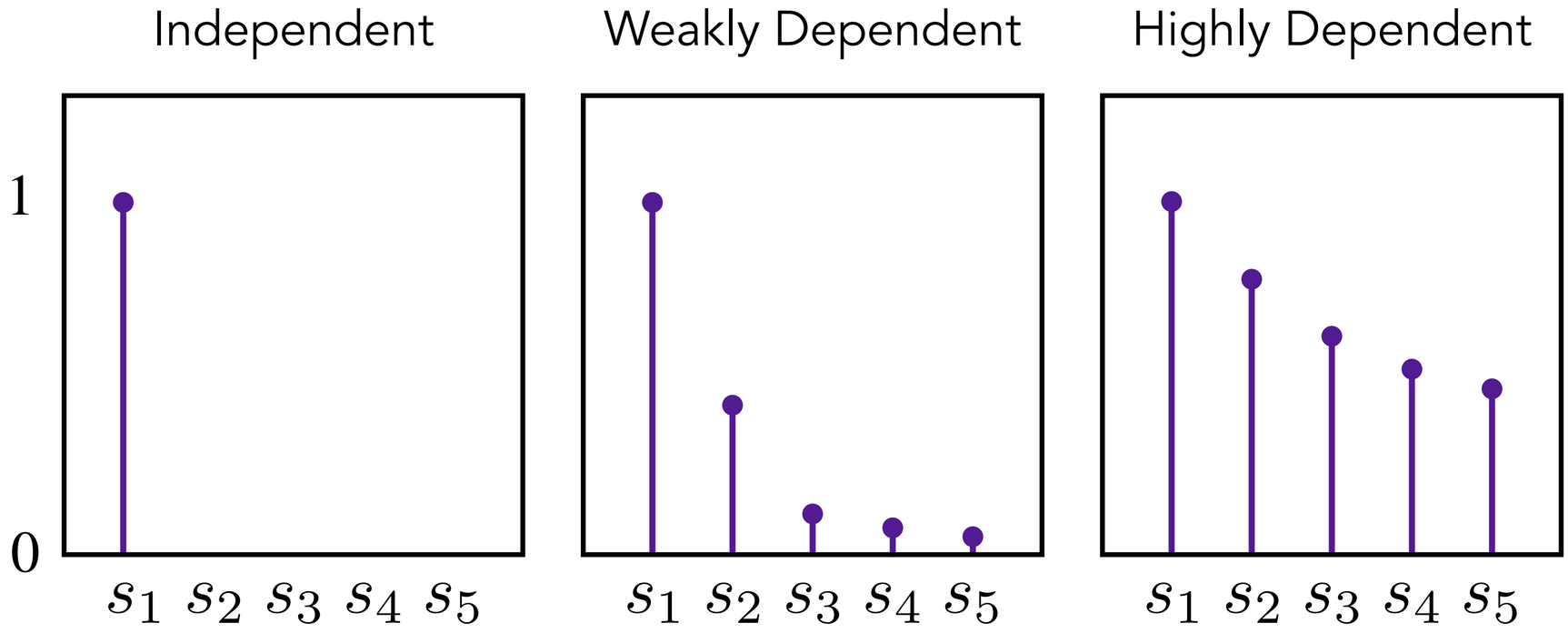
Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X\leftarrow Z}\beta_i = s_i\alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z\leftarrow X}\alpha_i = s_i\beta_i$$

Conditional Mean Operator: Singular Value Decomposition



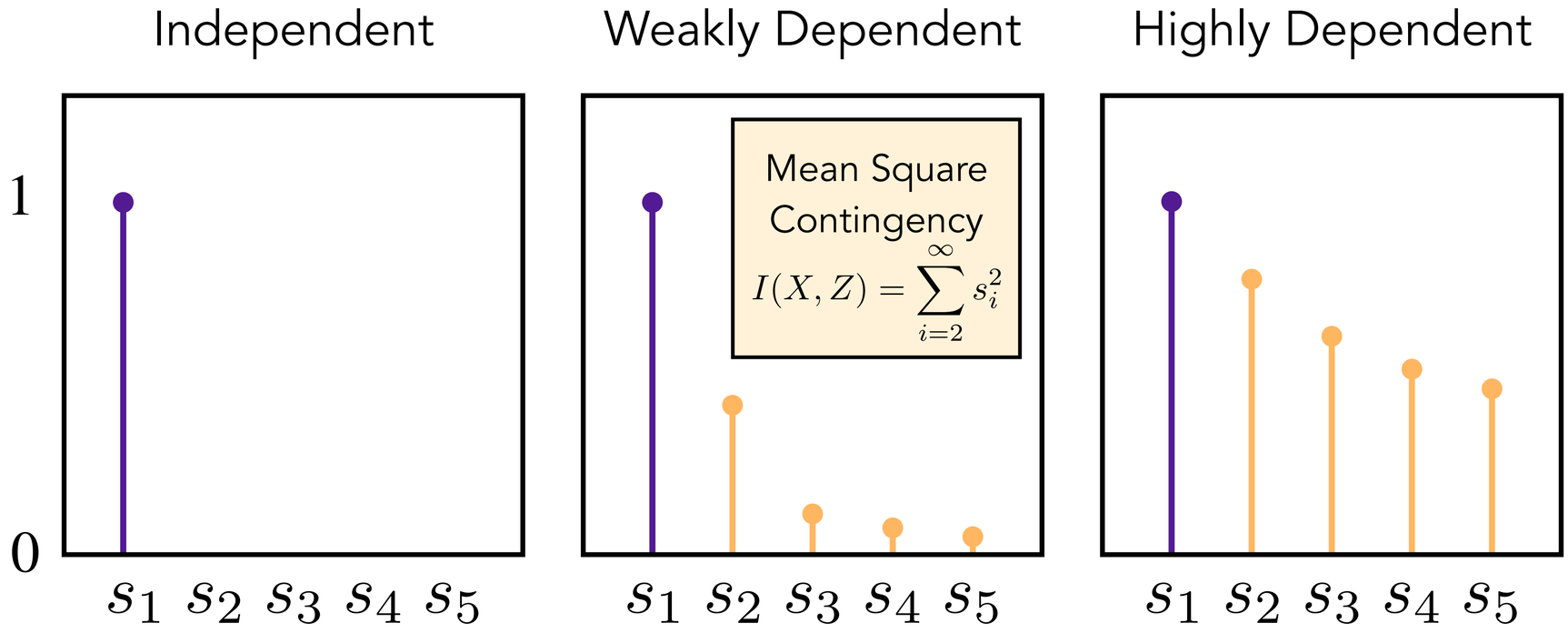
Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

Conditional Mean Operator: Singular Value Decomposition



Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

Conditional Mean $\Longleftrightarrow$ Information Density

$$\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x}, \boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})$$

Conditional Mean $\Longleftrightarrow$ Information Density

$$\boxed{\mathsf{R}(\boldsymbol{x},\boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x},\boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})}$$

1. $[\mu_{X\leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) = \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x},Z)\right]$

Conditional Mean $\Longleftrightarrow$ Information Density

$$\boxed{\mathsf{R}(\boldsymbol{x},\boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x},\boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})}$$

1. $[\mu_{X\leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}[g(Z)|X](\boldsymbol{x}) = \mathbb{E}_{P_Z}[g(Z)\mathsf{R}(\boldsymbol{x},Z)]$

Conditional Mean $\Longleftrightarrow$ Information Density

$$\boxed{\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x}, \boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})}$$

1. $[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}[g(Z)|X](\boldsymbol{x}) = \mathbb{E}_{P_Z}[g(Z)\mathsf{R}(\boldsymbol{x}, Z)]$

Conditional Mean $\Longleftrightarrow$ Information Density

$$\boxed{\mathsf{R}(\boldsymbol{x},\boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x},\boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})}$$

1. $[\mu_{X\leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) = \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x},Z)\right]$
2.

Conditional Mean $\Longleftrightarrow$ Information Density

$$\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x}, \boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})$$

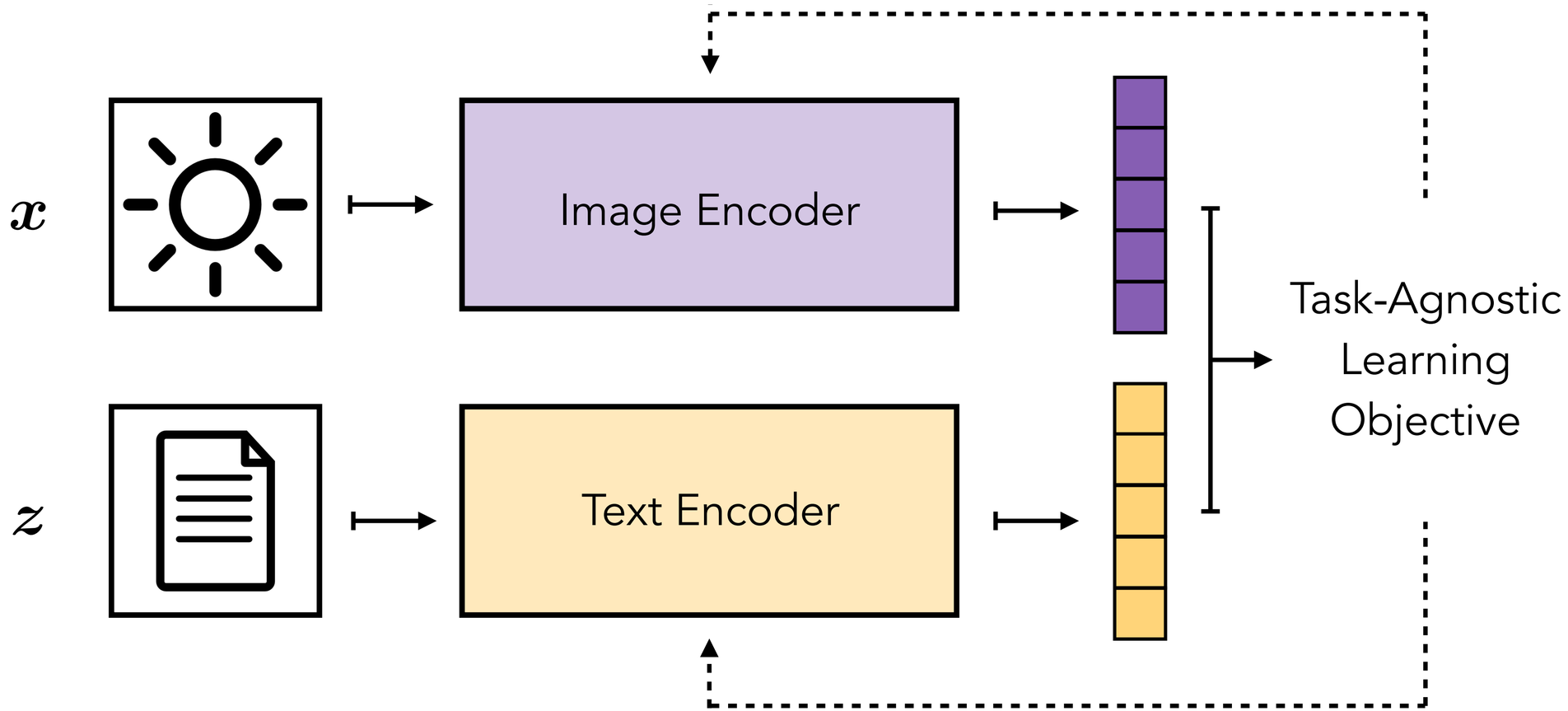
1. $[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) = \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x}, Z)\right]$
2. $I(X, Z) = \mathbb{E}_{P_{X,Z}}\left[(\mathsf{R}(X, Z) - 1)^2\right]$

Conditional Mean $\Longleftrightarrow$ Information Density

$$\boxed{\mathsf{R}(\boldsymbol{x},\boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x},\boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})}$$

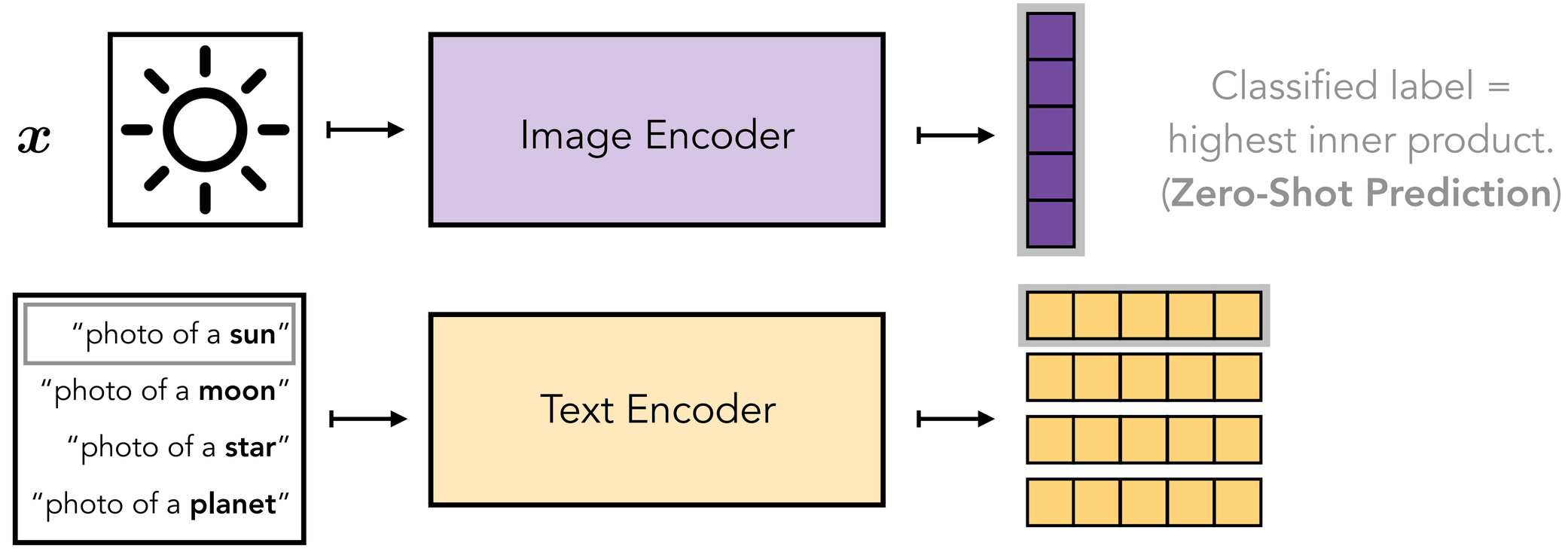
1. $[\mu_{X\leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[g(Z)|X\right](\boldsymbol{x}) = \mathbb{E}_{P_Z}\left[g(Z)\mathsf{R}(\boldsymbol{x},Z)\right]$
2. $I(X,Z) = \mathbb{E}_{P_{X,Z}}\left[(\mathsf{R}(X,Z)-1)^2\right] = \sum_{i=2}^{\infty} s_i^2$

Foundation Modeling



Pre-Training

Foundation Modeling



Idea: Convert labels into pseudo-captions (prompts)

Evaluation

Preliminaries

Balanced Pre-Training

Zero-Shot Prediction

Conclusion

Lang Liu
University of
Washington

Soumik Pal
University of
Washington

Zaid Harchaoui
University of
Washington

The Benefits of Balance: From Information Projections to Variance Reduction

Lang Liu* **Ronak Mehta*** **Soumik Pal** **Zaid Harchaoui**

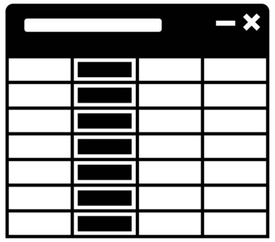
University of Washington

NeurIPS 2024

Motivating Questions

Foundation Modeling

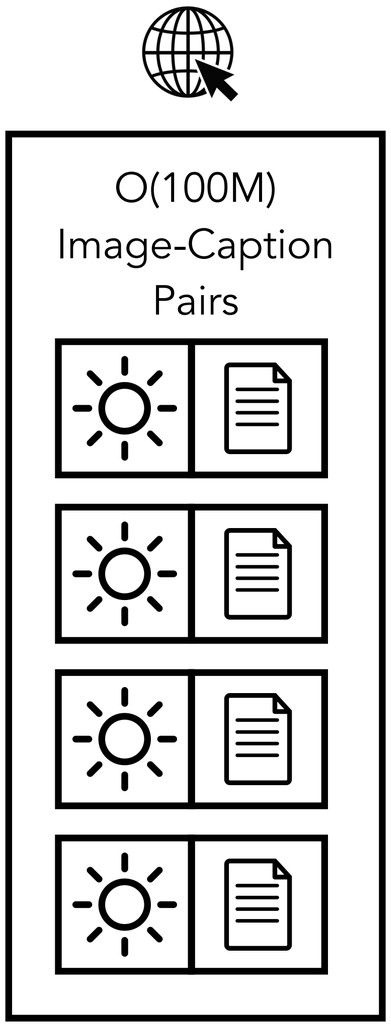
What is the statistical effect of **balancing** methods on the pre-training and the resulting foundation model?



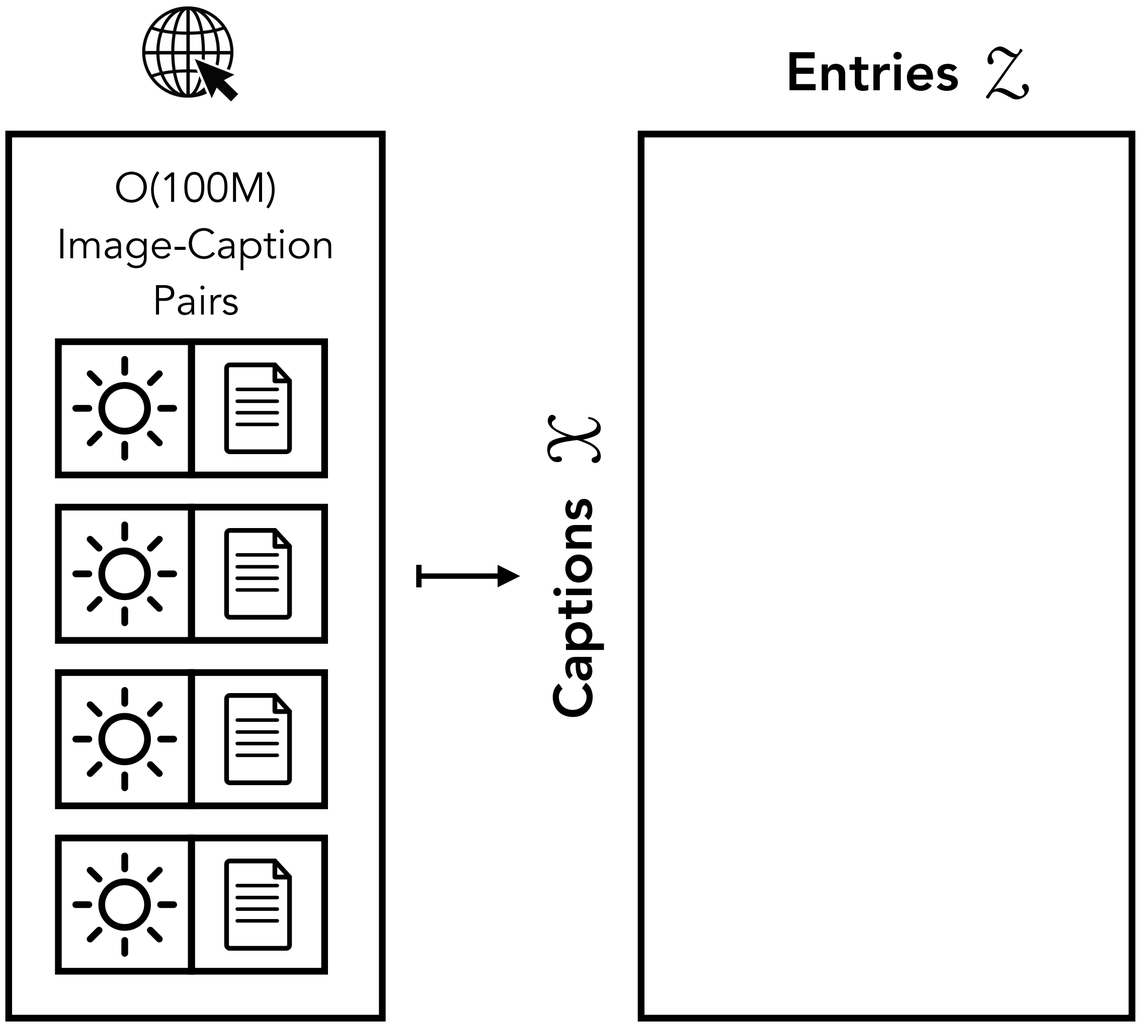
Zero-Shot Prediction

How do we analyze prompt-based **zero-shot prediction** as a statistical estimator, such as its comparison to direct supervision?

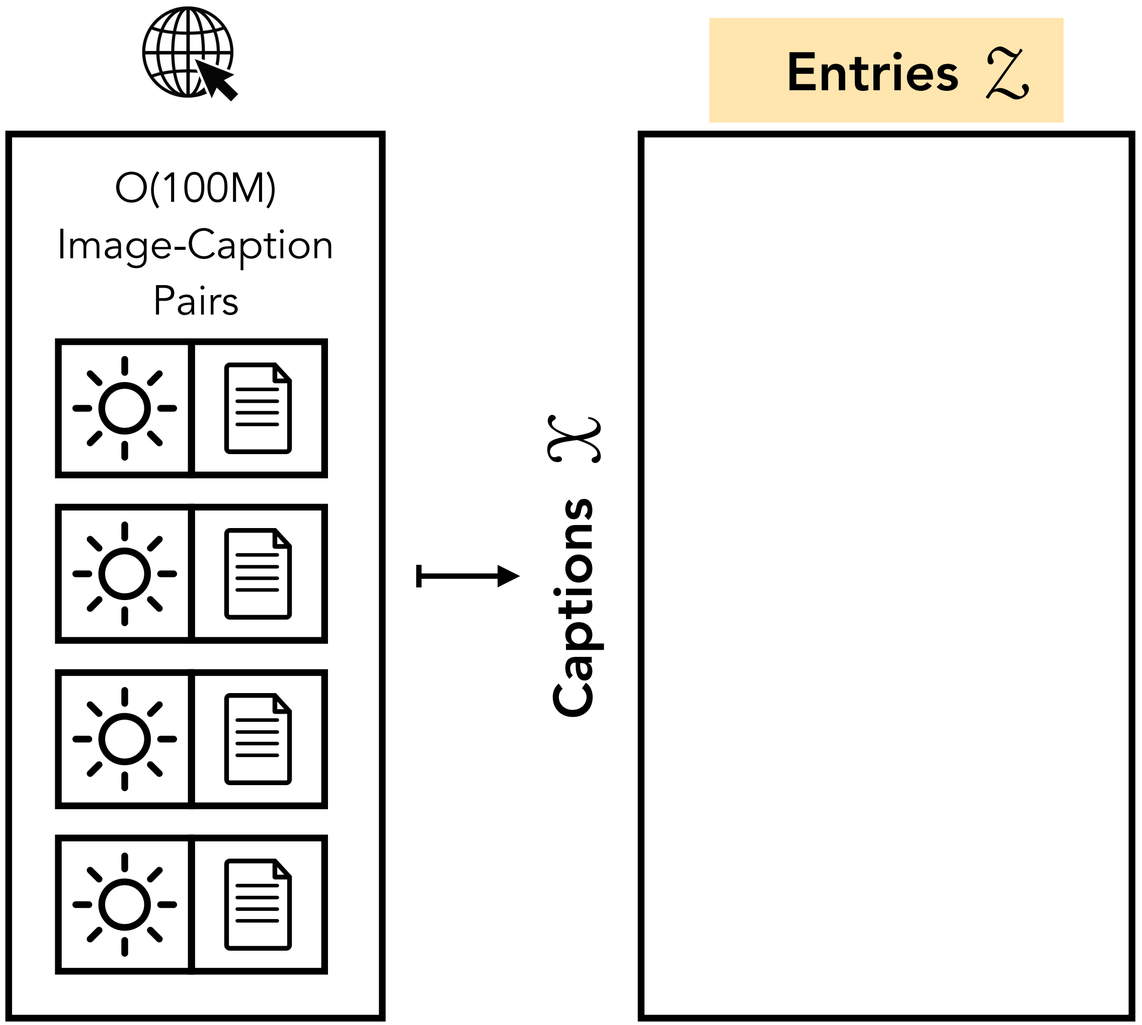
Pre-Training Data Curation: Balancing Keyword Distributions



Pre-Training Data Curation: Balancing Keyword Distributions



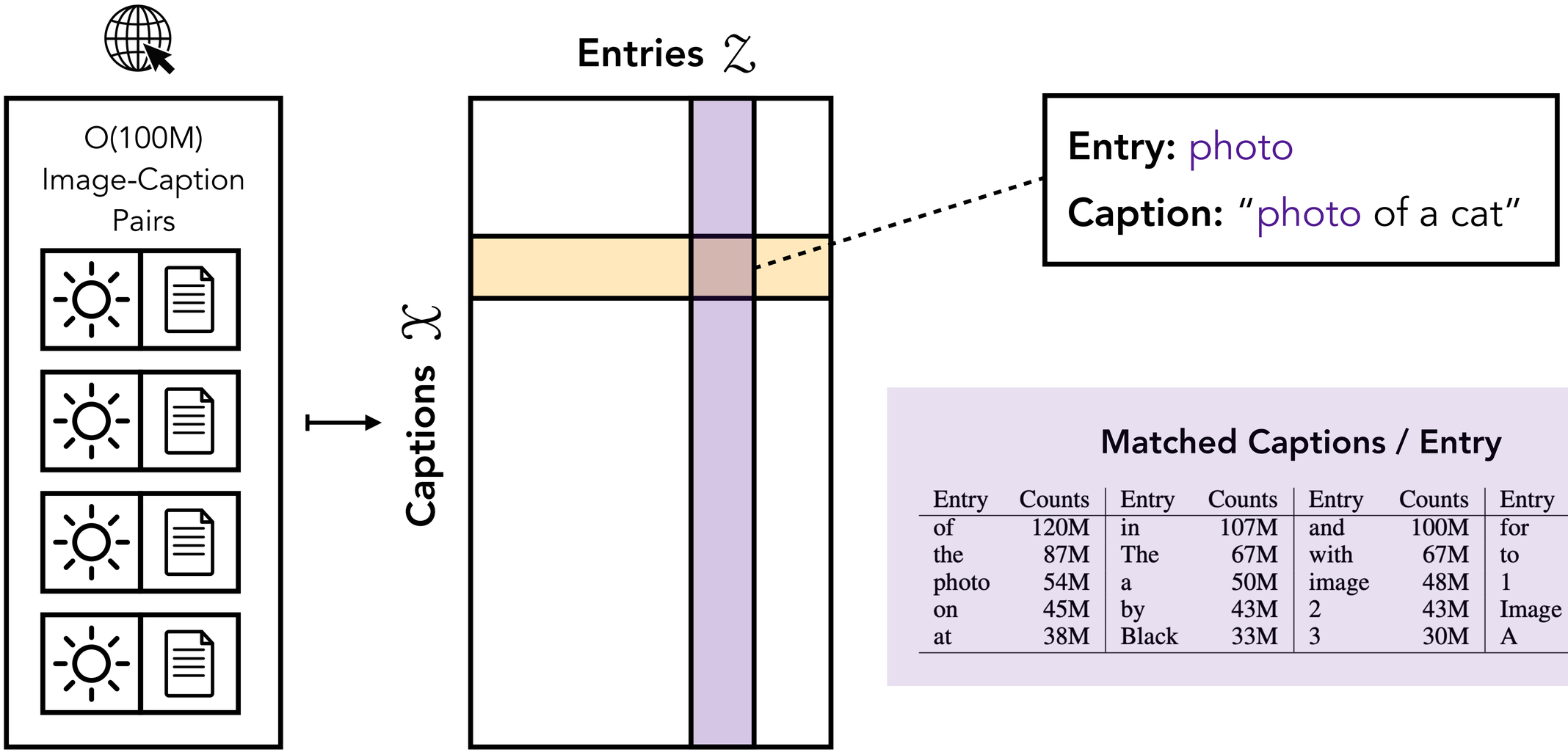
Pre-Training Data Curation: Balancing Keyword Distributions



Matched Captions / Entry

Entry	Counts	Entry	Counts	Entry	Counts	Entry	Counts
of	120M	in	107M	and	100M	for	89M
the	87M	The	67M	with	67M	to	61M
photo	54M	a	50M	image	48M	1	47M
on	45M	by	43M	2	43M	Image	39M
at	38M	Black	33M	3	30M	A	29M

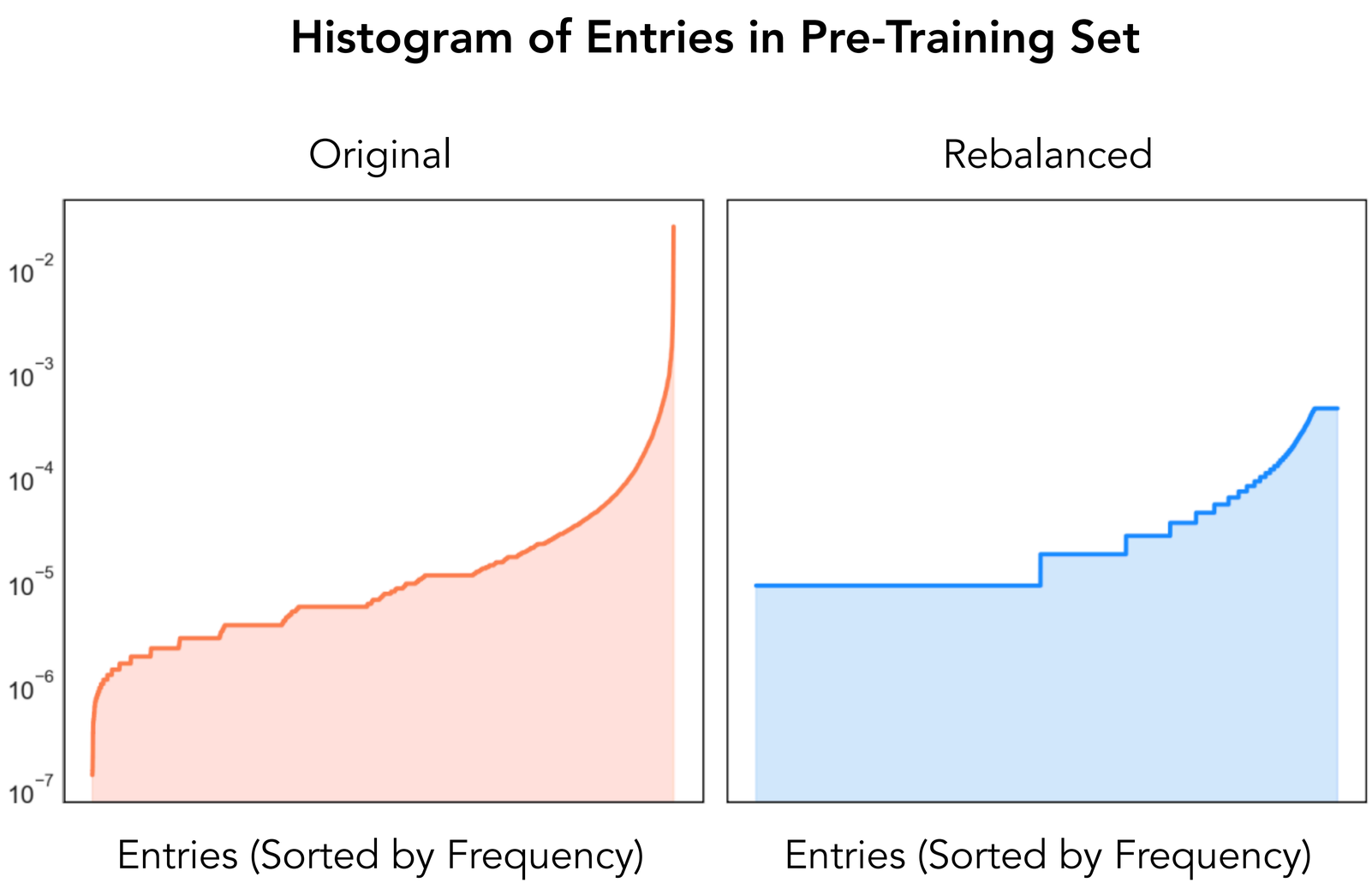
Pre-Training Data Curation: Balancing Keyword Distributions



Matched Captions / Entry

Entry	Counts	Entry	Counts	Entry	Counts	Entry	Counts
of	120M	in	107M	and	100M	for	89M
the	87M	The	67M	with	67M	to	61M
photo	54M	a	50M	image	48M	1	47M
on	45M	by	43M	2	43M	Image	39M
at	38M	Black	33M	3	30M	A	29M

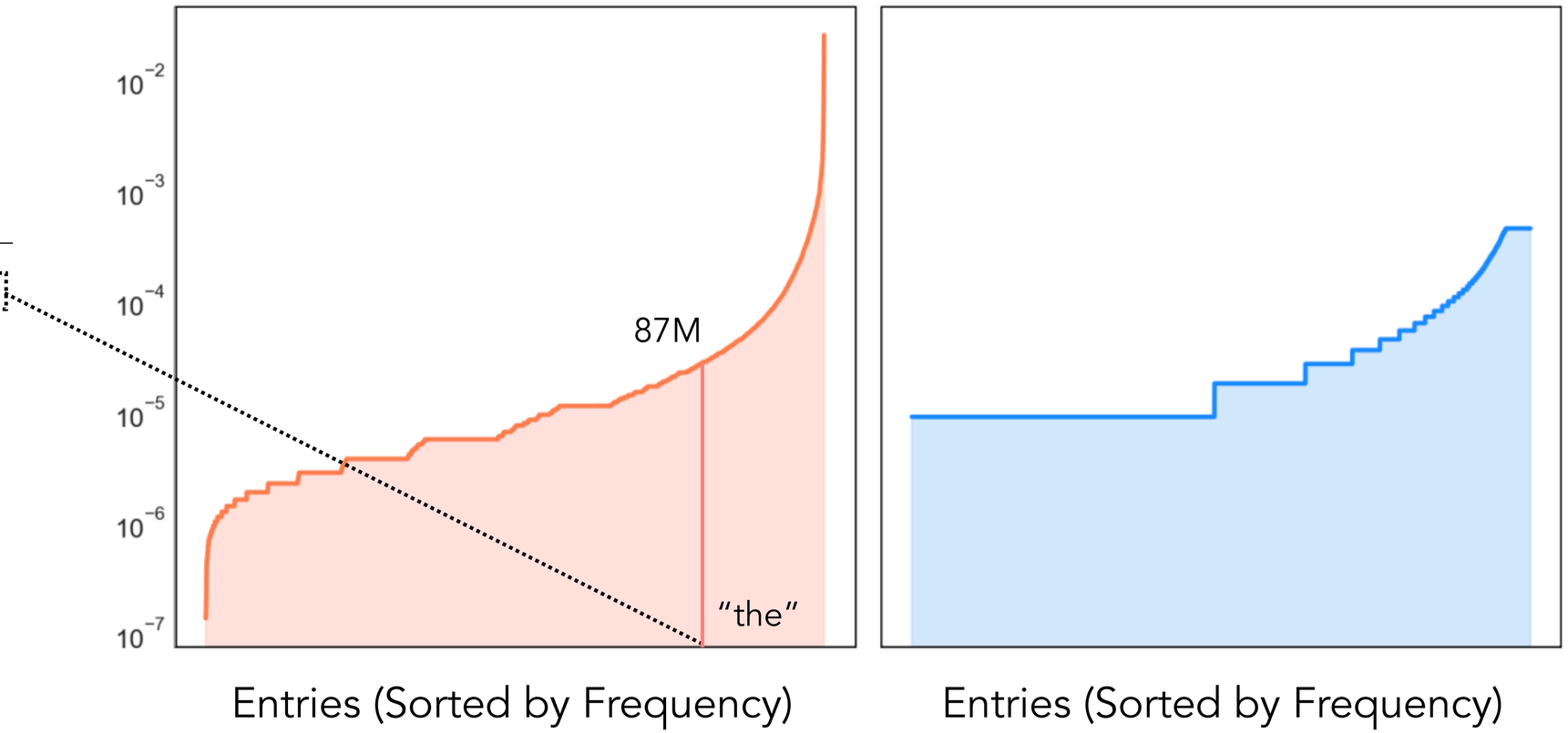
Pre-Training Data Curation: Balancing Keyword Distributions



Pre-Training Data Curation: Balancing Keyword Distributions

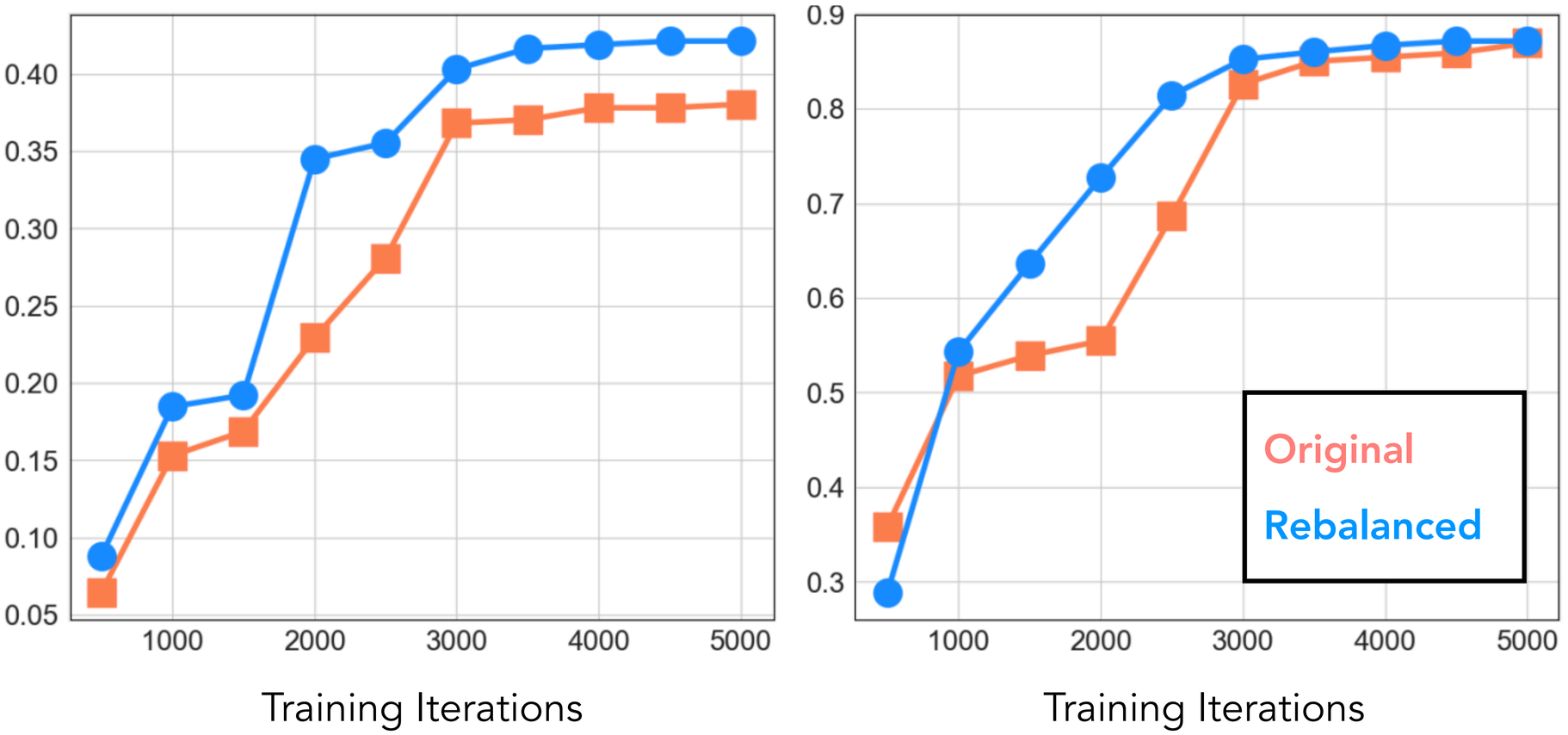
Histogram of Entries in Pre-Training Set

Entry	Counts
of	120M
the	87M
photo	54M
on	45M
at	38M



Pre-Training Data Curation: Balancing Keyword Distributions

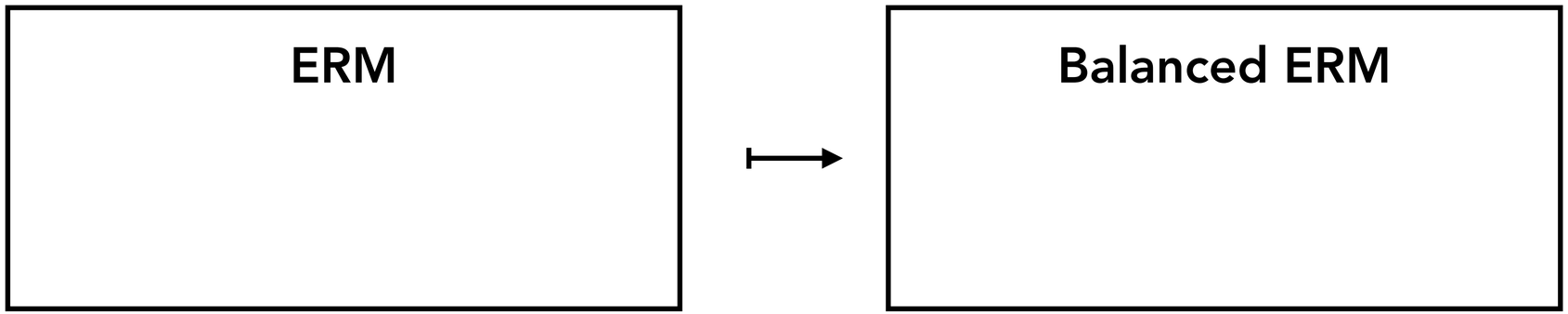
Zero-Shot Accuracy of Models with Different Pre-Training Data



Pre-Training Data Curation: Balancing Keyword Distributions

How should we interpret this empirically effective procedure statistically?

Empirical Risk Minimization with **Balancing**



$$(X_1, Z_1), \ldots, (X_n, Z_n) \overset{\text{i.i.d}}{\sim} P$$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Z_i)}$$

$h_\theta(\boldsymbol{x}, \boldsymbol{z}) =$ loss on pair $(\boldsymbol{x}, \boldsymbol{z})$ of model with parameter θ

Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Balanced ERM

$$(X_1, Z_1), \ldots, (X_n, Z_n) \overset{\text{i.i.d}}{\sim} P$$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Z_i)}$$

$h_\theta(\boldsymbol{x}, \boldsymbol{z}) =$ loss on pair $(\boldsymbol{x}, \boldsymbol{z})$ of model with parameter θ

Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Balanced ERM

$$(X_1, Z_1), \ldots, (X_n, Z_n) \overset{\text{i.i.d}}{\sim} P$$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Z_i)}$$

$$h_\theta(\boldsymbol{x}, \boldsymbol{z}) = \text{loss on pair } (\boldsymbol{x}, \boldsymbol{z}) \text{ of model with parameter } \theta$$

known marginals (P_X, P_Z)

Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Balanced ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^\star} \left[h_\theta(X, Z) \right]$$

$$(X_1, Z_1), \ldots, (X_n, Z_n) \overset{\text{i.i.d}}{\sim} P$$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Z_i)}$$

$h_\theta(\boldsymbol{x}, \boldsymbol{z}) =$ loss on pair $(\boldsymbol{x}, \boldsymbol{z})$ of model with parameter θ

known marginals (P_X, P_Z)

$P_n^\star =$ A distribution on $\mathcal{X} \times \mathcal{Z}$ that is close to P_n, but agrees with the known marginals.

Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Balanced ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^\star} \left[h_\theta(X, Z) \right]$$

$$(X_1, Z_1), \ldots, (X_n, Z_n) \overset{\text{i.i.d}}{\sim} P$$

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Z_i)}$$

$$h_\theta(\boldsymbol{x}, \boldsymbol{z}) = \text{loss on pair } (\boldsymbol{x}, \boldsymbol{z}) \text{ of model with parameter } \theta$$

known marginals (P_X, P_Z)

$$P_n^\star = \underset{Q \in \mathrm{Coup}(P_X, P_Z)}{\arg\min} \mathrm{KL}(Q \| P_n)$$

$$\mathrm{Coup}(P_X, P_Z) = \{Q : Q_X = P_X, Q_Z = P_Z\}$$

How do we perform **balancing** in practice?

(Sinkhorn Iterations, Iterative Proportional Fitting, Raking Ratio Estimation)

How do we perform **balancing** in practice?

(Sinkhorn Iterations, Iterative Proportional Fitting, Raking Ratio Estimation)

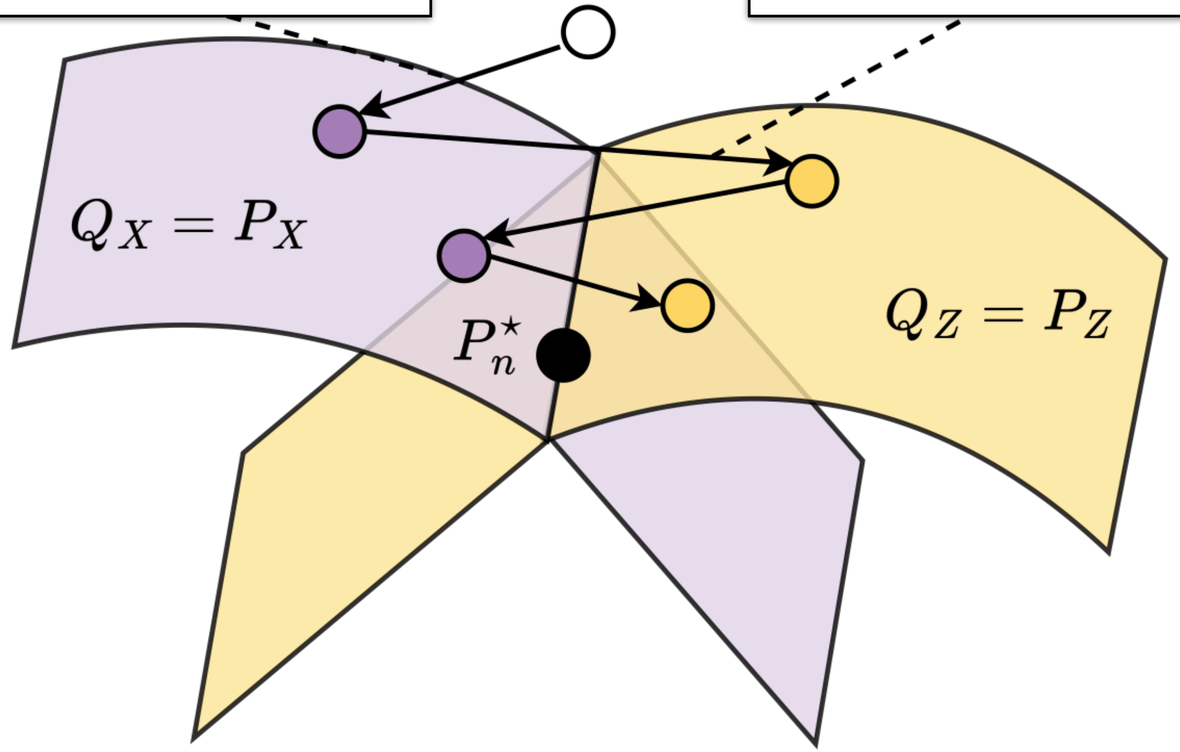
Odd Iterations

$$P_n^{(k)} = \underset{Q:Q_X=P_X}{\arg\min}\ \mathrm{KL}(Q\|P_n)$$

$P_n^{(0)} = P_n$

Even Iterations

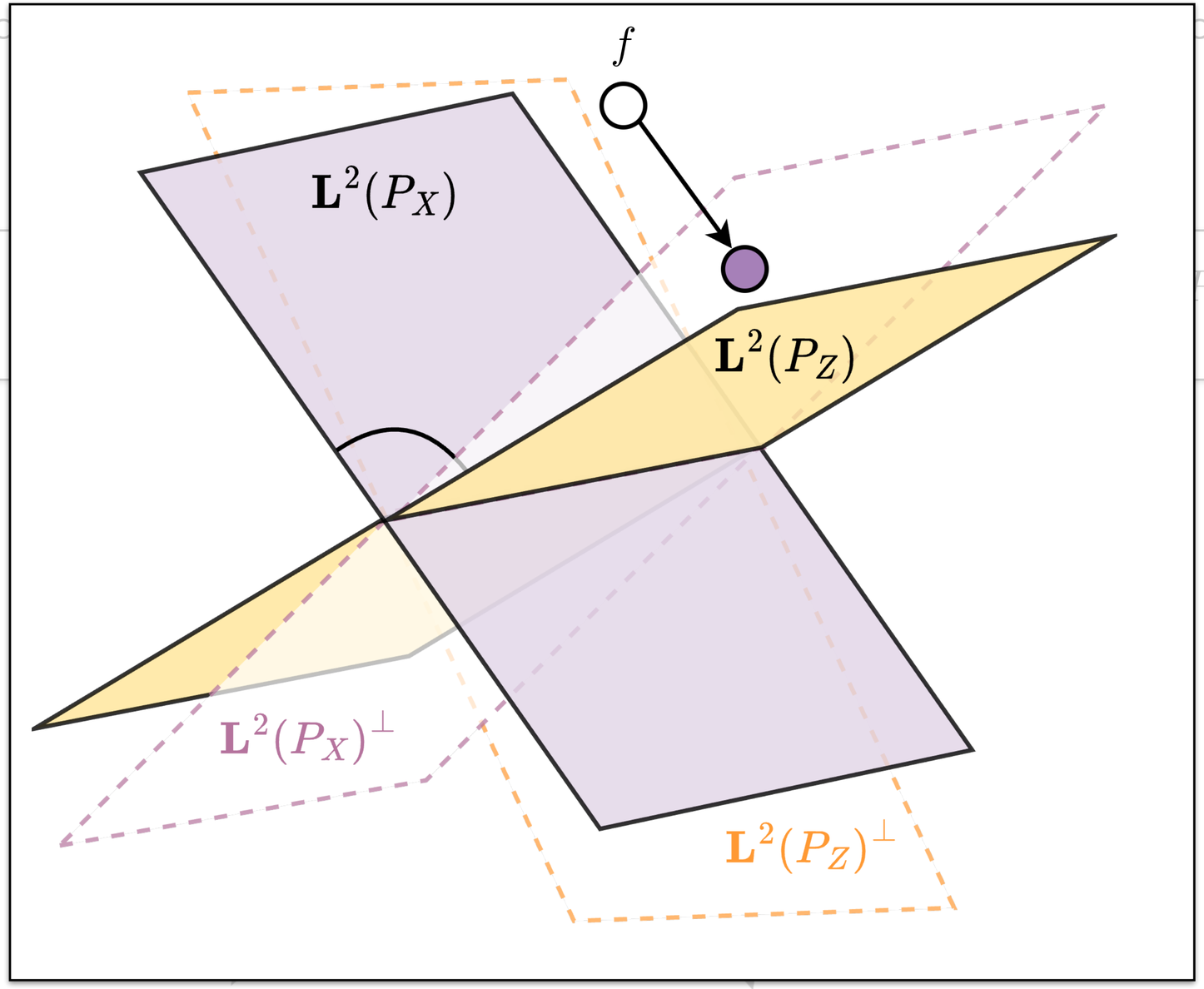
$$P_n^{(k)} = \underset{Q:Q_Z=P_Z}{\arg\min}\ \mathrm{KL}(Q\|P_n)$$



How do we perform **balancing** in practice?

(Sinkho … on)

$$P_n^{(k)} = \quad \ldots \quad P_n)$$



How do we perform **balancing** in practice?

(Sinkhorn Iterations, Iterative Proportional Fitting, Raking Ratio Estimation)

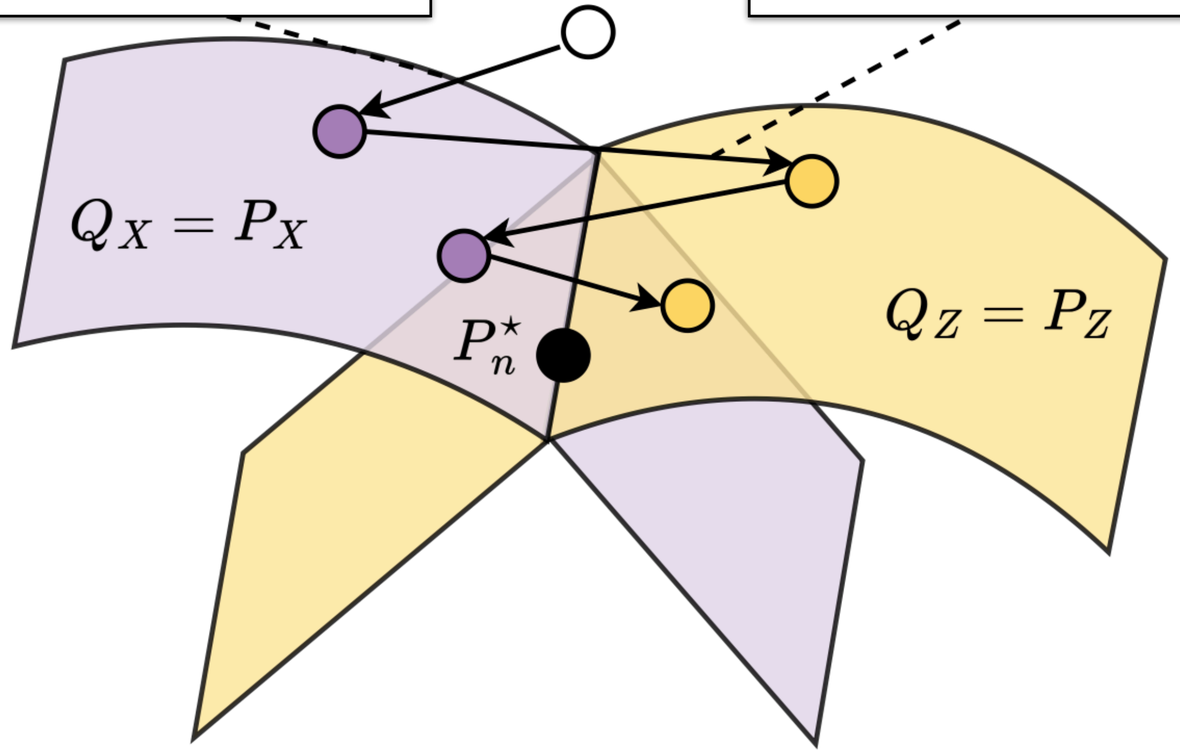
Odd Iterations

$$P_n^{(k)} = \underset{Q:Q_X=P_X}{\arg\min}\ \mathrm{KL}(Q\|P_n)$$

$P_n^{(0)} = P_n$

Even Iterations

$$P_n^{(k)} = \underset{Q:Q_Z=P_Z}{\arg\min}\ \mathrm{KL}(Q\|P_n)$$



Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Rebalanced ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^\star} \left[h_\theta(X, Z) \right]$$

Empirical Risk Minimization with **Balancing**

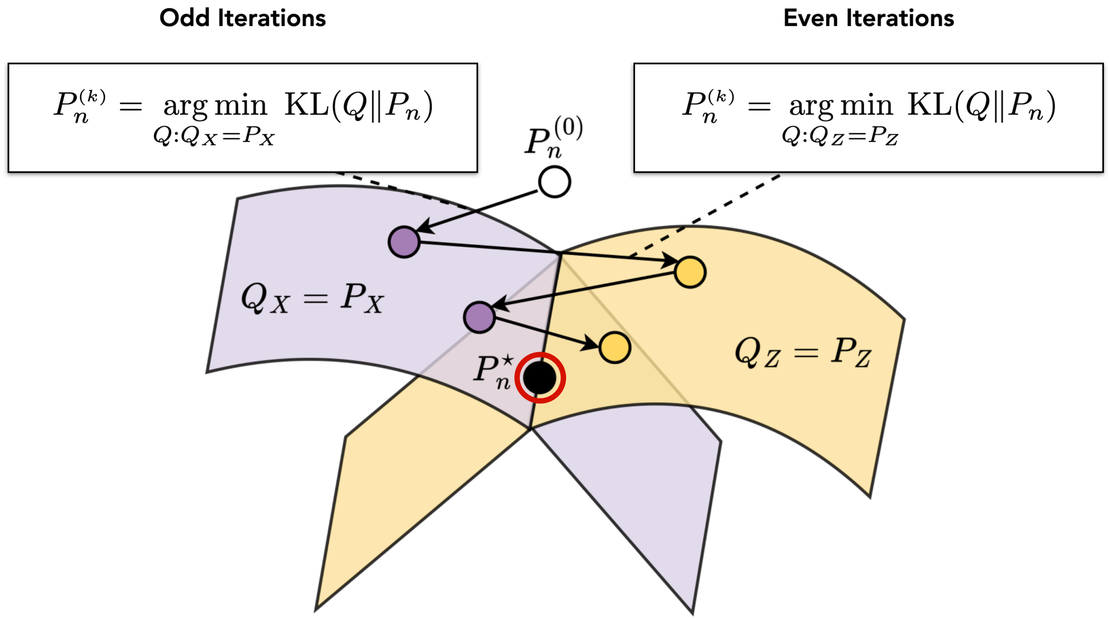
ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Rebalanced ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^\star} \left[h_\theta(X, Z) \right]$$



Empirical Risk Minimization with **Balancing**

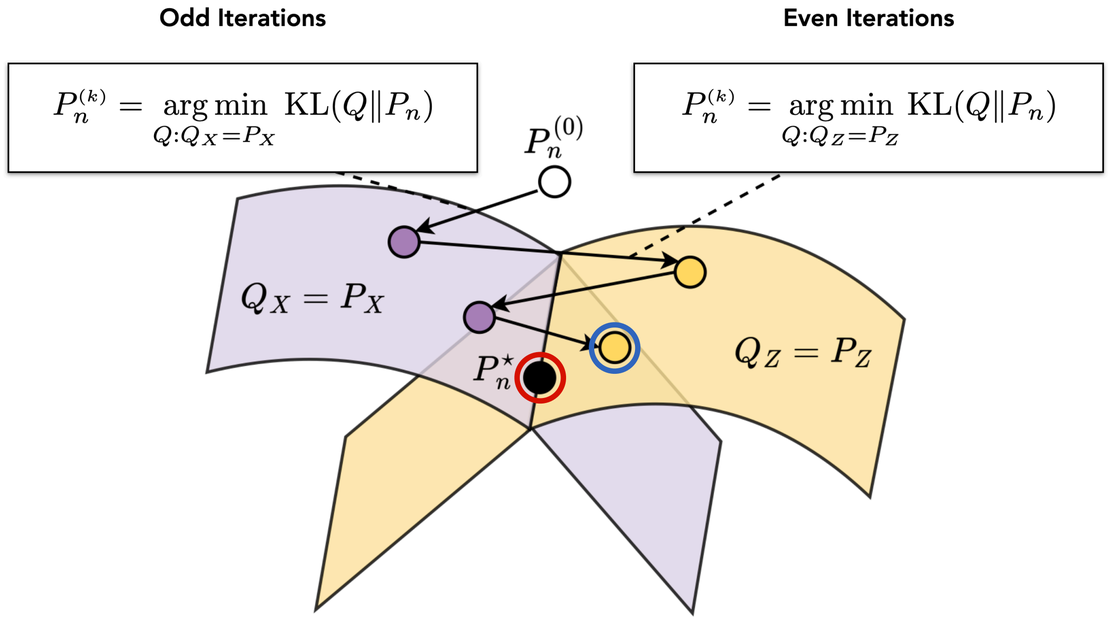
ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Rebalanced ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^{(k)}} \left[h_\theta(X, Z) \right]$$



Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Rebalanced ERM

$$\min_{\theta \in \mathbb{R}^d} \underbrace{\mathbb{E}_{P_n^{(k)}} \left[h_\theta(X, Z) \right]}$$

$$= P_n^{(k)}(h) \stackrel{?}{\approx} P(h)$$

We hide the dependence on θ and consider point-wise estimation for a fixed $h \equiv h_\theta$.

Empirical Risk Minimization with **Balancing**

ERM

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n} \left[h_\theta(X, Z) \right]$$

$\longmapsto$

Rebalanced ERM

$$\min_{\theta \in \mathbb{R}^d} \underbrace{\mathbb{E}_{P_n^{(k)}} \left[h_\theta(X, Z) \right]}$$

$$= P_n^{(k)}(h) \overset{?}{\approx} P(h)$$

We measure the benefit of balancing via **variance/MSE reduction** for estimating the expectation of a fixed test function.

$$\mathbb{E}_P \left[\left(P_n^{(k)}(h) - P(h) \right)^2 \right] \leq \ ? \ < \frac{\mathbb{V}\mathrm{ar}(h)}{n}$$

Empirical Risk Minimization with **Balancing**

Conditional Mean

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P) \to \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} h](\boldsymbol{x}) = \mathbb{E}_P \left[h(X, Z) | X \right](\boldsymbol{x})$$

Conditional Centering

$$\mu^{\perp}_{X \leftarrow Z} : \mathbf{L}^2(P) \to \mathbf{L}^2(P)$$

$$[\mu^{\perp}_{X \leftarrow Z} h](\boldsymbol{x}, \boldsymbol{z}) = h(\boldsymbol{x}, \boldsymbol{z}) - [\mu_{X \leftarrow Z} h](\boldsymbol{x})$$

$$= P_n^{(k)}(h) \approx \dot{P}(h)$$

We measure the benefit of balancing via **variance/MSE reduction** for estimating the expectation of a fixed test function.

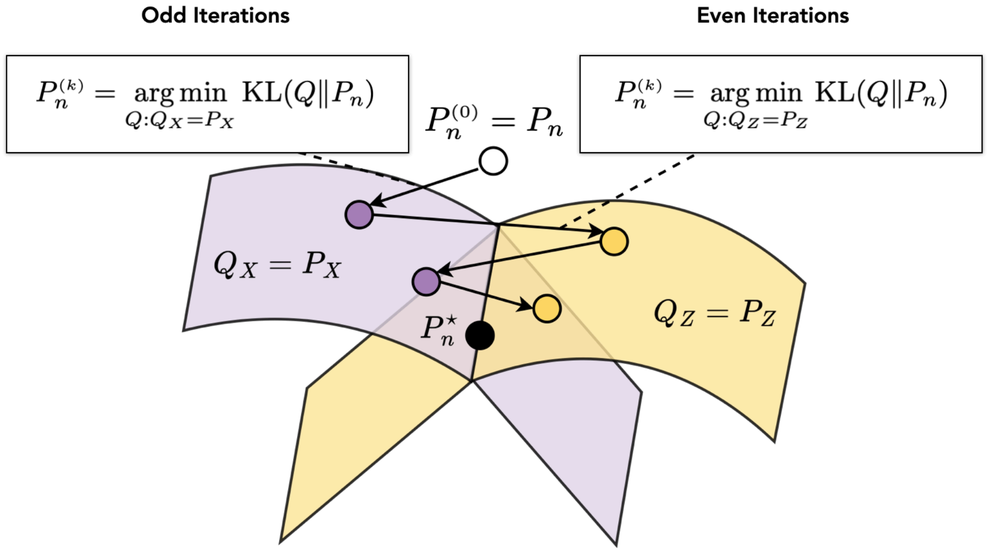
$$\mathbb{E}_P \left[(P_n^{(k)}(h) - P(h))^2 \right] \leq \ ? \ < \frac{\mathbb{V}\mathrm{ar}(h)}{n}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

Theorem (Liu, M., Pal, Harchaoui)

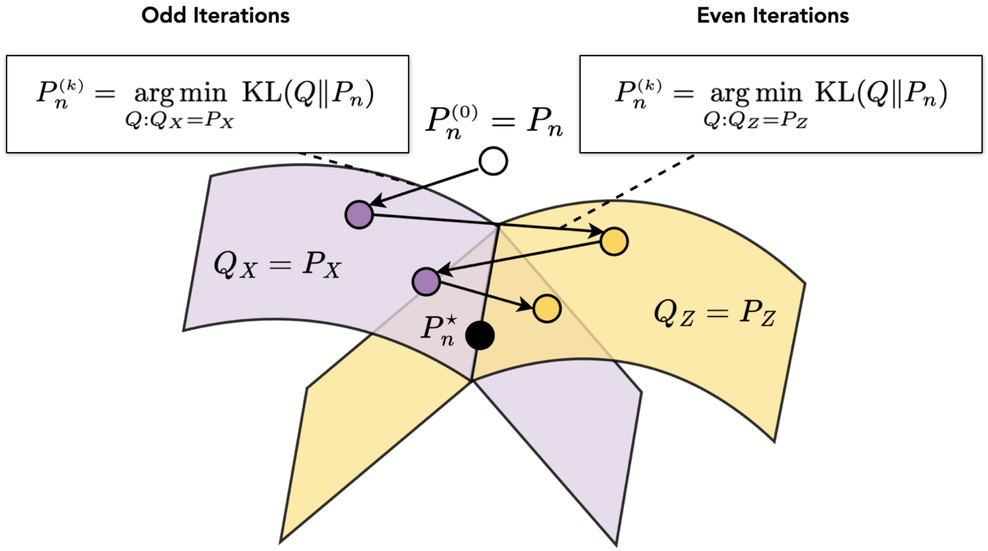
$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

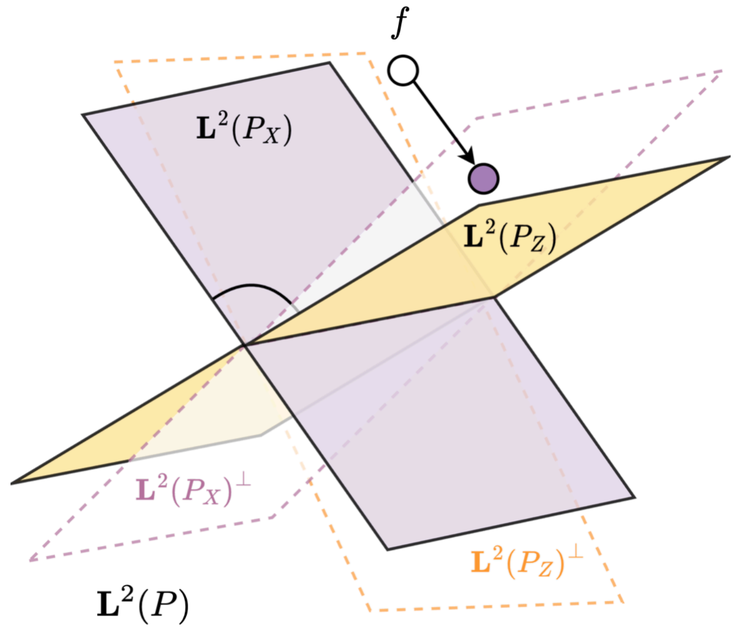


Nonlinear Projections ↦ Linear Projections ↦ Variance Reduction

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$



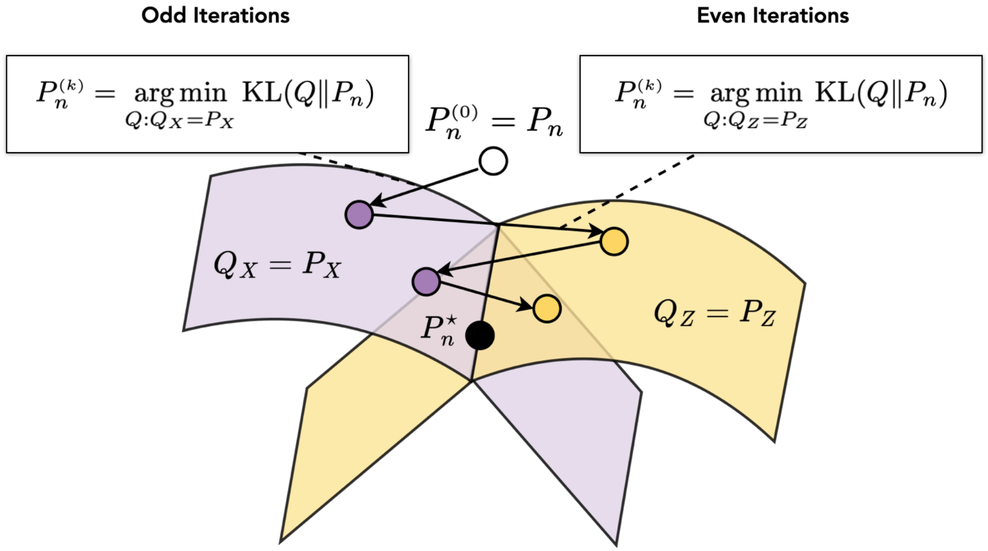


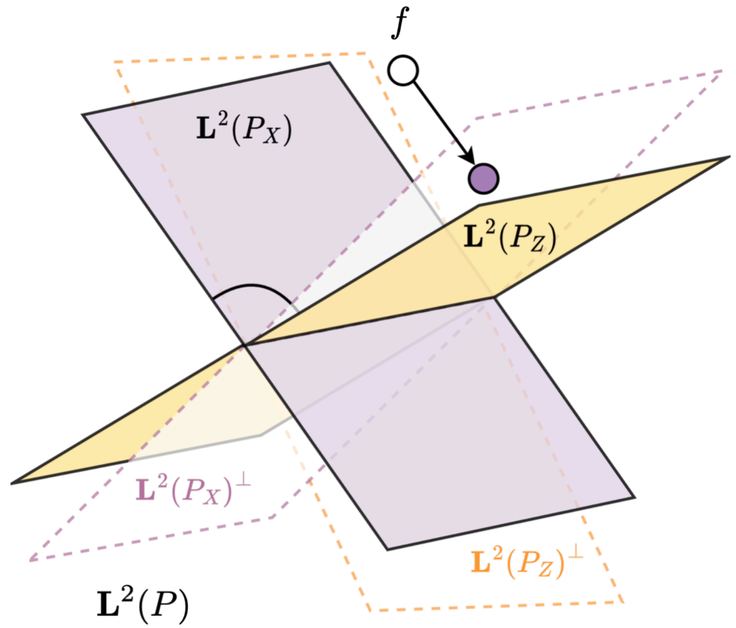
Nonlinear Projections ↦ Linear Projections ↦ Variance Reduction

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

(k times)





Nonlinear Projections ↦ Linear Projections ↦ Variance Reduction

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z} h) + [P_n^{(k)} - P](\mu_{X\leftarrow Z} h)$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z} h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z} h)}^{=0}$$

The X-marginal of both distributions match (assuming k odd)

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0} \\ &= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h)\end{aligned}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z} h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z} h)}^{=0}$$

$$= \underbrace{[P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z} h)}_{\text{Like LHS, but with a conditional centering}} + \underbrace{[P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z} h)}_{\text{error term}}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0} \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

unroll k times

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0} \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\underbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}_{k \text{ times}} h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

(k times)

$$\begin{aligned}
[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0} \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\underbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h}_{k \text{ times}}) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0} \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\underbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}_{k \text{ times}} h) + \sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z} h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z} h)}^{=0} \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z} h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z} h) \\
&= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h) + \textstyle\sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h)
\end{aligned}$$

$$[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h) = \sum_{\boldsymbol{x},\boldsymbol{z}} \underbrace{\left(\frac{P_X(\boldsymbol{x})}{P_{n,X}^{(\ell-1)}(\boldsymbol{x})} - 1\right)}_{\substack{\text{unmatched marginal} \\ \text{error } O(n^{-1/2})}} [\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h](\boldsymbol{x},\boldsymbol{z}) P_n^{(\ell-1)}(\boldsymbol{x},\boldsymbol{z})$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\overbrace{\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}}^{k \text{ times}} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= [P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h) + \overbrace{[P_n^{(k)} - P](\mu_{X\leftarrow Z}h)}^{=0} \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

$$[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h) = \sum_{\boldsymbol{x},\boldsymbol{z}}\left(\frac{P_X(\boldsymbol{x})}{P_{n,X}^{(\ell-1)}(\boldsymbol{x})} - 1\right)\underbrace{[\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h](\boldsymbol{x},\boldsymbol{z})P_n^{(\ell-1)}(\boldsymbol{x},\boldsymbol{z})}_{\text{control using similar recursion } O(n^{-1/2})}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= \overbrace{[P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h)}^{=0} + [P_n^{(k)} - P](\mu_{X\leftarrow Z}h) \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

What is the actual reduction in variance?

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$\begin{aligned}
[P_n^{(k)} - P](h) &= \overbrace{[P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z}h)}^{=0} + [P_n^{(k)} - P](\mu_{X\leftarrow Z}h) \\
&= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z}h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z}h) \\
&= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)
\end{aligned}$$

$$\begin{aligned}
\mu^{\perp}_{X\leftarrow Z} &= I - \mu_{X\leftarrow Z} \\
\mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} &= I - \mu_{X\leftarrow Z} - \mu_{Z\leftarrow X} + \mu_{Z\leftarrow X}\mu_{X\leftarrow Z} \\
\mu^{\perp}_{X\leftarrow Z}\mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} &= I - \mu_{X\leftarrow Z} - \mu_{Z\leftarrow X} + \mu_{Z\leftarrow X}\mu_{X\leftarrow Z} + \mu_{X\leftarrow Z}\mu_{Z\leftarrow X} - \mu_{X\leftarrow Z}\mu_{Z\leftarrow X}\mu_{X\leftarrow Z}
\end{aligned}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = \overbrace{[P_n^{(k)} - P](\mu^{\perp}_{x\leftarrow z} h)}^{=0} + [P_n^{(k)} - P](\mu_{x\leftarrow z} h)$$
$$= [P_n^{(k-1)} - P](\mu^{\perp}_{x\leftarrow z} h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{x\leftarrow z} h)$$
$$= [P_n^{(0)} - P](\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h)$$

$$\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z}$$
$$\mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z} - \mu_{z\leftarrow x} + \mu_{z\leftarrow x}\mu_{x\leftarrow z}$$
$$\mu^{\perp}_{x\leftarrow z}\mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z} - \mu_{z\leftarrow x} + \mu_{z\leftarrow x}\mu_{x\leftarrow z} + \mu_{x\leftarrow z}\mu_{z\leftarrow x} - \mu_{x\leftarrow z}\mu_{z\leftarrow x}\mu_{x\leftarrow z}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = \overbrace{[P_n^{(k)} - P](\mu^{\perp}_{x\leftarrow z} h)}^{=0} + [P_n^{(k)} - P](\mu_{x\leftarrow z} h)$$
$$= [P_n^{(k-1)} - P](\mu^{\perp}_{x\leftarrow z} h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{x\leftarrow z} h)$$
$$= [P_n^{(0)} - P](\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h)$$

$$\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z}$$
$$\mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z} - \mu_{z\leftarrow x} + \mu_{z\leftarrow x}\mu_{x\leftarrow z}$$
$$\mu^{\perp}_{x\leftarrow z}\mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z} - \mu_{z\leftarrow x} + \mu_{z\leftarrow x}\mu_{x\leftarrow z} + \mu_{x\leftarrow z}\mu_{z\leftarrow x} - \mu_{x\leftarrow z}\mu_{z\leftarrow x}\mu_{x\leftarrow z}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = \overbrace{[P_n^{(k)} - P](\mu^{\perp}_{x\leftarrow z} h)}^{=0} + [P_n^{(k)} - P](\mu_{x\leftarrow z} h)$$
$$= [P_n^{(k-1)} - P](\mu^{\perp}_{x\leftarrow z} h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{x\leftarrow z} h)$$
$$= [P_n^{(0)} - P](\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h) + \textstyle\sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\dots \mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} h)$$

$$\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z}$$
$$\mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z} - \mu_{z\leftarrow x} + \mu_{z\leftarrow x}\mu_{x\leftarrow z}$$
$$\mu^{\perp}_{x\leftarrow z}\mu^{\perp}_{z\leftarrow x}\mu^{\perp}_{x\leftarrow z} = I - \mu_{x\leftarrow z} - \mu_{z\leftarrow x} + \mu_{z\leftarrow x}\mu_{x\leftarrow z} + \mu_{x\leftarrow z}\mu_{z\leftarrow x} - \mu_{x\leftarrow z}\mu_{z\leftarrow x}\mu_{x\leftarrow z}$$

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\dots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

$$[P_n^{(k)} - P](h) = \overbrace{[P_n^{(k)} - P](\mu^{\perp}_{X\leftarrow Z} h)}^{=0} + [P_n^{(k)} - P](\mu_{X\leftarrow Z} h)$$
$$= [P_n^{(k-1)} - P](\mu^{\perp}_{X\leftarrow Z} h) + [P_n^{(k)} - P_n^{(k-1)}](\mu^{\perp}_{X\leftarrow Z} h)$$
$$= [P_n^{(0)} - P](\dots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h) + \sum_{\ell=1}^{k}[P_n^{(\ell)} - P_n^{(\ell-1)}](\dots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} h)$$

$$\mu^{\perp}_{X\leftarrow Z} = I - \mu_{X\leftarrow Z}$$
$$\mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} = I - \mu_{X\leftarrow Z} - \mu_{Z\leftarrow X} + \mu_{Z\leftarrow X}\mu_{X\leftarrow Z}$$
$$\mu^{\perp}_{X\leftarrow Z}\mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z} = I - \mu_{X\leftarrow Z} - \mu_{Z\leftarrow X} + \mu_{Z\leftarrow X}\mu_{X\leftarrow Z} + \mu_{X\leftarrow Z}\mu_{Z\leftarrow X} - \mu_{X\leftarrow Z}\mu_{Z\leftarrow X}\underbrace{\mu_{X\leftarrow Z} h}_{\mathbb{E}_{P_{X,Z}}[h(X,Z)|X](\cdot)}$$

heorem (Liu, M., Pal, Harchaoui)

Old Friend: Singular Value Decomposition

Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

$$= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) + \sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h)$$

$$\mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{X \leftarrow Z} \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} + \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} - \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} h$$

$$\mathbb{E}_{P_{X,Z}} [h(X, Z) | X] (\cdot)$$

heorem (Liu, M., Pal, Harchaoui)

Old Friend: Singular Value Decomposition

Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

$$= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) + \sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h)$$

$$\mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{X \leftarrow Z} \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} + \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} - \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} h$$

$$\mathbb{E}_{P_{X,Z}} [h(X, Z)|X] (\cdot)$$

$$\mu_{X \leftarrow Z} h = \sum_i u_i \alpha_i$$

heorem (Liu, M., Pal, Harchaoui)

Old Friend: Singular Value Decomposition

Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \dots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \dots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

$$= [P_n^{(0)} - P](\dots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) + \sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\dots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h)$$

$$\mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{X \leftarrow Z} \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} + \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} - \underbrace{\mu_{X \leftarrow Z} \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} h}$$

$$\mu_{X \leftarrow Z} h = \sum_i s_i u_i \beta_i$$

heorem (Liu, M., Pal, Harchaoui)

Old Friend: Singular Value Decomposition

Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

$$= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) + \sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h)$$

$$\mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{X \leftarrow Z} \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} + \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} - \underbrace{\mu_{X \leftarrow Z} \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} h}$$

$$\mu_{X \leftarrow Z} h = \sum_i s_i^2 u_i \alpha_i$$

heorem (Liu, M., Pal, Harchaoui)

Old Friend: Singular Value Decomposition

Basis of $\mathbf{L}^2(P_X)$: $\alpha_1, \alpha_2, \ldots$

$$\mu_{X \leftarrow Z} \beta_i = s_i \alpha_i$$

Basis of $\mathbf{L}^2(P_Z)$: $\beta_1, \beta_2, \ldots$

$$\mu_{Z \leftarrow X} \alpha_i = s_i \beta_i$$

$$= [P_n^{(0)} - P](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) + \sum_{\ell=1}^{k} [P_n^{(\ell)} - P_n^{(\ell-1)}](\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h)$$

$$\mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z}$$

$$\mu^{\perp}_{X \leftarrow Z} \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} = I - \mu_{X \leftarrow Z} - \mu_{Z \leftarrow X} + \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} + \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} - \mu_{X \leftarrow Z} \mu_{Z \leftarrow X} \mu_{X \leftarrow Z} h$$

$$\mu_{X \leftarrow Z} h = \sum_i s_i^{10} u_i \alpha_i$$

$$u_i = \langle \alpha_i, \mu_{X \leftarrow Z} h \rangle$$

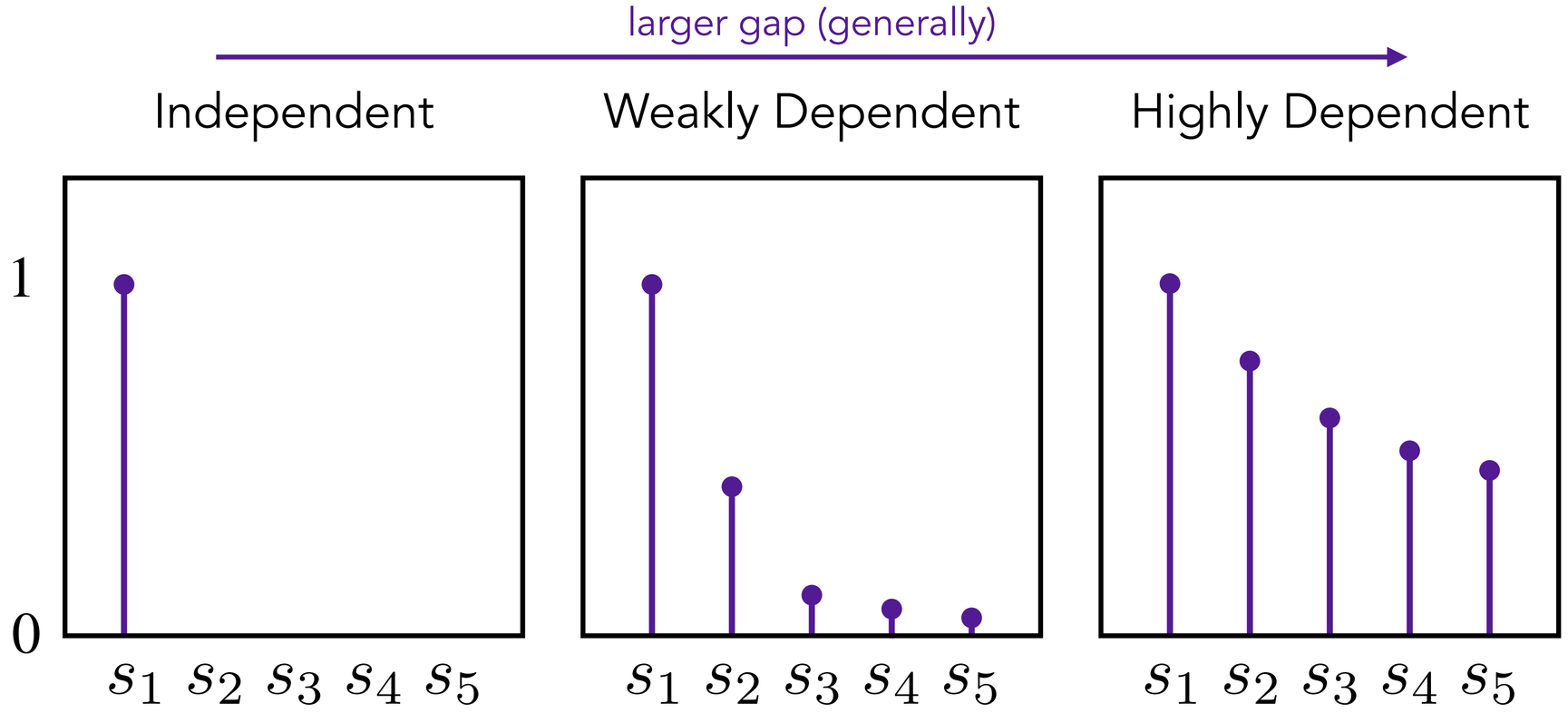
$$v_i = \langle \beta_i, \mu_{Z \leftarrow X} h \rangle$$

$$\mathbb{V}\mathrm{ar}(h) - \mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) = \sum_{i=2,3,\ldots} \frac{u_i^2 + v_i^2 - 2 s_i u_i v_i}{1 - s_i^2}$$

$$u_i = \langle \alpha_i, \mu_{X \leftarrow Z} h \rangle$$

$$v_i = \langle \beta_i, \mu_{Z \leftarrow X} h \rangle$$

$$\mathbb{V}\mathrm{ar}(h) - \mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h) = \sum_{i=2,3,\ldots} \frac{u_i^2 + v_i^2 - 2 s_i u_i v_i}{1 - s_i^2}$$



Research Contributions

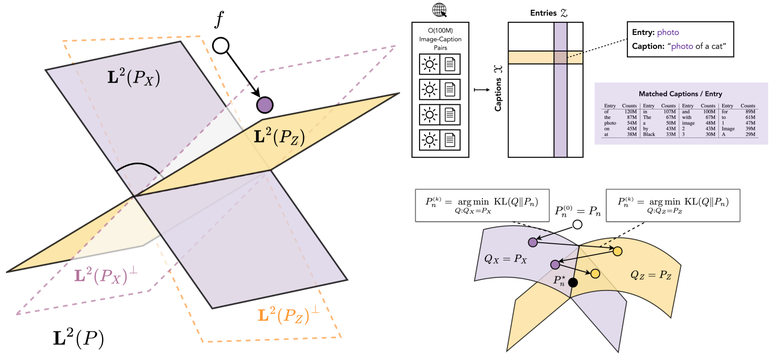
✅

Exact variance reduction analysis of balancing-based mean estimation!

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z\leftarrow X}\mu^{\perp}_{X\leftarrow Z}h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

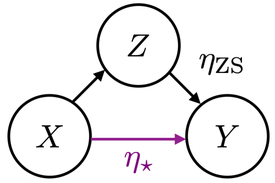
(k times)

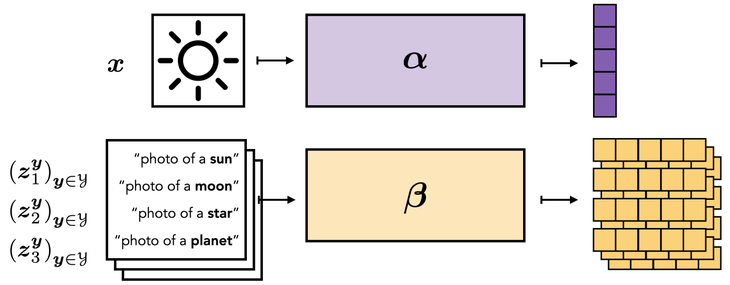


Theoretical framework for obtaining generalization guarantees for two broad classes of zero-shot prediction models!

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$





Preliminaries

Balanced Pre-Training

Zero-Shot Prediction

Conclusion

Zaid Harchaoui
University of
Washington

The Prediction Path to Zero-Shot Generalization

Ronak Mehta [1] **Zaid Harchaoui** [1]

ICML 2025 Spotlight (Top 2.6% of Submissions)

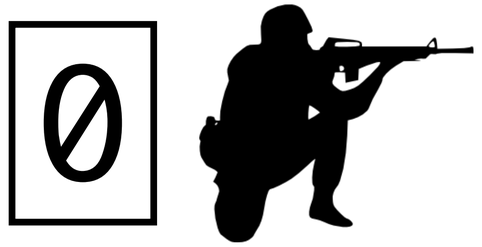
Motivating Questions

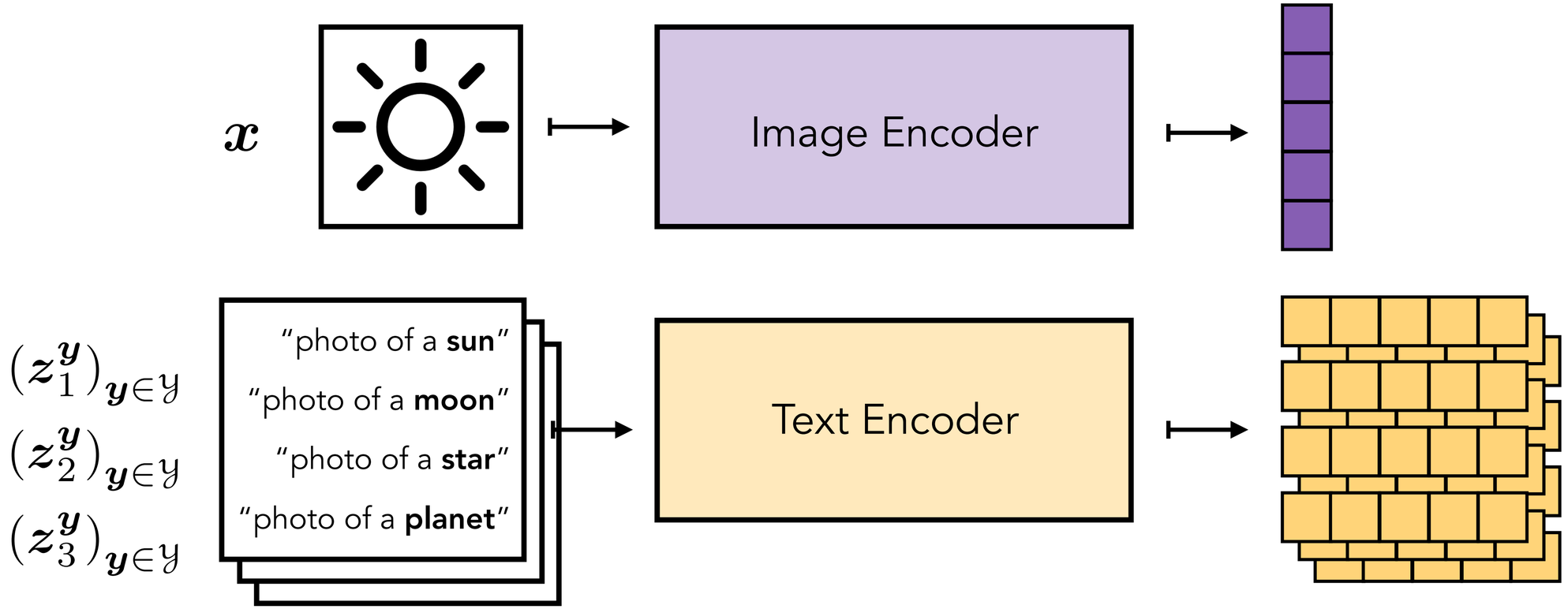
Foundation Modeling

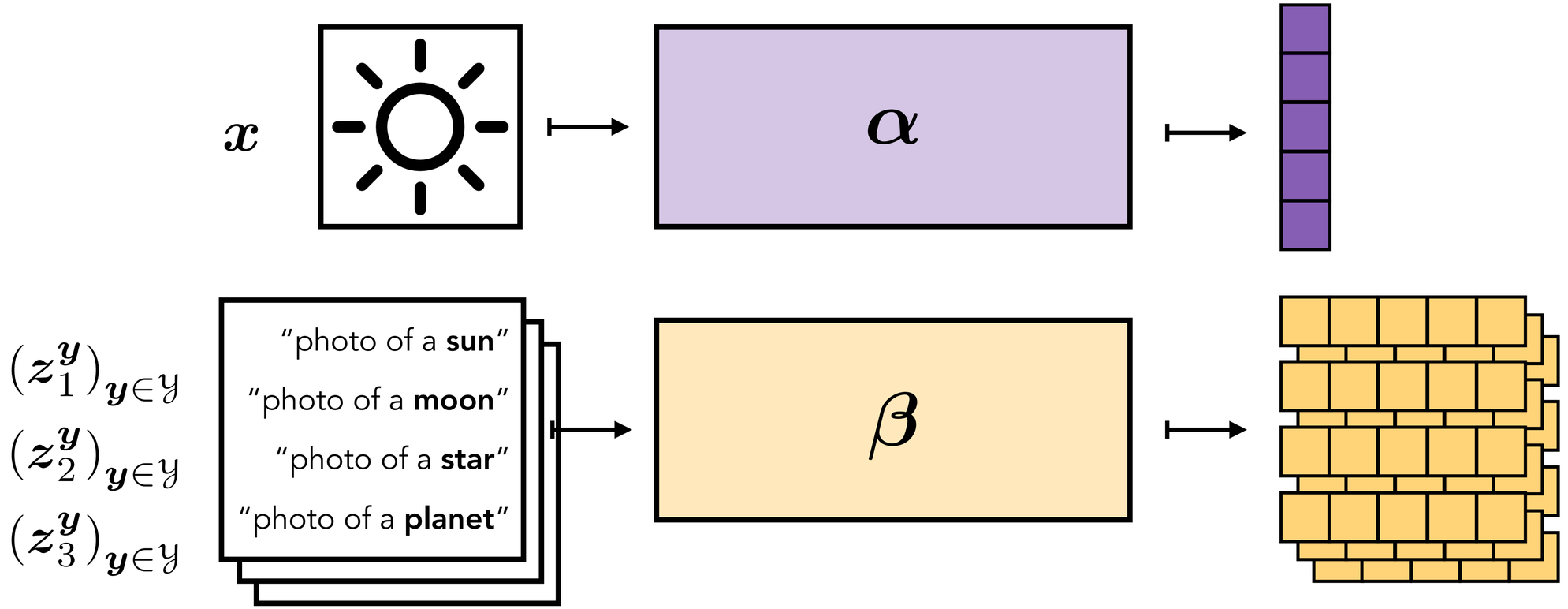
What is the statistical effect of **balancing** methods on the pre-training and the resulting foundation model?

Zero-Shot Prediction

How do we analyze prompt-based **zero-shot prediction** as a statistical estimator, such as its comparison to direct supervision?




x
Image Encoder
$(z_1^y)_{y \in \mathcal{Y}}$
$(z_2^y)_{y \in \mathcal{Y}}$
$(z_3^y)_{y \in \mathcal{Y}}$
"photo of a **sun**"
"photo of a **moon**"
"photo of a **star**"
"photo of a **planet**"
Text Encoder



$$x \mapsto \arg\max_{y\in\mathcal{Y}} \frac{1}{m}\sum_{k=1}^{m}\langle \alpha(x), \beta(z_k^y)\rangle$$

Zero-Shot Prediction: Compatibility of Three Distributions

How should we interpret this procedure from
the perspective of statistical learning theory?

Zero-Shot Prediction: Compatibility of Three Distributions

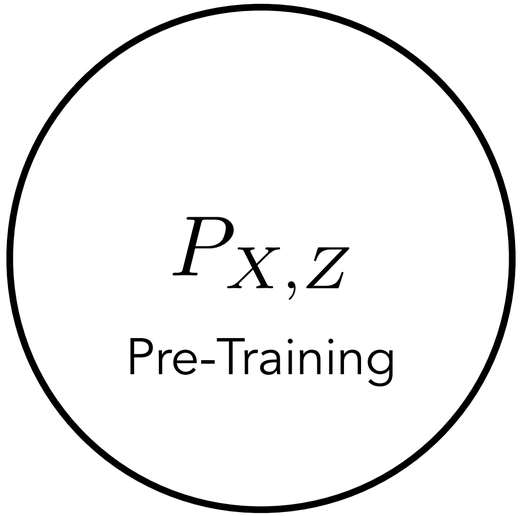
How should we interpret this procedure from the perspective of statistical learning theory?

Step 1: Define the downstream task we wish to solve (no different from supervised learning).

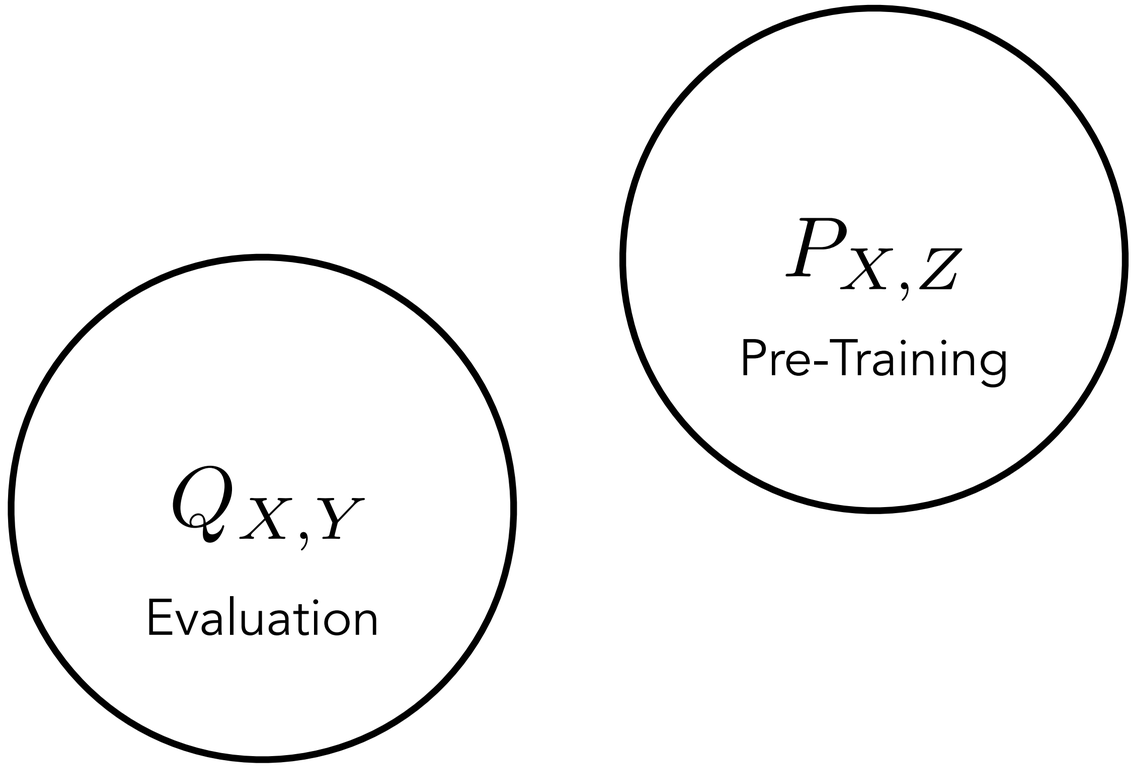
Step 2: Define the population version of the zero-shot prediction procedure and justify it.

Step 3: Provide two estimation frameworks and generic error decompositions that can be used to prove generalization risk bounds.

Zero-Shot Prediction: Compatibility of Three Distributions



Zero-Shot Prediction: Compatibility of Three Distributions



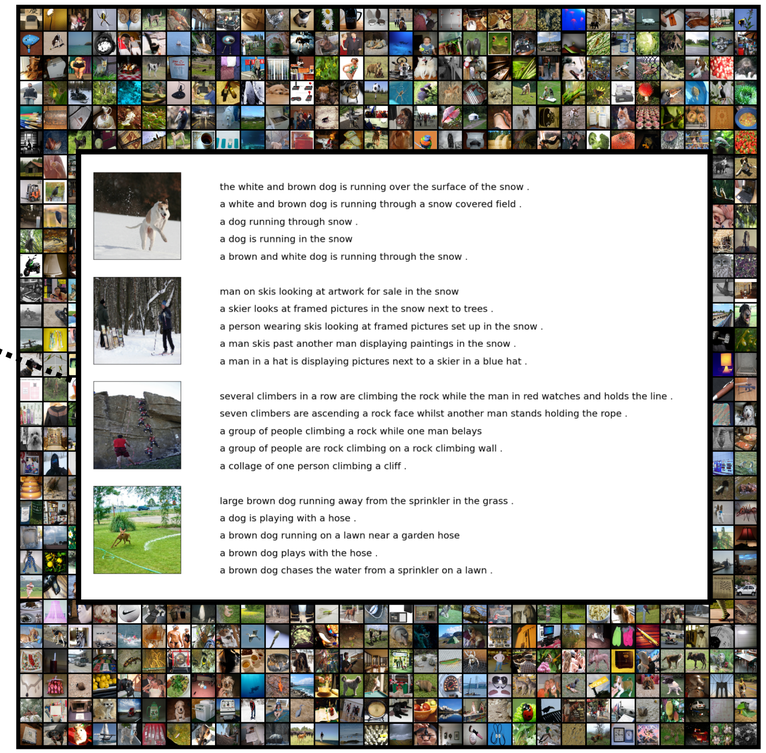
Zero-Shot Prediction: Compatibility of Three Distributions

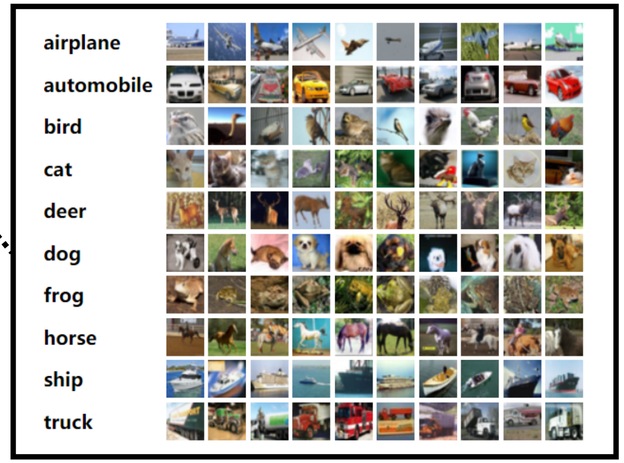
$P_{X,Z}$

Pre-Training

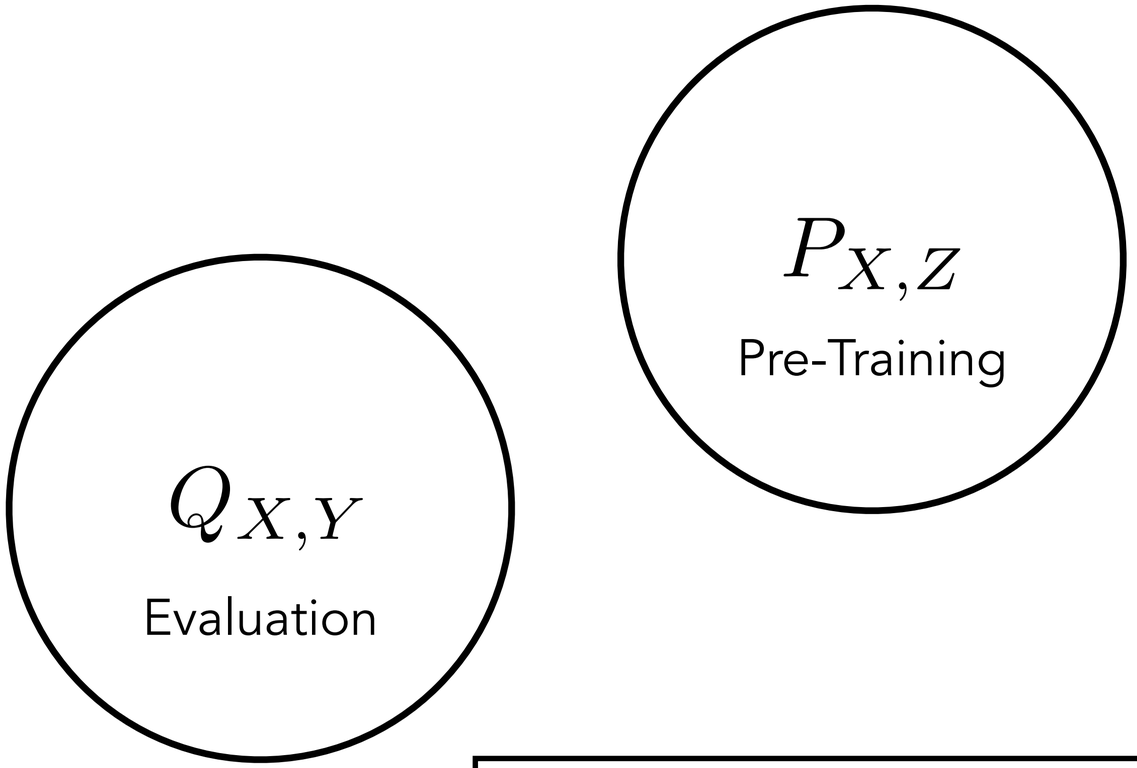
$Q_{X,Y}$

Evaluation





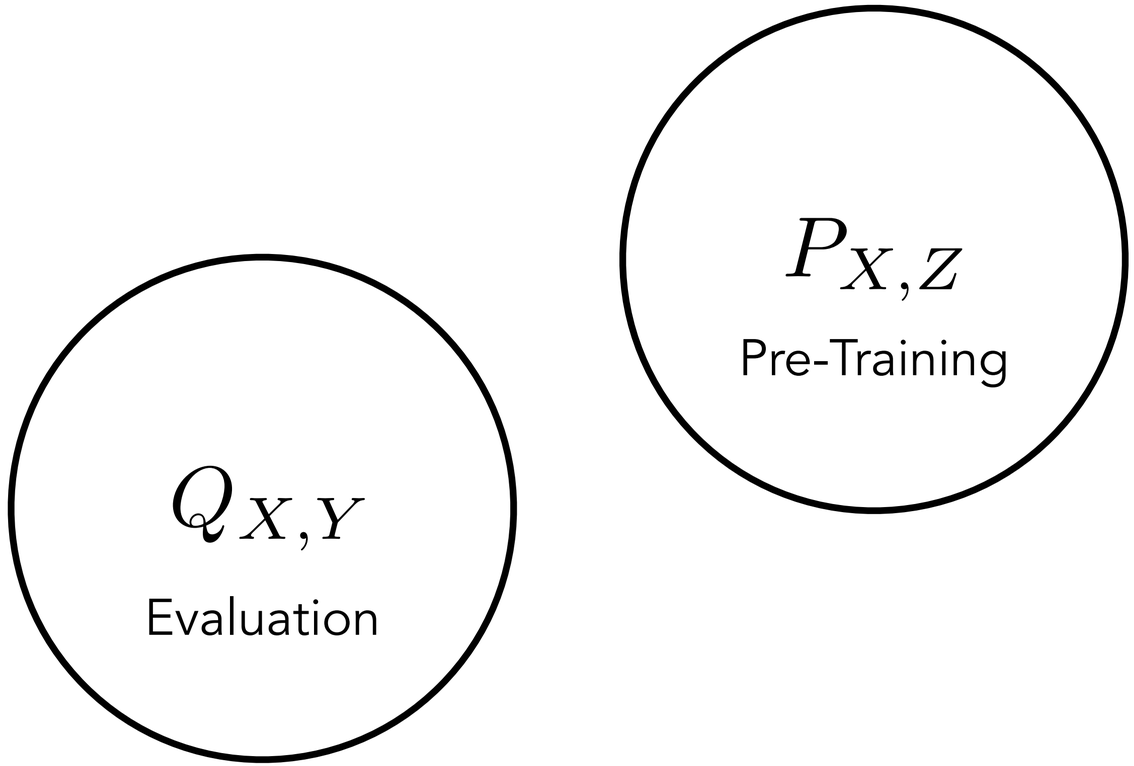
Zero-Shot Prediction: Compatibility of Three Distributions



Task: Least-Squares Regression

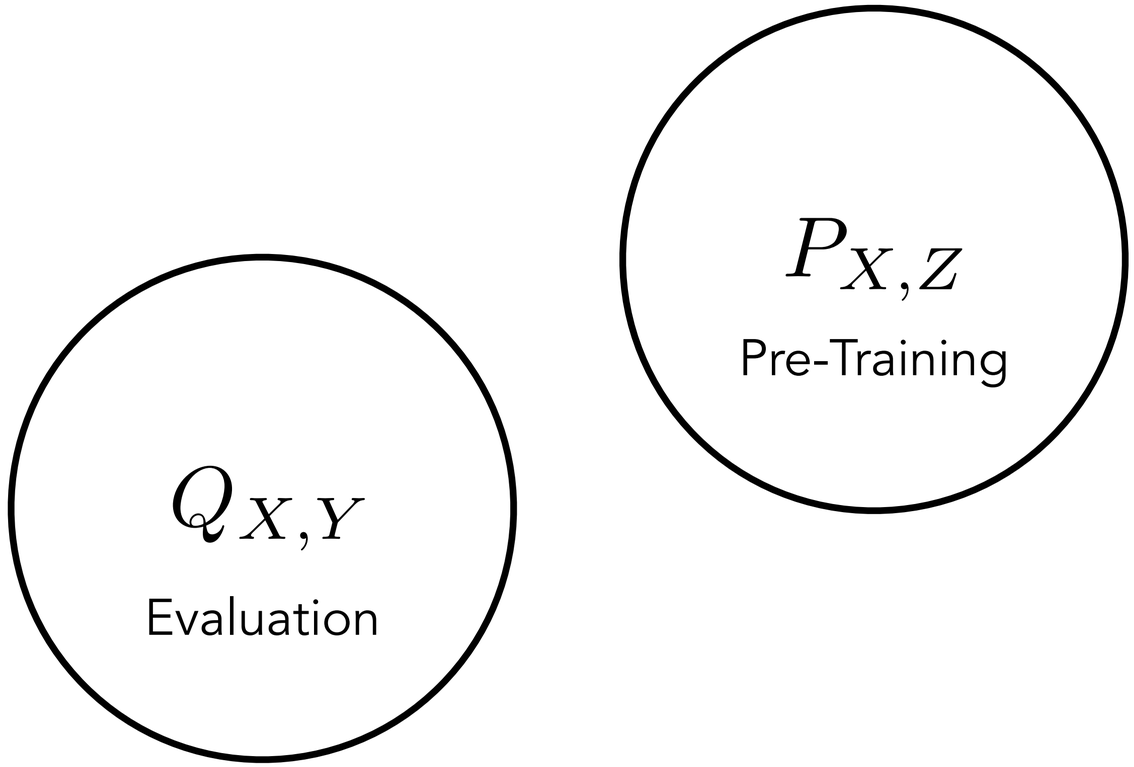
$$\min_{\eta:\mathcal{X}\to\mathbb{R}} \mathbb{E}_{Q_{X,Y}}\left[(Y-\eta(X))^2\right]$$

Zero-Shot Prediction: Compatibility of Three Distributions



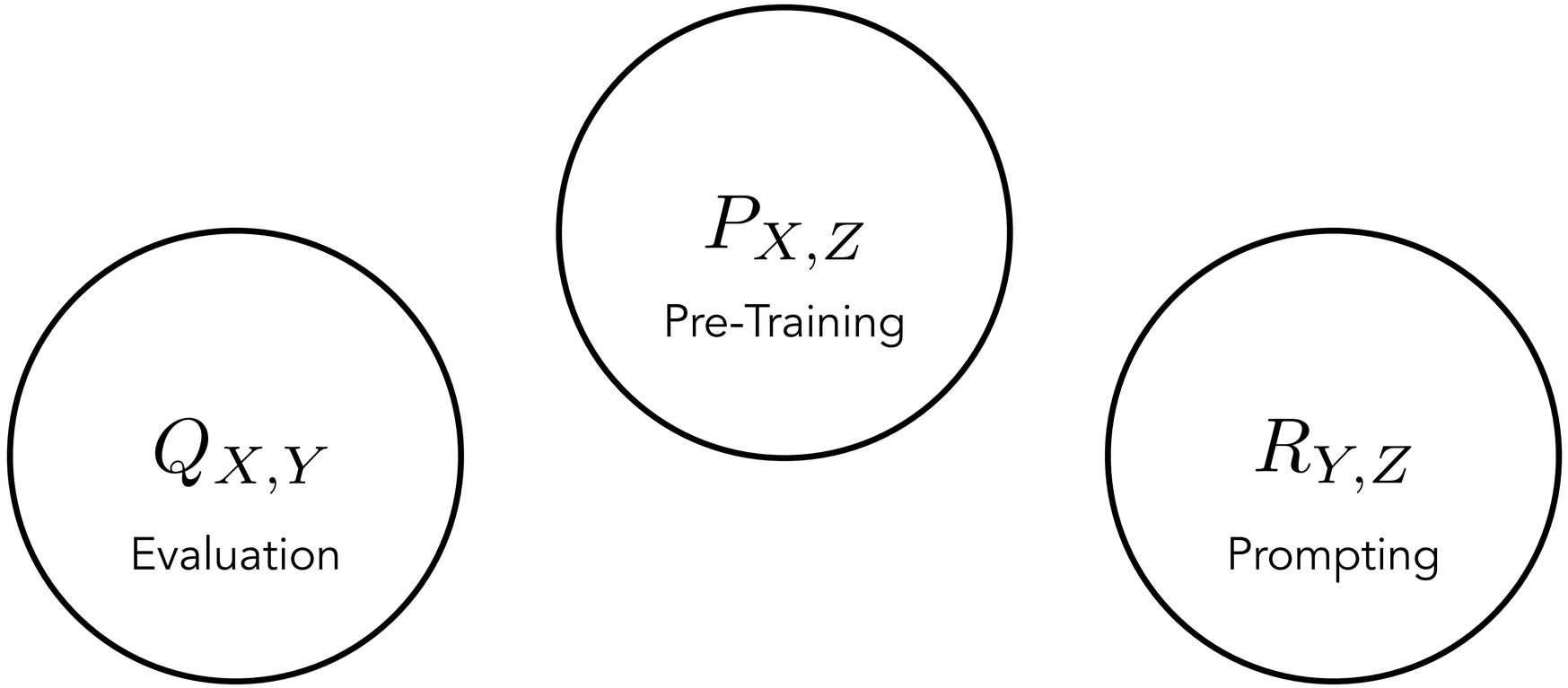
$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

Zero-Shot Prediction: Compatibility of Three Distributions



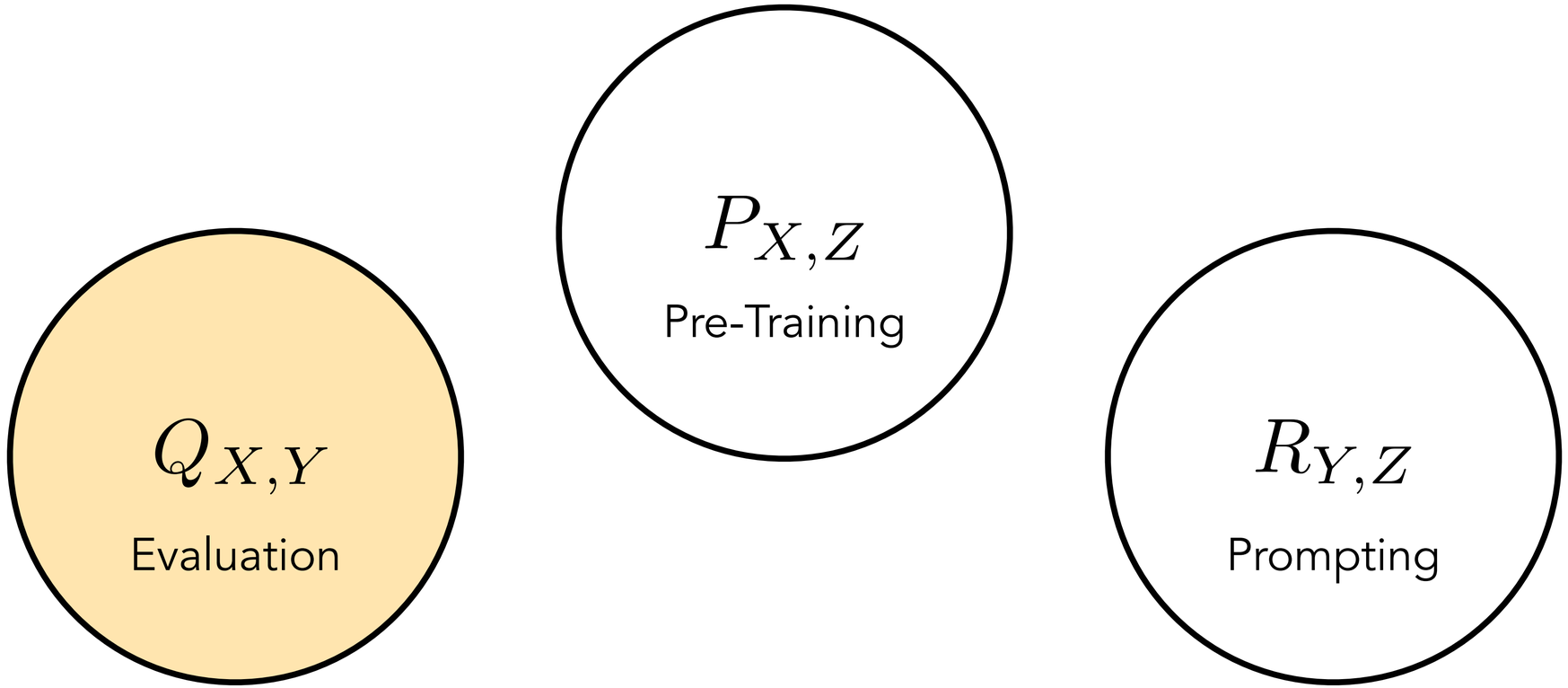
$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x}) \qquad \eta_{\text{ZS}}(\boldsymbol{x}) =$$

Zero-Shot Prediction: Compatibility of Three Distributions



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x}) \qquad \eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$

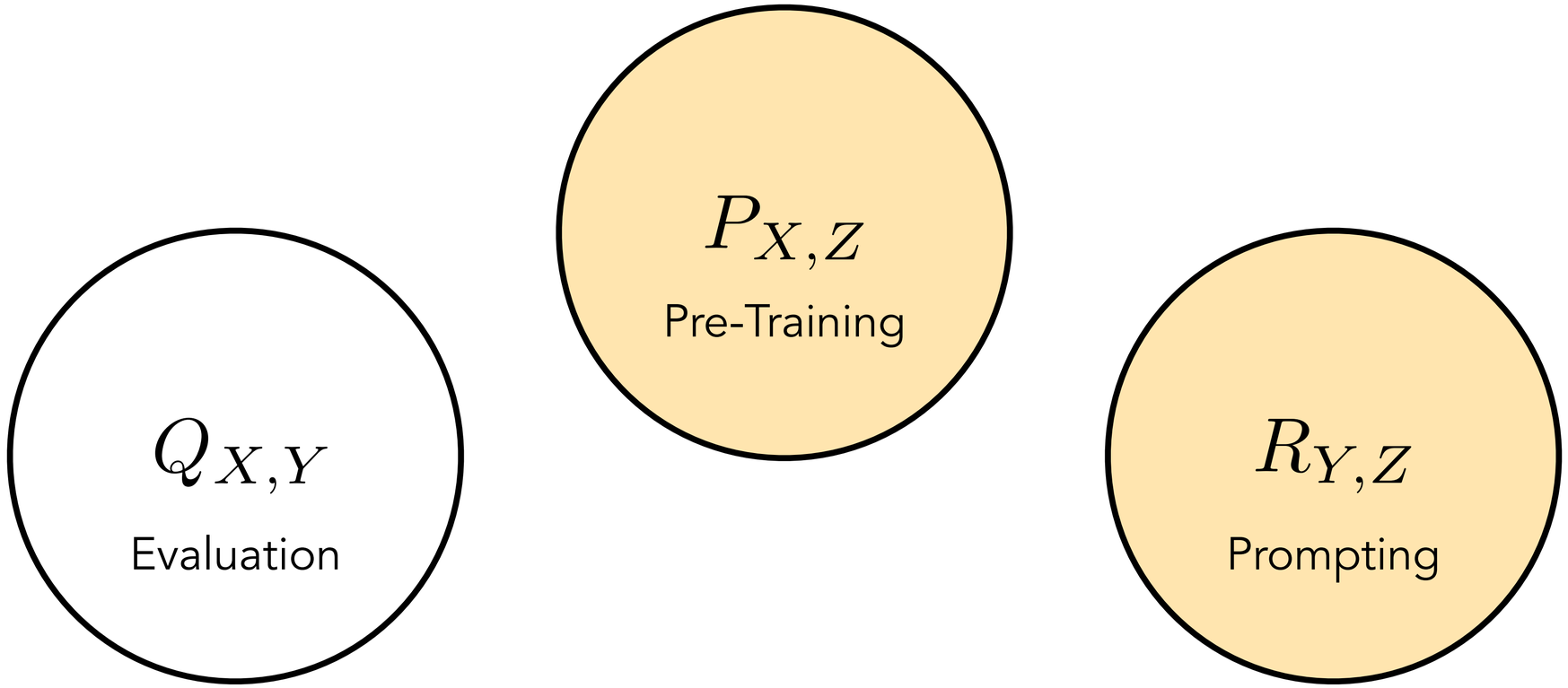
Zero-Shot Prediction: Compatibility of Three Distributions



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

$$\eta_{\text{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$

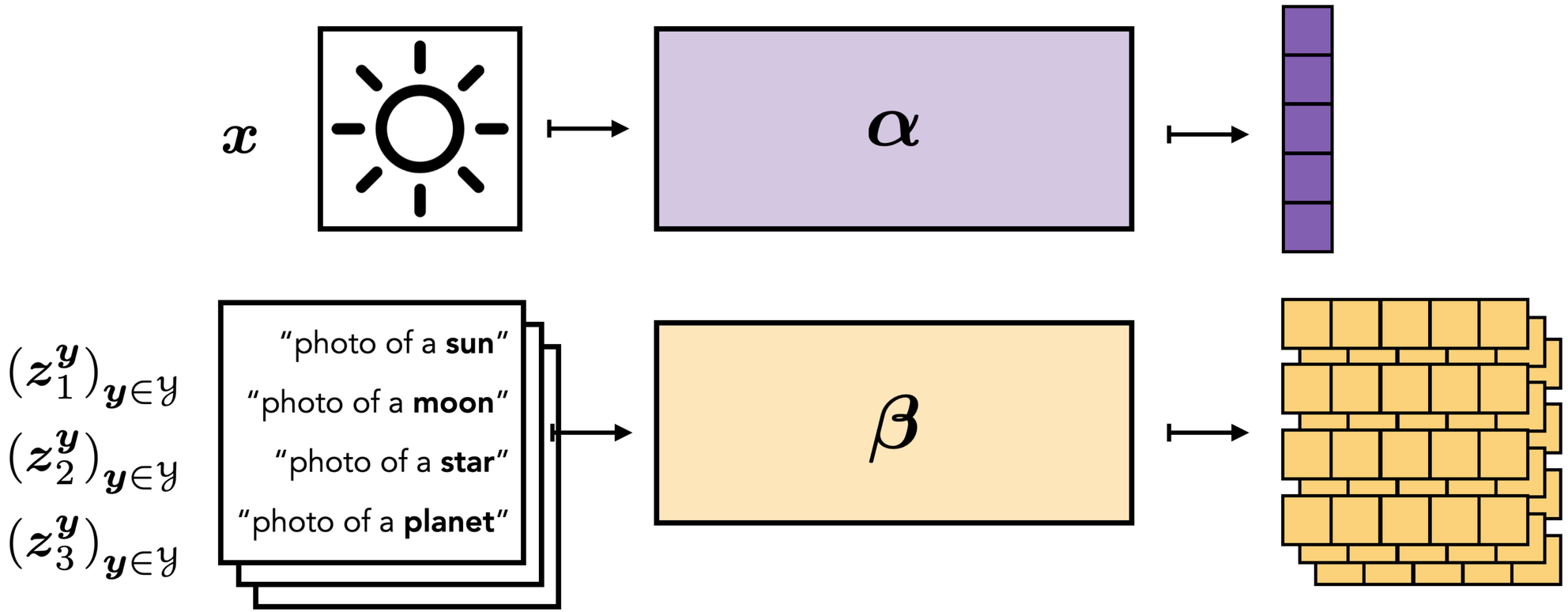
Zero-Shot Prediction: Compatibility of Three Distributions



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

$$\eta_{\text{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$

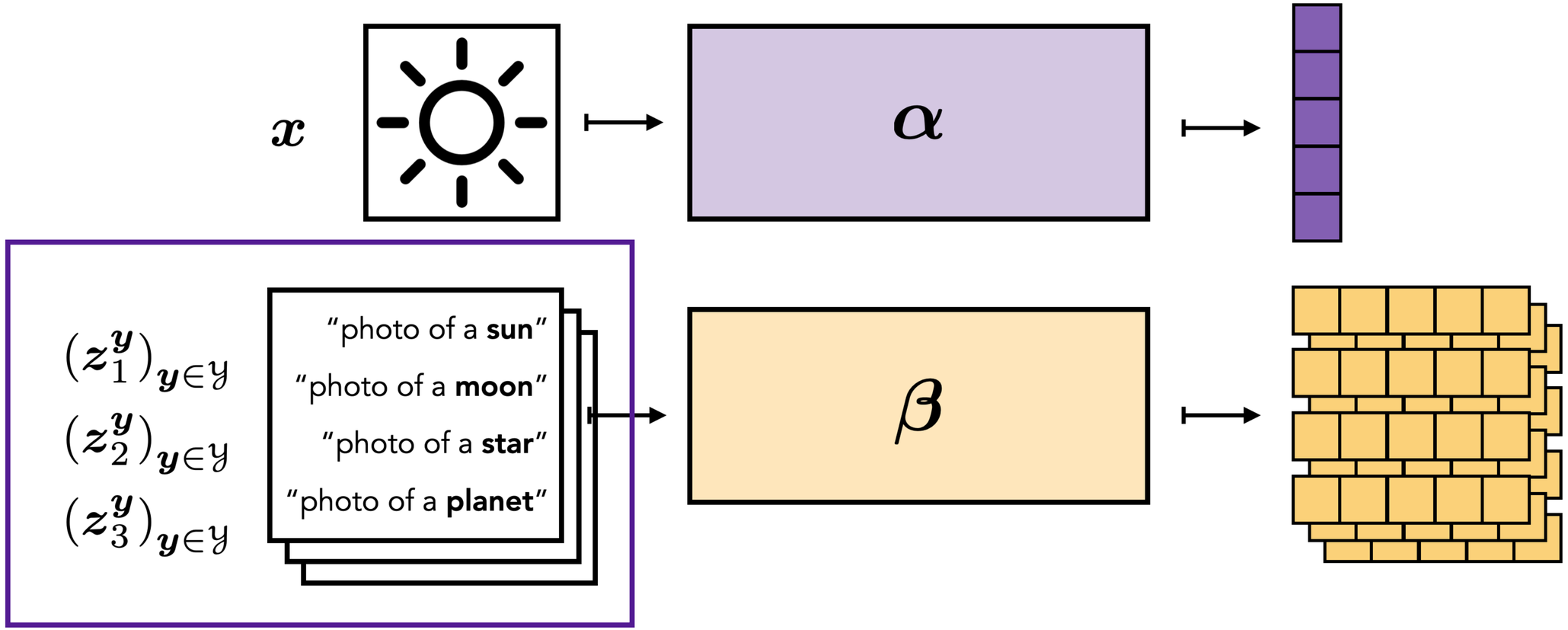
Zero-Shot Prediction: Compatibility of Three Distributions



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$

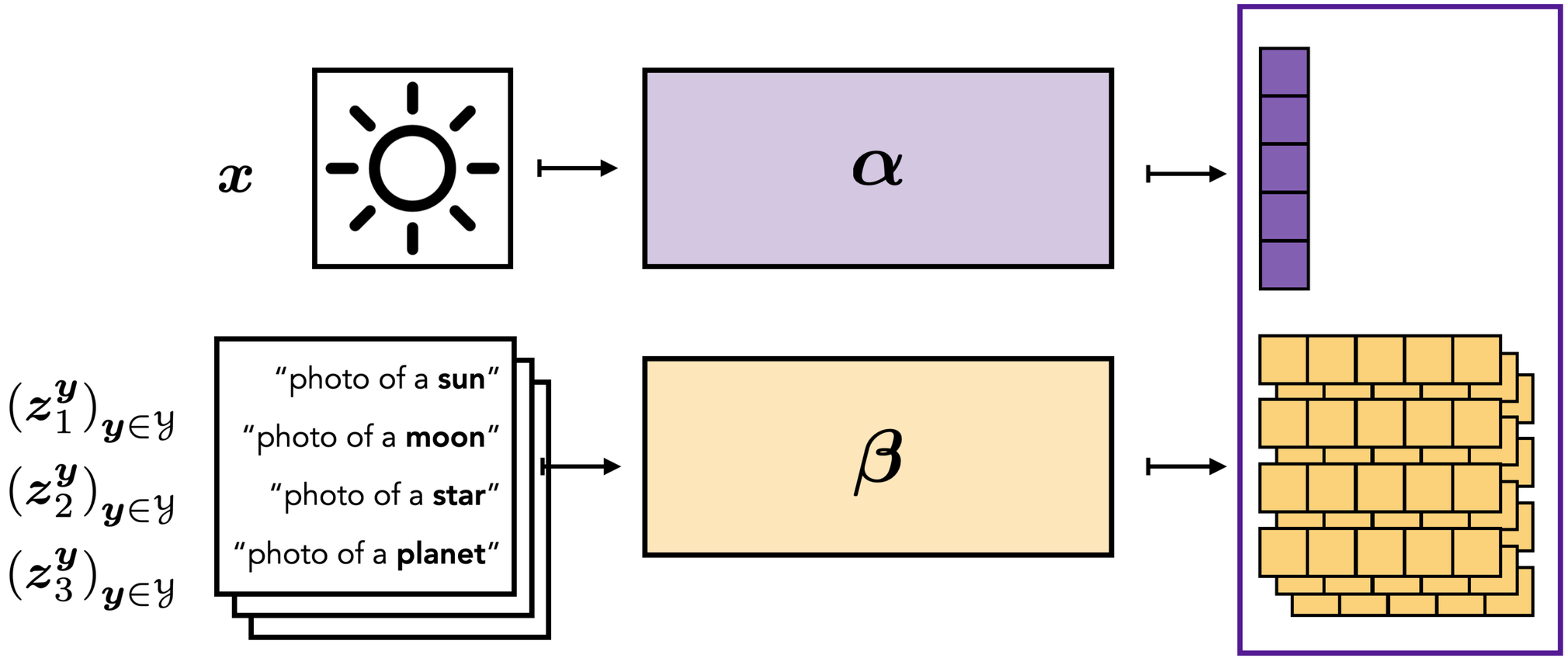
Zero-Shot Prediction: Compatibility of Three Distributions



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

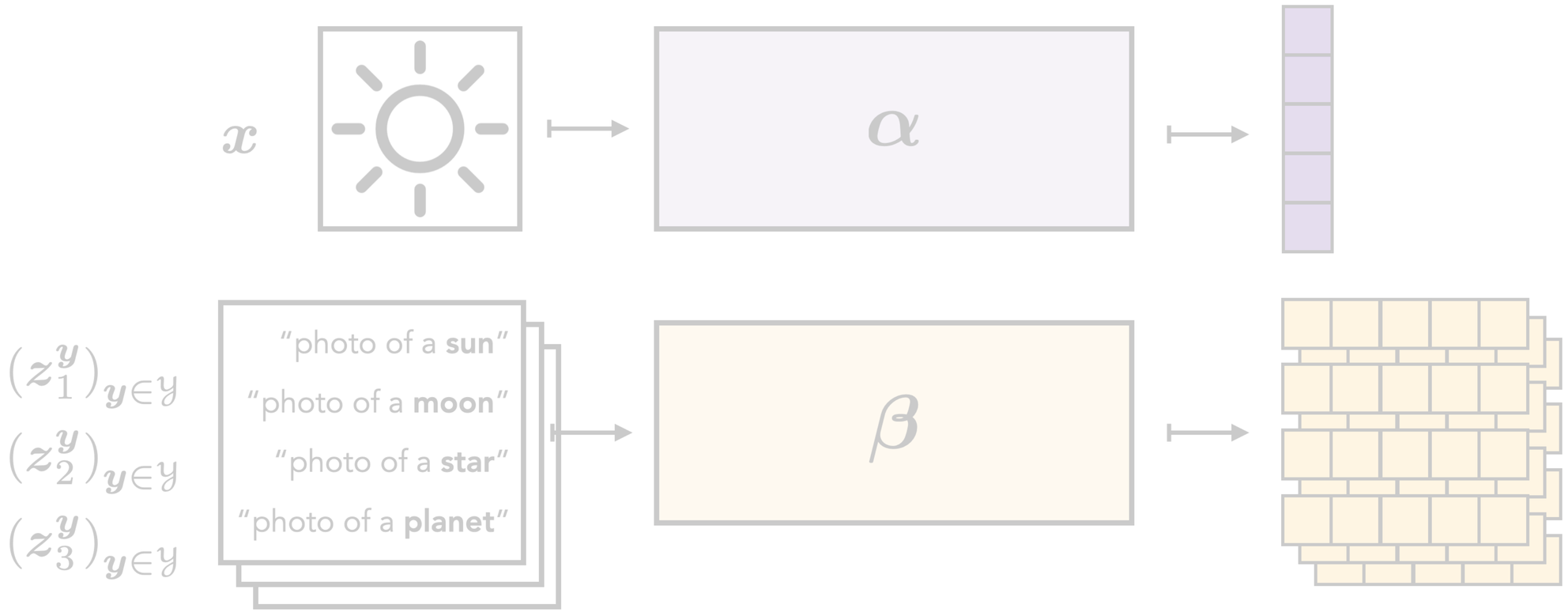
$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$

Zero-Shot Prediction: Compatibility of Three Distributions



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$



$$\eta_\star(\boldsymbol{x}) = \mathbb{E}_{Q_{X,Y}}\left[Y|X\right](\boldsymbol{x})$$

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[\mathbb{E}_{R_{Y,Z}}\left[Y|Z\right]|X\right](\boldsymbol{x})$$

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{P_{X,Z}}\left[g_R(Z)|X\right](\boldsymbol{x})$$

Conditional Mean

Information Density

Conditional Mean

$$\mu_{X \leftarrow Z} : \mathbf{L}^2(P_Z) \to \mathbf{L}^2(P_X)$$

$$[\mu_{X \leftarrow Z} g](\boldsymbol{x}) = \mathbb{E}_P \left[g(Z) | X \right](\boldsymbol{x})$$

Information Density

$$\mathsf{R}(\boldsymbol{x}, \boldsymbol{z}) := \frac{\mathrm{d}P_{X,Z}}{\mathrm{d}P_X P_Z}(\boldsymbol{x}, \boldsymbol{z}) = \frac{\mathrm{d}P_{Z|X=\boldsymbol{x}}}{\mathrm{d}P_Z}(\boldsymbol{z})$$

Conditional Mean

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = [\mu_{X \leftarrow Z} g_R](\boldsymbol{x})$$

Information Density

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{R_{Y,Z}}\left[Y \cdot \mathsf{R}(\boldsymbol{x}, Z)\right] + \mathrm{err}(P_Z, R_Z)$$

Conditional Mean

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = [\mu_{X \leftarrow Z} g_R](\boldsymbol{x})$$

Conditional Mean Approach

$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = [\hat{\mu}_{X \leftarrow Z} \hat{g}_R](\boldsymbol{x})$$

1. $(X_1, Z_1), \ldots, (X_n, Z_n) \sim P_{X,Z}$
   Estimate $\mu_{X \leftarrow Z}$ (operator regression)

2. $(Y_1, Z_1), \ldots, (Y_M, Z_M) \sim R_{Y,Z}$
   Estimate g_R (nonparametric regression)

Information Density

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{R_{Y,Z}} \left[Y \cdot \mathsf{R}(\boldsymbol{x}, Z) \right] + \mathrm{err}(P_Z, R_Z)$$

Conditional Mean

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = [\mu_{X \leftarrow Z} g_R](\boldsymbol{x})$$

Conditional Mean Approach

$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = [\hat{\mu}_{X \leftarrow Z} \hat{g}_R](\boldsymbol{x})$$

1. $(X_1, Z_1), \ldots, (X_n, Z_n) \sim P_{X,Z}$
   Estimate $\mu_{X \leftarrow Z}$ (operator regression)

2. $(Y_1, Z_1), \ldots, (Y_M, Z_M) \sim R_{Y,Z}$
   Estimate g_R (nonparametric regression)

Information Density

$$\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{R_{Y,Z}}\left[Y \cdot \mathsf{R}(\boldsymbol{x}, Z)\right] + \mathrm{err}(P_Z, R_Z)$$

Information Density Approach

$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{\hat{R}_{Y,Z}}[Y \cdot \widehat{\mathsf{R}}(\boldsymbol{x}, Z)]$$

1. $(X_1, Z_1), \ldots, (X_n, Z_n) \sim P_{X,Z}$
   $(X'_1, Z'_1), \ldots, (X'_n, Z'_n) \sim P_X P_Z$
   Estimate R (RND estimation)

2. $(Y_1, Z_1), \ldots, (Y_M, Z_M) \sim R_{Y,Z}$
   Estimate $\mathbb{E}_{R_{Y,Z}}[\cdot]$ (infinite-dimensional mean estimation)

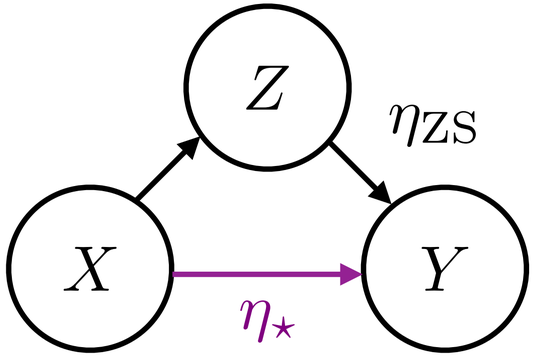
$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**

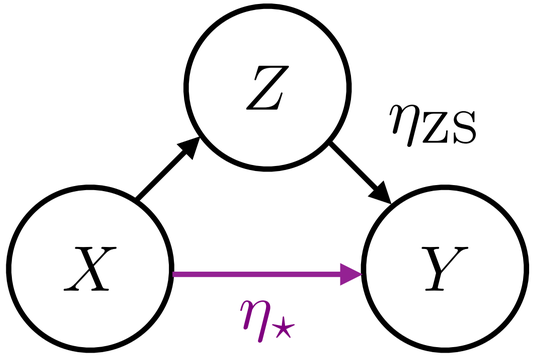


$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X, Y|Z)\right]$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



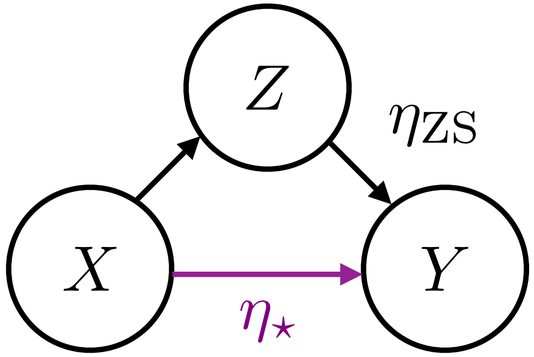
Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X, Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$g_R(\boldsymbol{z}) := \mathbb{E}_{R_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ prompt distribution

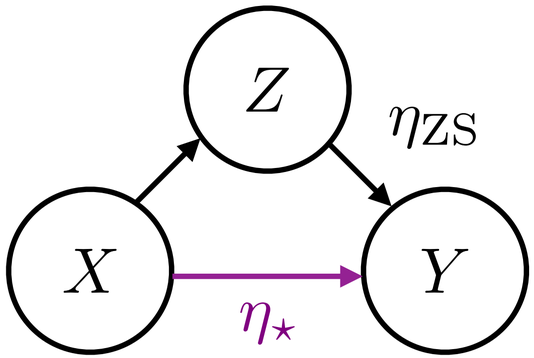
$g_Q(\boldsymbol{z}) := \mathbb{E}_{Q_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ true relationship

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim$$
$$\mathbb{E}_{Q_Z}\left[I(X, Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$g_R(\boldsymbol{z}) := \mathbb{E}_{R_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ prompt distribution

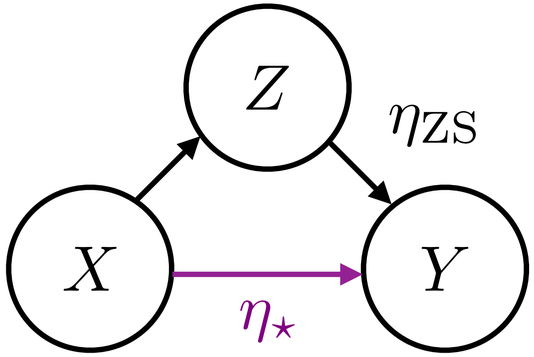
$g_Q(\boldsymbol{z}) := \mathbb{E}_{Q_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ true relationship

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim$$
$$\mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$g_R(\boldsymbol{z}) := \mathbb{E}_{R_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ prompt distribution

$g_Q(\boldsymbol{z}) := \mathbb{E}_{Q_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ true relationship

Thm. 2 (M., Harchaoui) Conditional Mean

$$\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(P_X)} \lesssim$$
$$\mathrm{plog}(1/\delta)\left[n^{-\frac{q}{q+1}} + M^{-\frac{2p}{2p-1}}\right] \text{ w.p. } 1-\delta$$

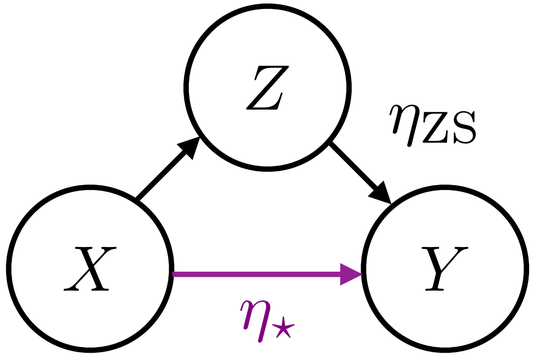
$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = [\hat{\mu}_{X\leftarrow Z}\hat{g}_R](\boldsymbol{x})$$

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim$$
$$\mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$g_R(\boldsymbol{z}) := \mathbb{E}_{R_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ prompt distribution

$g_Q(\boldsymbol{z}) := \mathbb{E}_{Q_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ true relationship

Thm. 2 (M., Harchaoui) Conditional Mean

$$\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(P_X)} \lesssim$$
$$\mathrm{plog}(1/\delta)\left[n^{-\frac{q}{q+1}} + M^{-\frac{2p}{2p-1}}\right] \text{ w.p. } 1-\delta$$

$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = [\hat{\mu}_{X\leftarrow Z}\hat{g}_R](\boldsymbol{x})$$

$n^{-\frac{q}{q+1}}$ (estimating $\mu_{X\leftarrow Z}$ using $\hat{\mu}_{X\leftarrow Z}$)

$1 \leftarrow q \rightarrow \infty$ effective dimension of Z is small relative to effective dimension of X

X and Z are (approximately) independent

$M^{-\frac{2p}{2p+1}}$ (estimating g_R using $\hat{g}_R$)

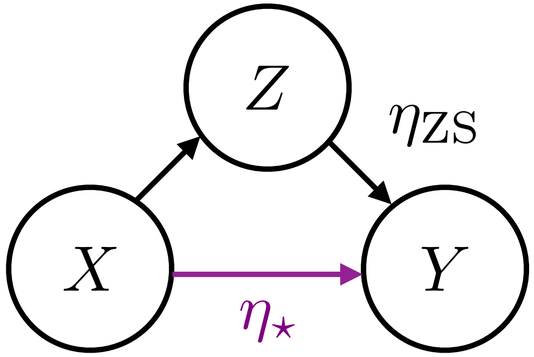
$1 \leftarrow p \rightarrow \infty$ effective dimension of Z is small

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$g_R(\boldsymbol{z}) := \mathbb{E}_{R_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ prompt distribution

$g_Q(\boldsymbol{z}) := \mathbb{E}_{Q_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ true relationship

Thm. 3 (M., Harchaoui) Information Density

$$\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(P_X)} \lesssim \mathrm{err}(P_Z, R_Z)$$

$$\mathrm{plog}(1/\delta)\left[n^{-\frac{q}{q+1}} + M^{-1}\right] \text{ w.p. } 1-\delta$$

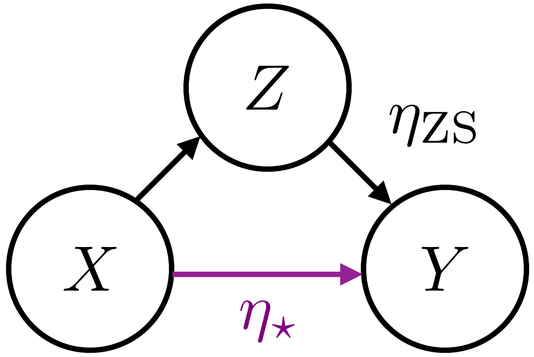
$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{\hat{R}_{Y,Z}}[Y \cdot \widehat{\mathsf{R}}(\boldsymbol{x}, Z)]$$

$$\|\eta_\star - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \leq 2\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} + 2\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)}$$

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$

$Q_{X,Y,Z}$ joint evaluation distribution with **latent caption**



Direct and indirect paths differ when X and Y are **conditionally dependent** given Z.

$g_R(\boldsymbol{z}) := \mathbb{E}_{R_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ prompt distribution

$g_Q(\boldsymbol{z}) := \mathbb{E}_{Q_{Y,Z}}\left[Y|Z\right](\boldsymbol{z})$ true relationship

Thm. 3 (M., Harchaoui) Information Density

$$\|\eta_{\mathrm{ZS}} - \hat{\eta}_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(P_X)} \lesssim \mathrm{err}(P_Z, R_Z)$$

$$\mathrm{plog}(1/\delta)\left[n^{-\frac{q}{q+1}} + M^{-1}\right] \text{ w.p. } 1-\delta$$

$$\hat{\eta}_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{\hat{R}_{Y,Z}}[Y \cdot \widehat{\mathsf{R}}(\boldsymbol{x}, Z)]$$

$n^{-\frac{q}{q+1}}$ (estimating R using $\widehat{\mathsf{R}}$)

$1 \leftarrow q \rightarrow \infty$ X and Z are (approximately) independent

M^{-1} (estimating $R_{Y,Z}$ using $\hat{R}_{Y,Z}$)

$\mathrm{err}\left(P_Z, R_Z\right)$ (error term from $\eta_{\mathrm{ZS}}(\boldsymbol{x}) = \mathbb{E}_{R_{Y,Z}}\left[Y \cdot \mathsf{R}(\boldsymbol{x}, Z)\right] + \mathrm{err}(P_Z, R_Z)$)

Research Contributions

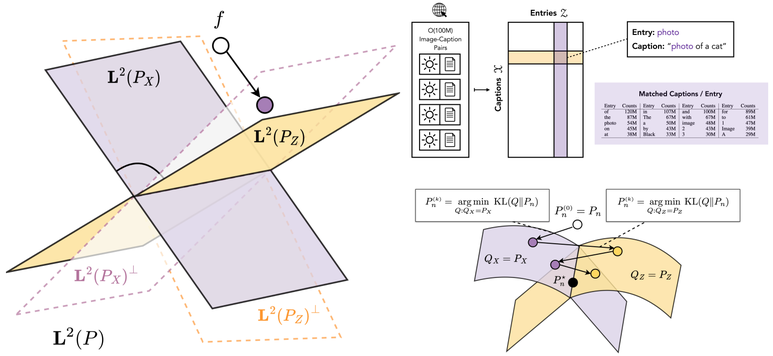
✅

Exact variance reduction analysis of balancing-based mean estimation!

Theorem (Liu, M., Pal, Harchaoui)

$$\mathbb{E}_P\left[(P_n^{(k)}(h) - P(h))^2\right] = \frac{\mathbb{V}\mathrm{ar}(\ldots \mu^{\perp}_{Z \leftarrow X} \mu^{\perp}_{X \leftarrow Z} h)}{n} + \tilde{O}\left(\frac{k^6}{n^{3/2}}\right)$$

(k times)

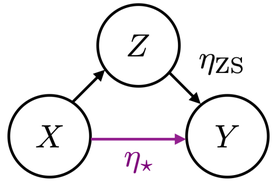


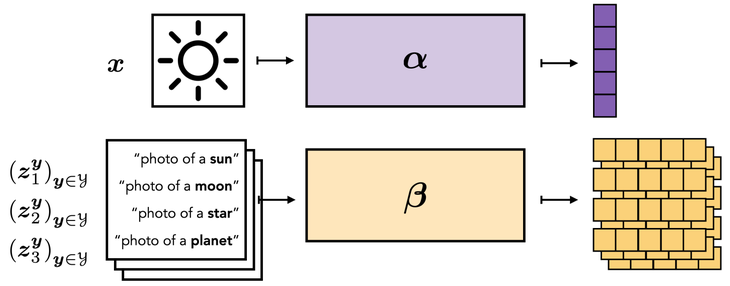
✅

Theoretical framework for obtaining generalization guarantees for two broad classes of zero-shot prediction models!

Thm. 1 (M., Harchaoui)

$$\|\eta_\star - \eta_{\mathrm{ZS}}\|^2_{\mathbf{L}^2(Q_X)} \lesssim \mathbb{E}_{Q_Z}\left[I(X,Y|Z)\right] + \|g_Q - g_R\|^2_{\mathbf{L}^2(Q_Z)}$$





Preliminaries

Balanced Pre-Training

Zero-Shot Prediction

Conclusion

Perspectives: Balanced Pre-Training

- Non-asymptotic mean squared error bounds for **nonlinear** functionals.

$$\Psi(P_n^{(k)}) + P_n(\phi_n^{(k)}) - \Psi(P) = [P_n - P](\phi_0) + \mathcal{R}_n^{(k)} + \mathcal{D}_n^{(k)}$$

-

Perspectives: Balanced Pre-Training

- Non-asymptotic mean squared error bounds for **nonlinear** functionals.

$$\Psi(P_n^{(k)}) + P_n(\phi_n^{(k)}) - \Psi(P) = [P_n - P](\phi_0) + \mathcal{R}_n^{(k)} + \mathcal{D}_n^{(k)}$$

- tangent space for "known marginal" model known (Bickel et al, '90)
- $\Psi(P_n^{(k)}) + P(\phi_n^{(k)}) - \Psi(P)$ — $o_p(n^{-1/2})$ remainder?
- $[P_n - P](\phi_n^{(k)} - \phi_0)$ — $o_p(n^{-1/2})$ drift?

Perspectives: Balanced Pre-Training

- Non-asymptotic mean squared error bounds for **nonlinear** functionals.

$$\Psi(P_n^{(k)}) + P_n(\phi_n^{(k)}) - \Psi(P) = [P_n - P](\phi_0) + \mathcal{R}_n^{(k)} + \mathcal{D}_n^{(k)}$$

tangent space for "known marginal" model known (Bickel et al, '90)

$\Psi(P_n^{(k)}) + P(\phi_n^{(k)}) - \Psi(P)$ — $o_p(n^{-1/2})$ remainder?

$[P_n - P](\phi_n^{(k)} - \phi_0)$ — $o_p(n^{-1/2})$ drift?

- Simultaneously optimizing (P_X, P_Z) for downstream performance (bilevel).

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^\star}\left[h_\theta(X, Z)\right] \longrightarrow \mathbb{E}_P[h_{\hat{\theta}_n}(X, Z)] - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}_P\left[h_\theta(X, Z)\right]$$

Perspectives: Balanced Pre-Training

- Non-asymptotic mean squared error bounds for **nonlinear** functionals.

$$\Psi(P_n^{(k)}) + P_n(\phi_n^{(k)}) - \Psi(P) = [P_n - P](\phi_0) + \mathcal{R}_n^{(k)} + \mathcal{D}_n^{(k)}$$

tangent space for "known marginal" model known (Bickel et al, '90)

$\Psi(P_n^{(k)}) + P(\phi_n^{(k)}) - \Psi(P)$ — $o(n^{-1/2})$ remainder

$[P_n - P](\phi_n^{(k)} - \phi_0)$ — $o(n^{-1/2})$ drift

- Simultaneously optimizing (P_X, P_Z) for downstream performance (bilevel).

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{P_n^\star}[h_\theta(X, Z)] \longrightarrow \mathbb{E}_P[h_{\hat{\theta}_n}(X, Z)] - \inf_{\theta \in \mathbb{R}^d} \mathbb{E}_P[h_\theta(X, Z)]$$

(P_X, P_Z)

Perspectives: Zero-Shot Prediction

- Incorporating ERM-based statistical results for models based optimizing foundation modeling objectives.

$$\widehat{\mathsf{R}} = \min_{\mathsf{R}} L_n(\mathsf{R})$$

Perspectives: Zero-Shot Prediction

- Incorporating ERM-based statistical results for models based optimizing foundation modeling objectives.

$$\widehat{\mathsf{R}} = \min_{\mathsf{R}} L_n(\mathsf{R})$$

$$\|\widehat{\mathsf{R}}_{\mathrm{opt}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)} \leq$$

Perspectives: Zero-Shot Prediction

- Incorporating ERM-based statistical results for models based optimizing foundation modeling objectives.

$$\widehat{\mathsf{R}} = \min_{\mathsf{R}} L_n(\mathsf{R})$$

$$\|\widehat{\mathsf{R}}_{\text{opt}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)} \leq \underbrace{2\|\widehat{\mathsf{R}}_{\text{opt}} - \widehat{\mathsf{R}}\|^2_{\mathbf{L}^2(P_X P_Z)}}_{\text{optimization error (non-convex)}} + 2\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)}$$

Perspectives: Zero-Shot Prediction

- Incorporating ERM-based statistical results for models based optimizing foundation modeling objectives.

$$\widehat{\mathsf{R}} = \min_{\mathsf{R}} L_n(\mathsf{R})$$

$$\|\widehat{\mathsf{R}}_{\text{opt}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)} \leq \underbrace{2\|\widehat{\mathsf{R}}_{\text{opt}} - \widehat{\mathsf{R}}\|^2_{\mathbf{L}^2(P_X P_Z)}}_{\text{optimization error (non-convex)}} + \underbrace{2\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)}}_{\text{statistical error}}$$

$$\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)} \leq C(L(\widehat{\mathsf{R}}) - L(\mathsf{R}))$$

Perspectives: Zero-Shot Prediction

- Incorporating ERM-based statistical results for models based optimizing foundation modeling objectives.

$$\widehat{\mathsf{R}} = \min_{\mathsf{R}} L_n(\mathsf{R})$$

$$\|\widehat{\mathsf{R}}_{\text{opt}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)} \leq \underbrace{2\|\widehat{\mathsf{R}}_{\text{opt}} - \widehat{\mathsf{R}}\|^2_{\mathbf{L}^2(P_X P_Z)}}_{\text{optimization error (non-convex)}} + \underbrace{2\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)}}_{\text{statistical error}}$$

$$\begin{aligned}\|\widehat{\mathsf{R}} - \mathsf{R}\|^2_{\mathbf{L}^2(P_X P_Z)} &\leq C(L(\widehat{\mathsf{R}}) - L(\mathsf{R})) \\ &\leq 2C \sup_{\mathsf{R}} (L_n(\mathsf{R}) - L(\mathsf{R}))\end{aligned}$$

Preliminaries

Balanced Pre-Training

Zero-Shot Prediction

Conclusion

Thank you!

To my co-authors, mentors,

and to all my friends and family.

NSF
IFDS Institute for Foundations of Data Science
IFML
NIH National Institutes of Health