Stochastic Optimization for Spectral Risk Measures

Ronak Mehta

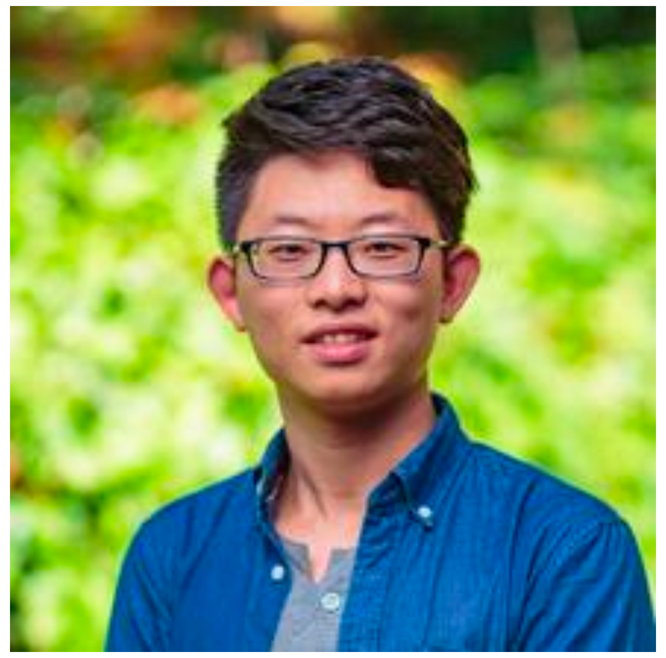
June 03, 2023

Team

Ronak Mehta
University of Washington

Vincent Roulet
Google Research

Krishna Pillutla
Google Research



Lang Liu
University of Washington

Zaid Harchaoui
University of Washington

IFDS
Institute for Foundations of Data Science

NSF

W
UNIVERSITY of WASHINGTON

Stochastic Programming is the prevailing model for machine learning.

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{Z \sim P}[\ell(w, Z)]$$

Stochastic Programming is the prevailing model for machine learning.

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{Z \sim P}[\ell(w, Z)]$$

model parameters

Stochastic Programming is the prevailing model for machine learning.

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{Z \sim P}[\ell(w, Z)]$$

data generating distribution

data instance

Stochastic Programming is the prevailing model for machine learning.

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{Z \sim P}[\ell(w, Z)]$$

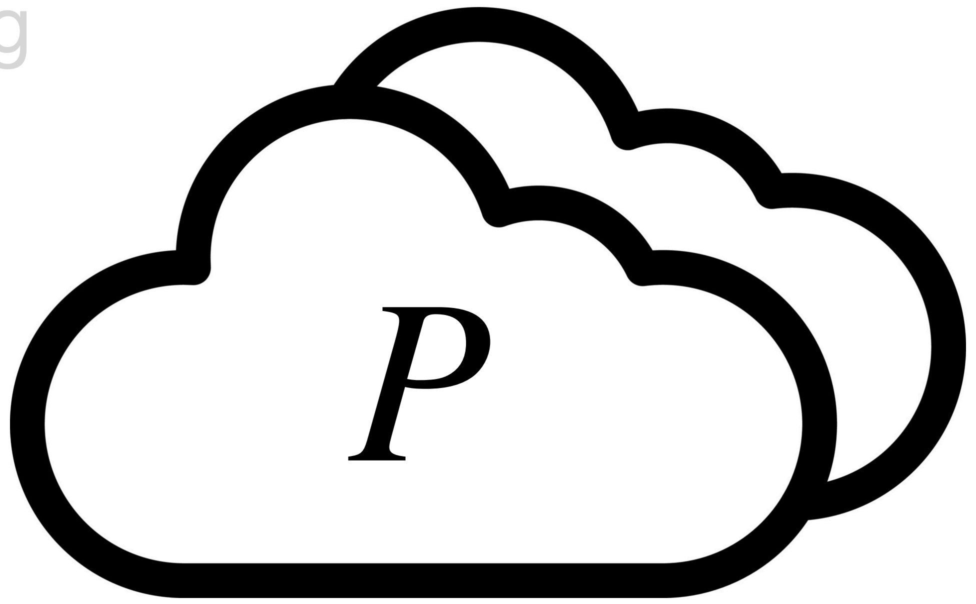
loss function

Stochastic Programming is the prevailing model for machine learning.

$$\min_{w\in\mathbb{R}^d} \mathbb{E}_{Z\sim P}[\ell(w,Z)]$$

$$\approx$$

Training



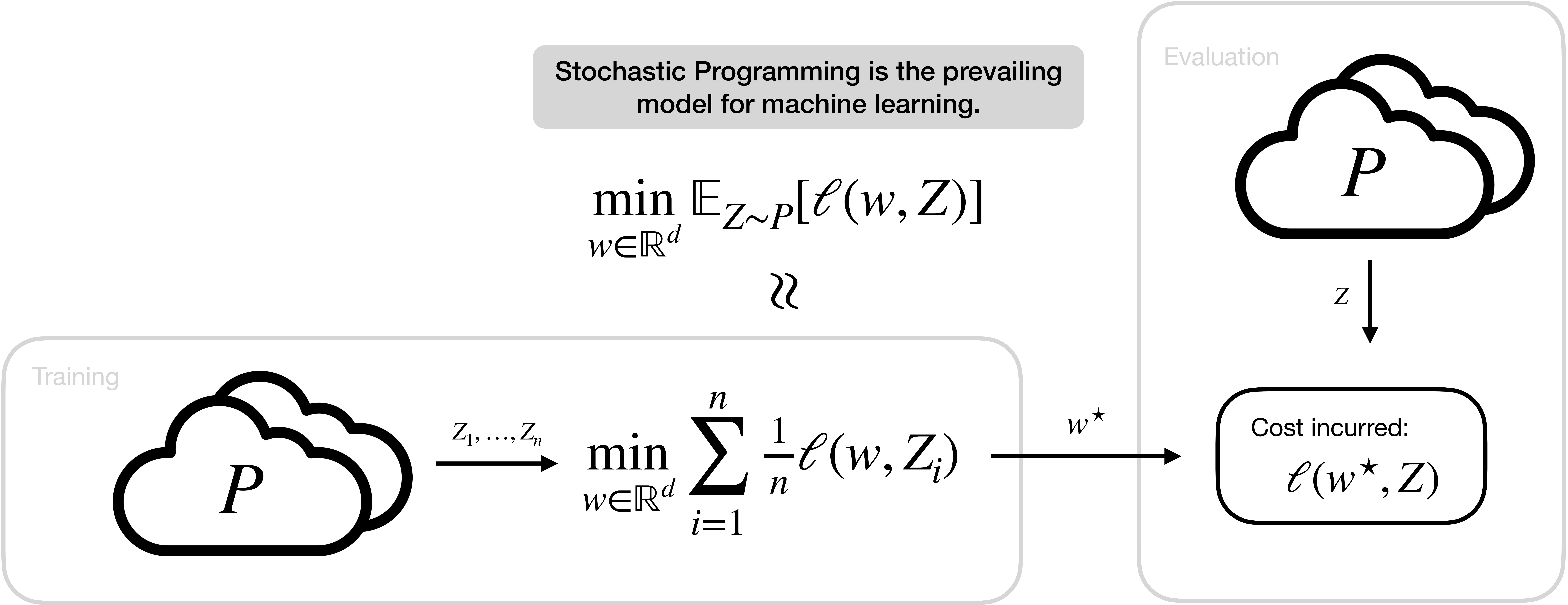
$\xrightarrow{Z_1,\ldots,Z_n}$

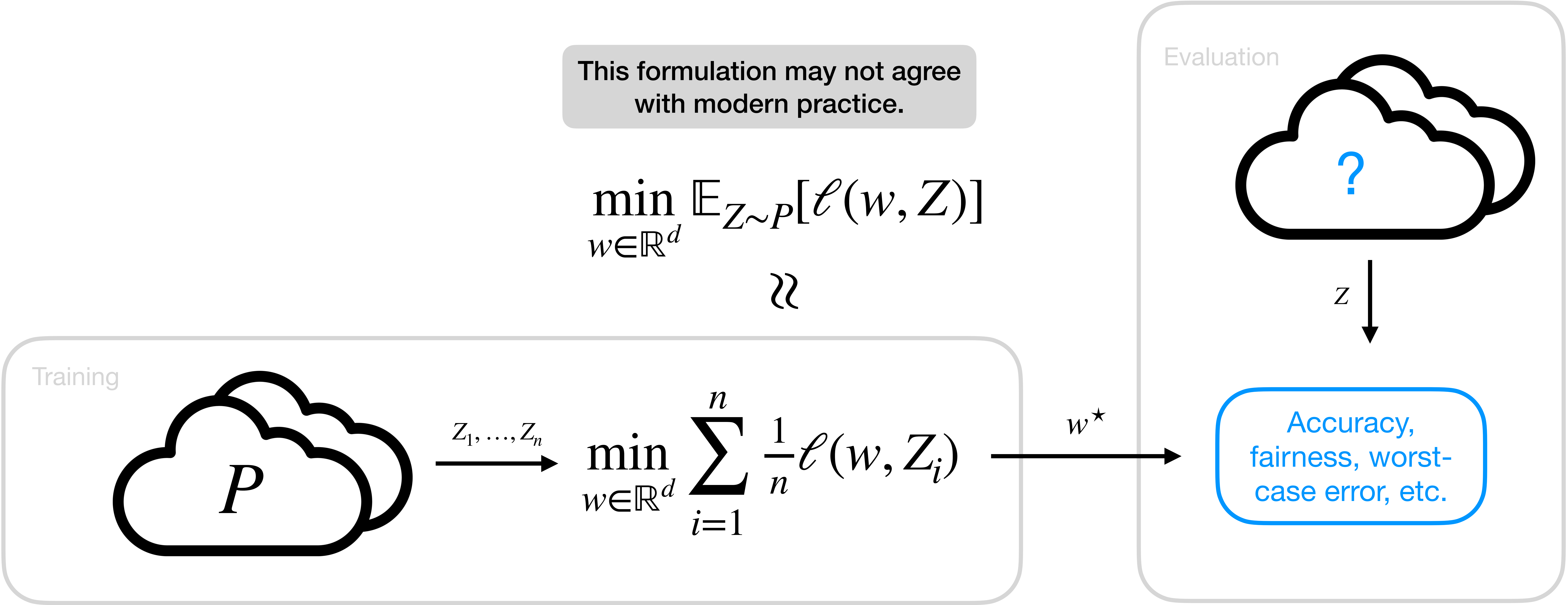
$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{n}\ell(w,Z_i)$$

Stochastic Programming is the prevailing model for machine learning.

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{Z \sim P}[\ell(w, Z)]$$

$$\approx$$




This formulation may not agree with modern practice.
$\min_{w \in \mathbb{R}^d} \mathbb{E}_{Z \sim P}[\ell(w, Z)]$
$\approx$
Training
P
$Z_1, \ldots, Z_n$
$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{n} \ell(w, Z_i)$
$w^\star$
Evaluation
?
Z
Accuracy, fairness, worst-case error, etc.

Distributionally robust objectives explicitly account for subpopulation shifts.

$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathcal{U}} \sum_{i=1}^{n} q_i \ell(w, Z_i) - \nu D(q \| \mathbf{1}_n / n)$$

Distributionally robust objectives explicitly account for subpopulation shifts.

$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathcal{U}} \sum_{i=1}^{n} q_i \ell(w, Z_i) - \nu D(q \| \mathbf{1}_n / n)$$

ambiguity set of possible distributions, i.e. each $q_i \geq 0$ and $\sum_{i=1}^{n} q_i = 1$

Distributionally robust objectives explicitly account for subpopulation shifts.

$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathcal{U}} \sum_{i=1}^{n} q_i \ell(w, Z_i) - \nu D(q \| \mathbf{1}_n / n)$$

shift cost

deviation of q from original distribution

Spectral risk measures are generated by letting $\mathcal{U}$ be a permutahedron in $\mathbb{R}^n$.

$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathcal{U}} \sum_{i=1}^{n} q_i \ell(w, Z_i) - \nu D(q \| \mathbf{1}_n / n)$$

Stochastic optimization is an essential ingredient for ERM, but implementing these algorithms for SRMs is a key challenge.

$$w_{t+1} = w_t - \eta_t g_t$$

stepsize sequence

stochastic gradient estimate that only depends on $O(1)$ calls to oracles $\{\ell(\cdot, Z_i), \nabla \ell(\cdot, Z_i)\}_{i=1}^n$

Notation

$R =$ objective function

$P_n =$ sampling distribution used for g_t (e.g. mini-batch sampling)

Stochastic optimization is an essential ingredient for ERM, but implementing these algorithms for SRMs is a key challenge.

$$w_{t+1} = w_t - \eta_t g_t$$

Bias

$$\mathbb{E}_{P_n}[g_t] - \nabla R(w_t)$$

Variance

$$\mathbb{E}_{P_n}\|g_t - \mathbb{E}[g_t]\|_2^2$$

Notation

$R =$ objective function

$P_n =$ sampling distribution used for g_t (e.g. mini-batch sampling)

Stochastic optimization is an essential ingredient for ERM, but implementing these algorithms for SRMs is a key challenge.

$$w_{t+1} = w_t - \eta_t g_t$$

Bias

$$\mathbb{E}_{P_n}[g_t] - \nabla R(w_t)$$

Variance

$$\mathbb{E}_{P_n}\|g_t - \mathbb{E}[g_t]\|_2^2$$

Problem in ERM as well, usually handled by decreasing learning rate or variance-reduced methods.

Notation

$R =$ objective function

$P_n =$ sampling distribution used for g_t (e.g. mini-batch sampling)

Stochastic optimization is an essential ingredient for ERM, but implementing these algorithms for SRMs is a key challenge.

$$w_{t+1} = w_t - \eta_t g_t$$

Unbiased estimates are used in ERM, but this is impossible for SRMs, resulting in poor convergence.

Bias

$$\mathbb{E}_{P_n}[g_t] - \nabla R(w_t)$$

Variance

$$\mathbb{E}_{P_n} \| g_t - \mathbb{E}[g_t] \|_2^2$$

Is there an optimizer that converges to the spectral risk minimizer using only $O(1)$ oracle calls per iterate?

Contributions

1. Characterize the smoothness properties of the objective as a function of the underlying losses.
2. Quantify the bias of current stochastic approaches.
3. Propose LSVRG, a stochastic optimization algorithm and establish its linear convergence rate.
4. Demonstrate superior convergence of LSVRG experimentally via numerical evaluations.

Outline

Properties of SRM Objective

LSVRG Algorithm

Theoretical Guarantees

Numerical Performance

Conclusion & Future Work

$$R(w) := \max_{q \in \mathscr{P}(\sigma)} q^\top \ell(w) - \nu n \|q - \mathbf{1}_n/n\|_2^2 + \frac{\mu}{2}\|w\|_2^2$$

$D_{\chi^2}(q \| \mathbf{1}_n/n) = n\|q - \mathbf{1}_n/n\|_2^2.$

strongly convex regularizer

$$R(w) := \max_{q \in \mathscr{P}(\sigma)} q^\top \ell(w) - \nu n \|q - \mathbf{1}_n/n\|_2^2 + \frac{\mu}{2}\|w\|_2^2$$

$\ell(w) := (\ell_1(w), \ldots, \ell_n(w)) \in \mathbb{R}^n$

$\ell_i(w) := \ell_i(w, Z_i) \quad i = 1, \ldots, n\,.$

Assumptions

Each loss $\ell_i : \mathbb{R}^d \to \mathbb{R}$ is convex, G-Lipschitz continuous, and L-smooth, i.e. $w \mapsto \nabla \ell(w)$ is well-defined and L-Lipschitz continuous w.r.t. $\|\cdot\|_2$.

The regularization parameter μ and shift cost ν satisfy $\mu > 0$ and $\nu > 0$.

Proposition 1

$$q^*(l) := \mathrm{argmax}_{q \in \mathscr{P}(\sigma)}\, q^\top l - \nu n \|q - \mathbf{1}_n/n\|_2^2$$

$$\nabla R(w) = \nabla \ell(w)^\top q^*(\ell(w)) + \mu w$$

$$= \sum_{i=1}^{n} q_i^*(\ell(w))(\nabla \ell_i(w) + \mu w)\,.$$

The gradient of R is a weighted average of the gradients of individual (regularized) losses, weighed by the "most unfavorable" distribution shift $q^*(\ell(w))$.

Proposition 1

$$q^*(l) := \mathrm{argmax}_{q \in \mathcal{P}(\sigma)}\, q^\top l - \nu n \|q - \mathbf{1}_n/n\|_2^2$$

$$\begin{aligned} \nabla R(w) &= \nabla \ell(w)^\top q^*(\ell(w)) + \mu w \\ &= \sum_{i=1}^{n} q_i^*(\ell(w))(\nabla \ell_i(w) + \mu w)\,. \end{aligned}$$

The gradient of R is a weighted average of the gradients of individual (regularized) losses, weighed by the "most unfavorable" distribution shift $q^*(\ell(w))$.

One could construct an unbiased estimator of $\nabla R(w)$… if $q^*(\ell(w))$ was known!

Outline

Properties of SRM Objective

LSVRG Algorithm

Theoretical Guarantees

Numerical Performance

Conclusion & Future Work

LSVRG

Choose an epoch length $N > 0$, and at the start of each epoch, store a checkpoint iterate $\bar{w}$ along with $\bar{q} := q^*(\ell(\bar{w}))$ and

$$\nabla R(\bar{w}) = \sum_{i1=}^{n} \bar{q}_i(\nabla \ell_i(\bar{w}) + \mu \bar{w}).$$

At iterate t, sample i_t uniformly from $\{1, \ldots, n\}$ and compute

$$g_t := n\bar{q}_{i_t}(\nabla \ell_{i_t}(w_t) + \mu w_t) - n\bar{q}_{i_t} \nabla \ell_{i_t}(\bar{w}) + \sum_{i=1}^{n} \bar{q}_i \nabla \ell_i(\bar{w}) \, .$$

zero-mean term used for variance reduction

LSVRG

Choose an epoch length $N > 0$, and at the start of each epoch, store a checkpoint iterate $\bar{w}$ along with $\bar{q} := q^*(\ell(\bar{w}))$ and

$$\nabla R(\bar{w}) = \sum_{i1=}^{n} \bar{q}_i(\nabla \ell_i(\bar{w}) + \mu \bar{w}).$$

At iterate t, sample i_t uniformly from $\{1,\ldots,n\}$ and compute

$$g_t := n\bar{q}_{i_t}(\nabla \ell_{i_t}(w_t) + \mu w_t) - n\bar{q}_{i_t} \nabla \ell_{i_t}(\bar{w}) + \sum_{i=1}^{n} \bar{q}_i \nabla \ell_i(\bar{w})\,.$$

Still biased, but bias decreases asymptotically.

$$\mathbb{E}_{P_n}[n\bar{q}_{i_t} \nabla \ell_i(w_t)] = \sum_{i=1}^{n} \bar{q}_i \nabla \ell_i(w) \neq \sum_{i=1}^{n} q_i^*(\ell(w_t)) \nabla \ell_i(w)$$

LSVRG

Choose an epoch length $N > 0$, and at the start of each epoch, store a checkpoint iterate $\bar{w}$ along with $\bar{q} := q^*(\ell(\bar{w}))$ and

$$\nabla R(\bar{w}) = \sum_{i1=}^{n} \bar{q}_i(\nabla \ell_i(\bar{w}) + \mu \bar{w}).$$

At iterate t, sample i_t uniformly from $\{1, \ldots, n\}$ and compute

$$g_t := n\bar{q}_{i_t}(\nabla \ell_{i_t}(w_t) + \mu w_t) - n\bar{q}_{i_t} \nabla \ell_{i_t}(\bar{w}) + \sum_{i=1}^{n} \bar{q}_i \nabla \ell_i(\bar{w}) \, .$$

Perform the update:

$$w_{t+1} = w_t - \eta g_t$$

constant stepsize, as update direction combines bias reduction and variance reduction

Outline

Properties of SRM Objective

Bias and Noise of Current Methods

LSVRG Algorithm

Theoretical Guarantees

Numerical Performance

Conclusion & Future Work

Notation

$R =$ objective function

$P_n =$ sampling distribution used for g_t (e.g. mini-batch sampling)

$w^\star = \operatorname{argmin}_w R(w)$

$\kappa = n\sigma_n L/\mu + 1$

Theorem 1

Assume that $\nu \geq O(G^2/\mu)$. The output of LSVRG with epoch length $N = O(n + \kappa)$ and stepsize $\eta = O(1/(N\mu))$ achieves

$$\mathbb{E}_{P_n^t}\|w_t - w^\star\|_2^2 \lesssim 2^{-\frac{t}{4(n+8\kappa)}}$$

Notation

$R =$ objective function

$P_n =$ sampling distribution used for g_t (e.g. mini-batch sampling)

$w^\star = \mathrm{argmin}_w R(w)$

$\kappa = n\sigma_n L/\mu + 1$

Theorem 1

Assume that $\nu \geq O(G^2/\mu)$. The output of LSVRG with epoch length $N = O(n + \kappa)$ and stepsize $\eta = O(1/(N\mu))$ achieves

$$\mathbb{E}_{P_n^t} \|w_t - w^\star\|_2^2 \lesssim 2^{-\frac{t}{4(n+8\kappa)}}$$

condition number and sample size decoupled, as in variance-reduced algorithms for ERM

Outline

Properties of SRM Objective

LSVRG Algorithm

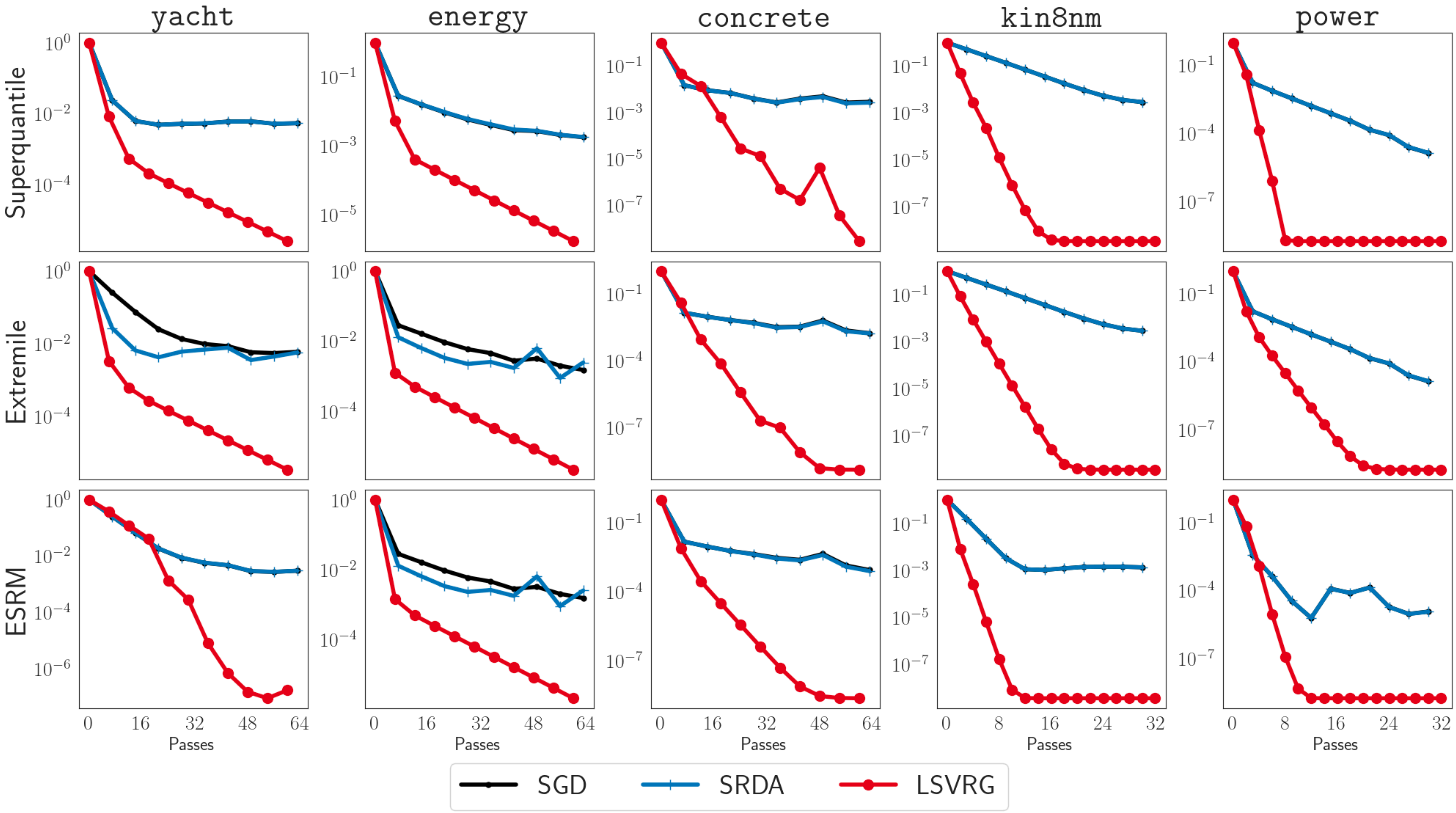
Theoretical Guarantees

Numerical Performance

Conclusion & Future Work

Regression Benchmarks

- We consider five regression tasks, for which we use squared loss under a linear prediction model.
- Datasets are labeled as *yacht*, *energy*, *concrete*, *kin8nm*, and *power*.
- Main metric is training suboptimality $(R(w_t) - R(w^\star))/(R(w_0) - R(w^\star))$.
- Baselines are stochastic gradient descent (SGD), and stochastic regularized dual averaging (SRDA).



yacht
energy
concrete
kin8nm
power
Superquantile
Extremile
ESRM
Passes
SGD
SRDA
LSVRG

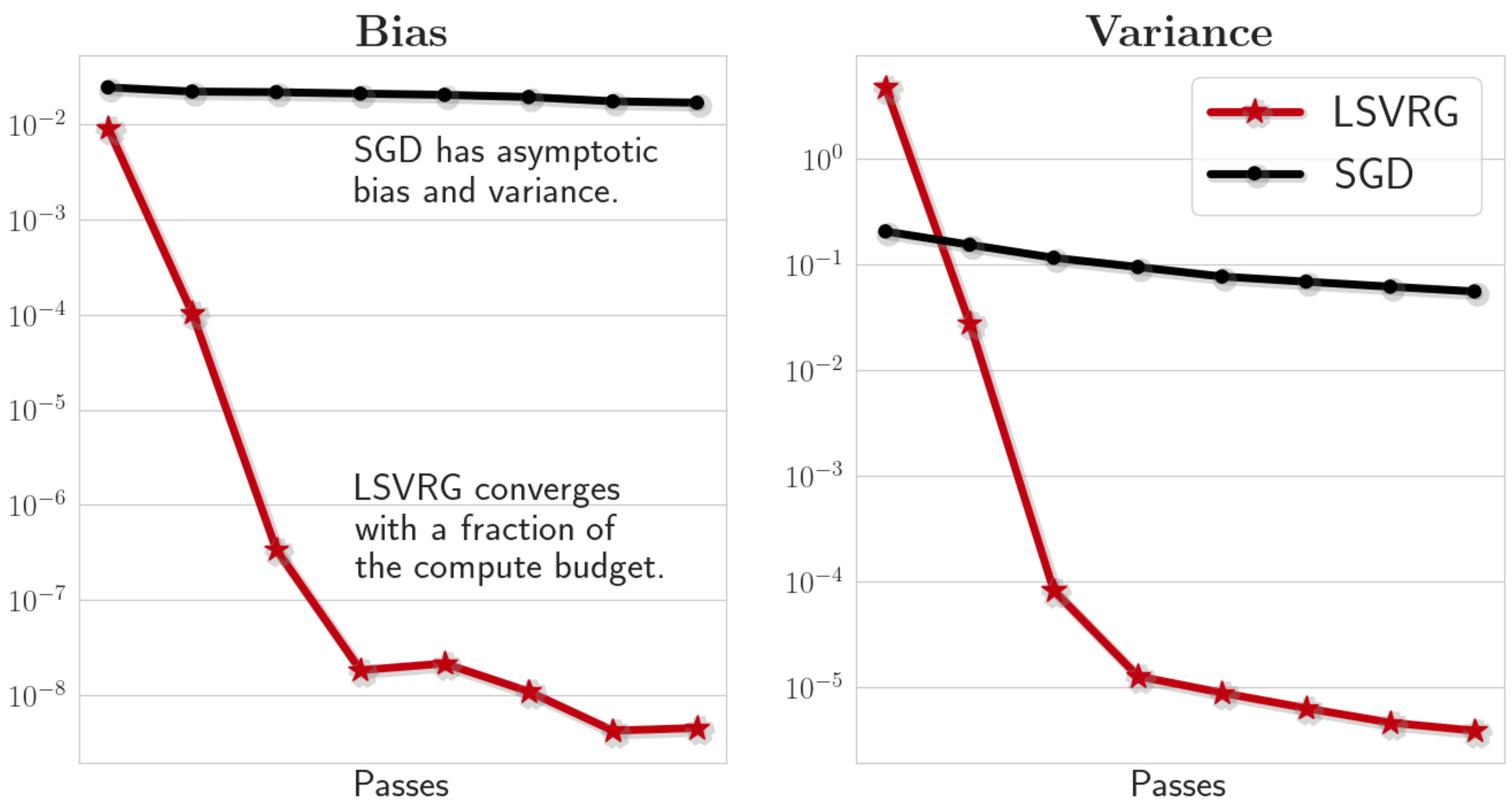
Bias

$\|\mathbb{E}_{P_n}[g_t] - \nabla R(w_t)\|_2^2$

Variance

$\mathbb{E}_{P_n}\|g_t - \mathbb{E}[g_t]\|_2^2$

Superquantile on yacht Benchmark



Outline

Properties of SRM Objective

LSVRG Algorithm

Theoretical Guarantees

Numerical Performance

Conclusion & Future Work

Summary

- We present a stochastic algorithm to optimize spectral risks measures of the empirical loss distribution that:
    - finds an exact minimizer/is asymptotically unbiased
    - makes $O(1)$ calls to a function/gradient oracle per update, and
    - outperforms out-of-the-box convex optimizers on real data.
- Future work includes extensions to the non-convex setting and exploring statistical properties of learned minimizers.

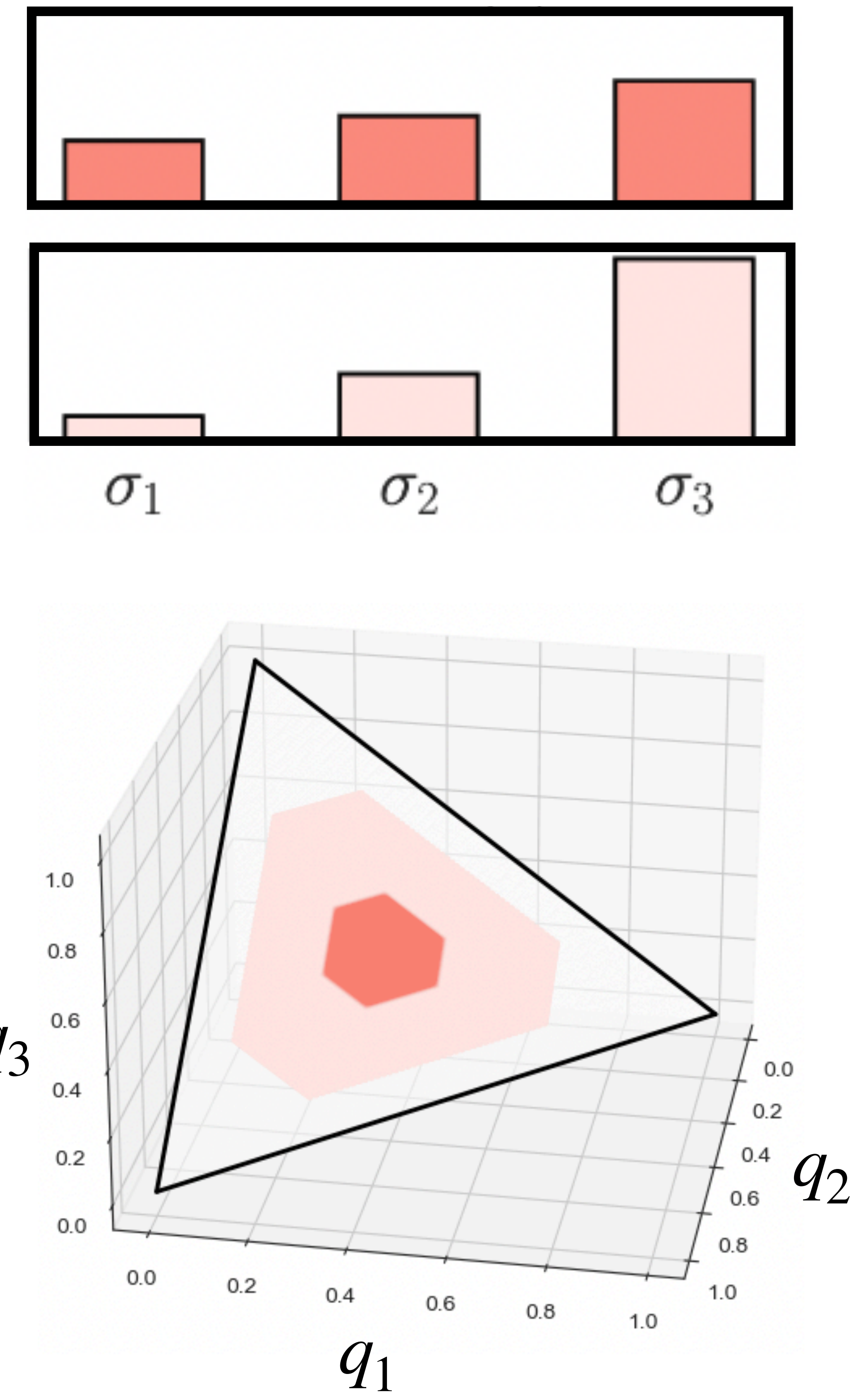
Thank you!

SCAN ME

Spectral risk measures are generated by letting $\mathscr{U}$ be a permutahedron in $\mathbb{R}^n$.

$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathscr{U}} \sum_{i=1}^{n} q_i \ell(w, Z_i) - \nu D(q \| \mathbf{1}_n / n)$$

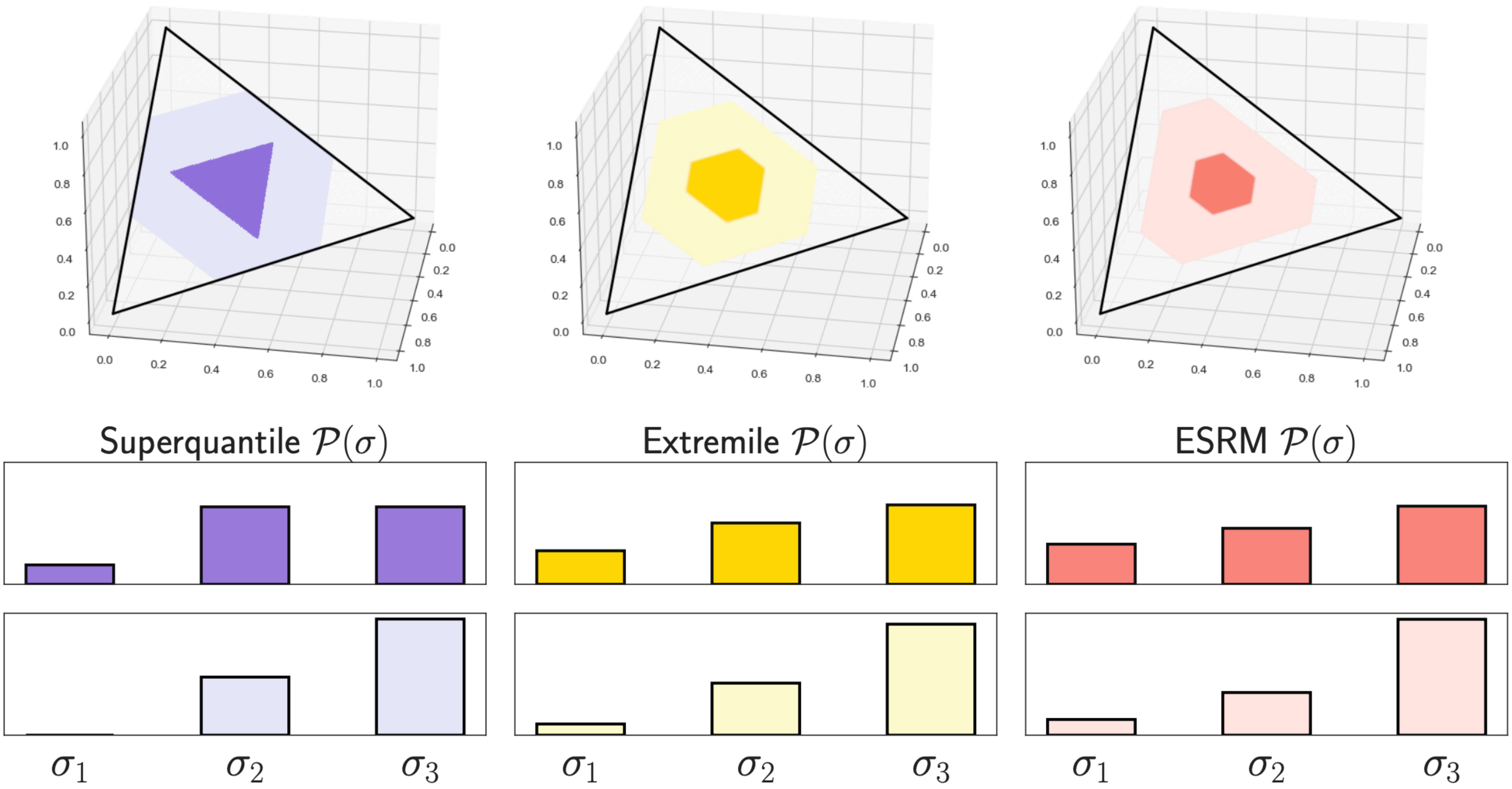
Example for $n = 3$



Spectral Risk Measure

Specify hyperparameter $\sigma = (\sigma_1, \ldots, \sigma_n)$ such that $\sigma_1 \leq \ldots \leq \sigma_n$ and $\sum_{i=1}^{n} \sigma_i = 1$, and use ambiguity set $\mathscr{P}(\sigma)$ by

$$\mathscr{P}(\sigma) = \text{ConvexHull}\{(\sigma_{\pi(1)}, \ldots, \sigma_{\pi(n)}) : \pi \textbf{ is a permutation on } [n]\}$$


Superquantile $\mathcal{P}(\sigma)$
Extremile $\mathcal{P}(\sigma)$
ESRM $\mathcal{P}(\sigma)$
σ_1 σ_2 σ_3
σ_1 σ_2 σ_3
σ_1 σ_2 σ_3

Quantitative Finance & Econometrics

Alternative risk measures (functionals of the loss distribution) and their axiomatic properties are well-studied.

He, 2018; Rockafellar 2007; Cotter, 2006; Acerbi, 2002; Daouia, 2019

Statistics

When $\nu = 0$, SRMs reduce to linear combinations of order statistics, or L-estimators.

Huber, 2009; Shorack, 2017

Spectral Risk Objectives in Machine Learning

Many recent examples of spectral risk-based objectives have appeared in ML, with focus on the superquantile.

Maurer, 2021; Laguel, 2021; Khim, 2020; Holland, 2022

Distributionally Robust Optimization Methods

Optimization approaches rely on full-batch gradient descent, biased SGD, or saddle-point formulations.

Levy 2020; Yu 2022; Yang 2020; Palaniappan, 2016; Kawaguchi & Lu, 2020;

Quantitative Finance & Econometrics

Alternative risk measures (functionals of the loss distribution) and their axiomatic properties are well-studied.

He, 2018; Rockafellar 2007; Cotter, 2006; Acerbi, 2002; Daouia, 2019

Statistics

When $\nu = 0$, SRMs reduce to linear combinations of order statistics, or L-estimators.

Huber, 2009; Shorack, 2017

Spectral Risk Objectives in Machine Learning

Many recent examples of spectral risk-based objectives have appeared in ML, with focus on the superquantile.

Maurer, 2021; Laguel, 2021; Khim, 2020; Holland, 2022

Distributionally Robust Optimization Methods

Optimization approaches rely on full-batch gradient descent, biased SGD, or saddle-point formulations.

Levy 2020; Yu 2022; Yang 2020; Palaniappan, 2016; Kawaguchi & Lu, 2020;

$$R(w) := h_\nu(\ell(w)) + \frac{\mu}{2}\|w\|_2^2$$

$$h_\nu(l) := \max_{q \in \mathscr{P}(\sigma)} q^\top l - \nu n \|q - \mathbf{1}_n/n\|_2^2, \; l \in \mathbb{R}^n$$

$\ell : \mathbb{R}^d \to \mathbb{R}^n$

$$R(w) := h_\nu(\ell(w)) + \frac{\mu}{2}\|w\|_2^2$$

$$h_\nu(l) := \max_{q \in \mathcal{P}(\sigma)} q^\top l - \nu n \|q - \mathbf{1}_n/n\|_2^2, \; l \in \mathbb{R}^n$$

$h_\nu : \mathbb{R}^n \to \mathbb{R}$

$$R(w) := h_\nu(\ell(w)) + \frac{\mu}{2}\|w\|_2^2$$

$$h_\nu(l) := \max_{q \in \mathcal{P}(\sigma)} q^\top l - \nu n \|q - \mathbf{1}_n/n\|_2^2,\ l \in \mathbb{R}^n$$