Stochastic Ordered Empirical Risk Minimization

Joint Statistical Meetings 2022

Ronak Mehta

April 8, 2022

Team

Ronak Mehta
Ph.D. Student

Krishna Pillutla
Ph.D. Student

Vincent Roulet
Acting Assistant Professor

Zaid Harchaoui
Professor

IFDS
Institute for Foundations of Data Science

NSF

UNIVERSITY of WASHINGTON

Motivation: Average-Case ⟶ Worst-Case

- **Current learning paradigm:** optimize average loss across training examples.
- Worst-case performance can often be more important than average-case.

‘I’m the Operator’: The Aftermath of a Self-Driving Tragedy

In 2018, an Uber autonomous vehicle fatally struck a pedestrian. In a WIRED exclusive, the human behind the wheel finally speaks.

2 Killed in Driverless Tesla Car Crash, Officials Say

“No one was driving the vehicle” when the car crashed and burst into flames, killing two men, a constable said.

A Tesla driver is charged in a crash involving Autopilot that killed 2 people

January 18, 2022 · 3:00 PM ET

Usual Setting

- $\ell_i(w) :=$ loss on example i with parameters/weights $w \in \mathbb{R}^d$.
- $\ell_{(i)}(w) :=$ i-th order statistic of $\ell(w) := \big(\ell_1(w), \ldots, \ell_n(w)\big)$.

Empirical Risk Minimization (ERM):

$$\min_w \sum_{i=1}^{n} \frac{1}{n} \ell_i(w)$$

Our Setting

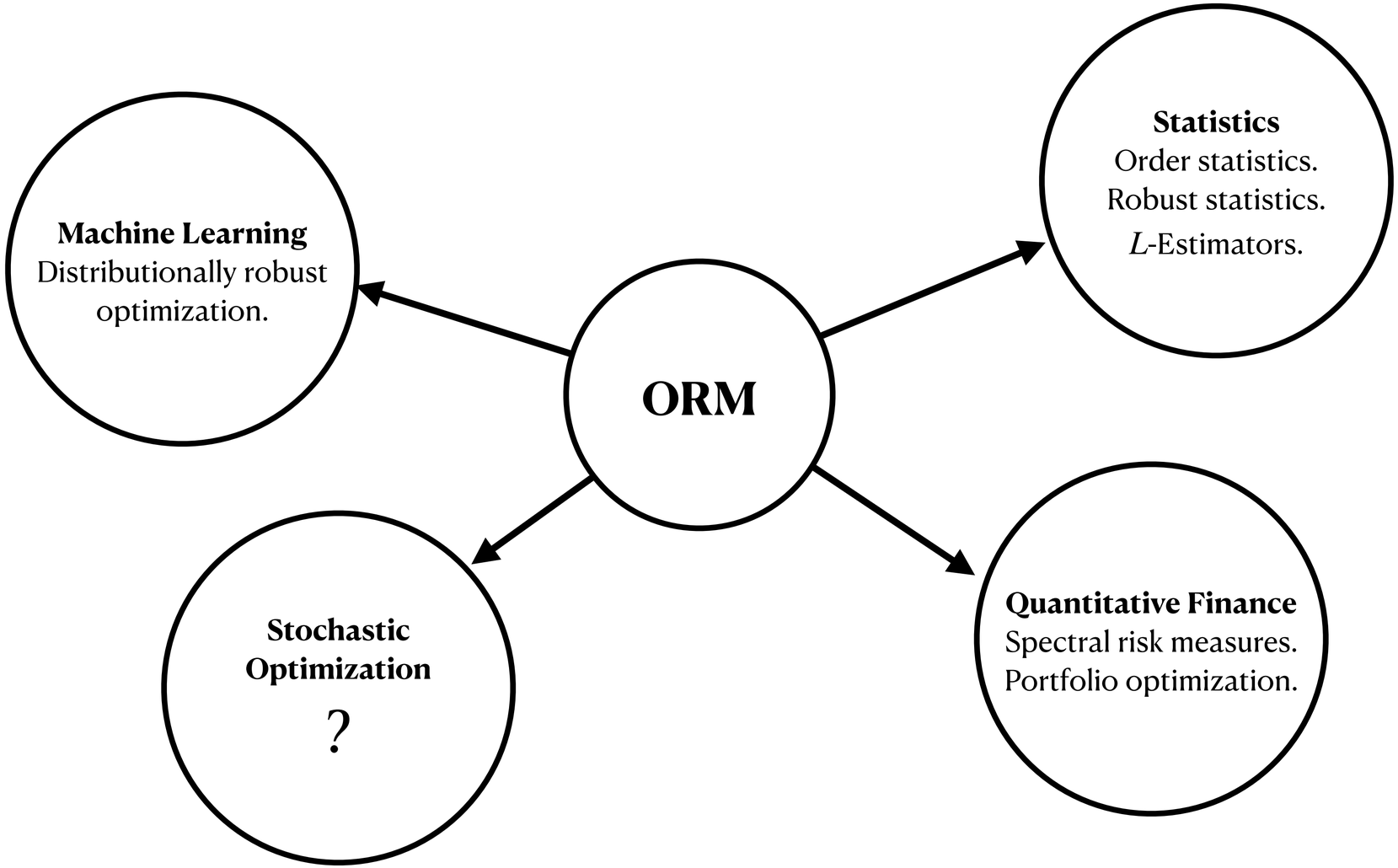
- $\ell_i(w) :=$ loss on example i with parameters/weights $w \in \mathbb{R}^d$.
- $\ell_{(i)}(w) :=$ i-th order statistic of $\ell(w) := \big(\ell_1(w), \ldots, \ell_n(w)\big)$.

Ordered Empirical Risk Minimization (ORM):

$$\min_w \sum_{i=1}^n \frac{1}{n} \ell_i(w) \quad \longrightarrow \quad \min_w \sum_{i=1}^n \alpha_i \ell_{(i)}(w)$$

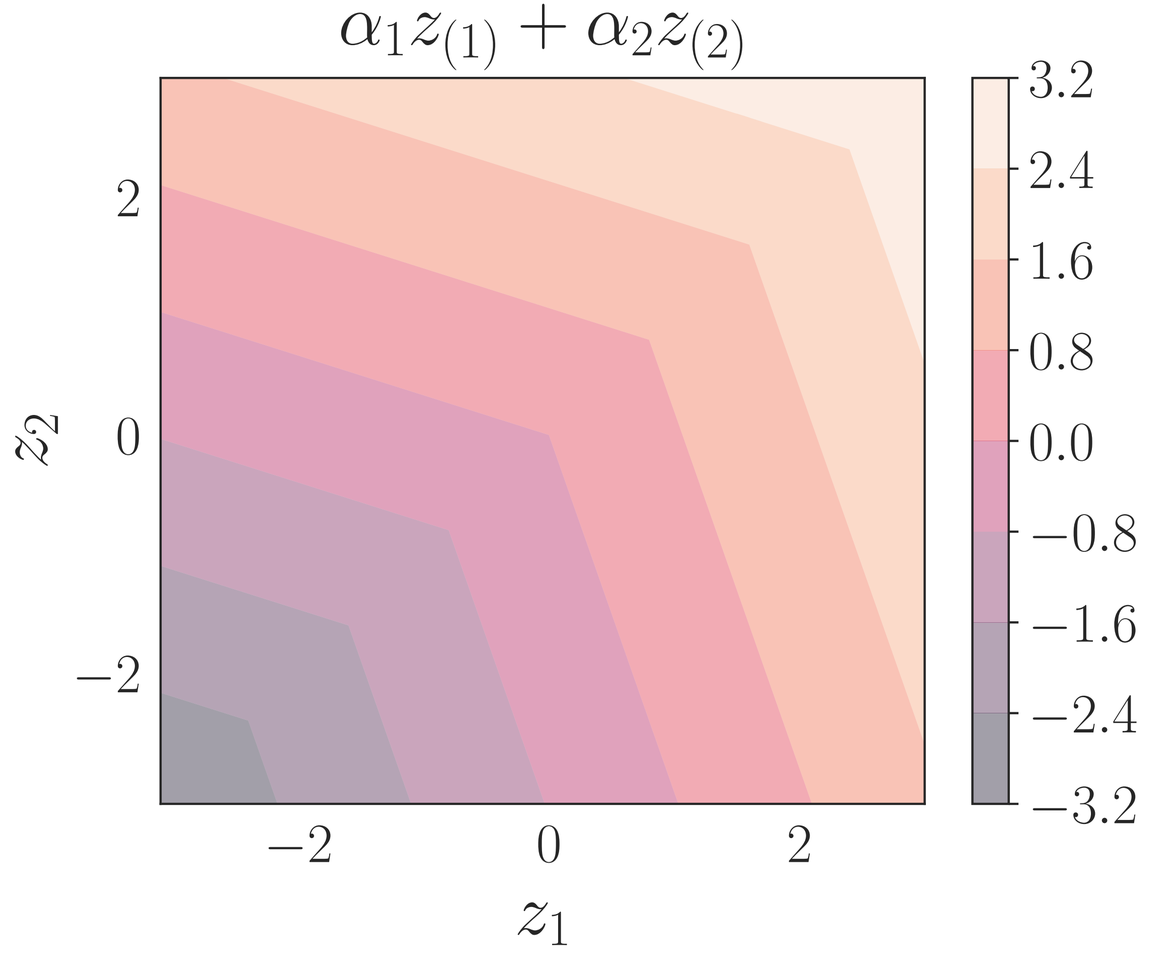
$$0 \le \alpha_1 \le \ldots \le \alpha_n, \sum_{i=1}^n \alpha_i = 1$$

Related Work



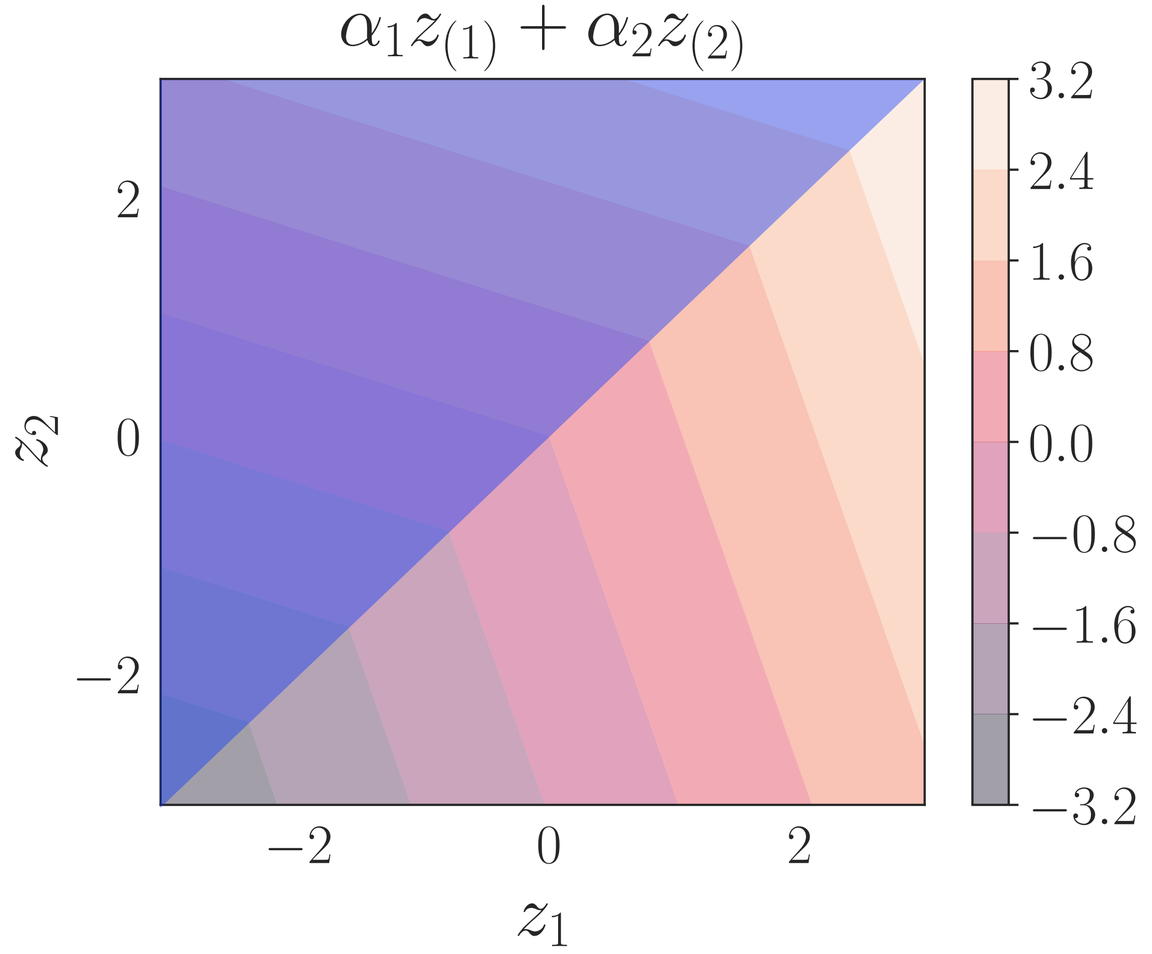
Objective is Piecewise Linear

$$f(z_1, z_2) = 0.3z_{(1)} + 0.7z_{(2)}$$



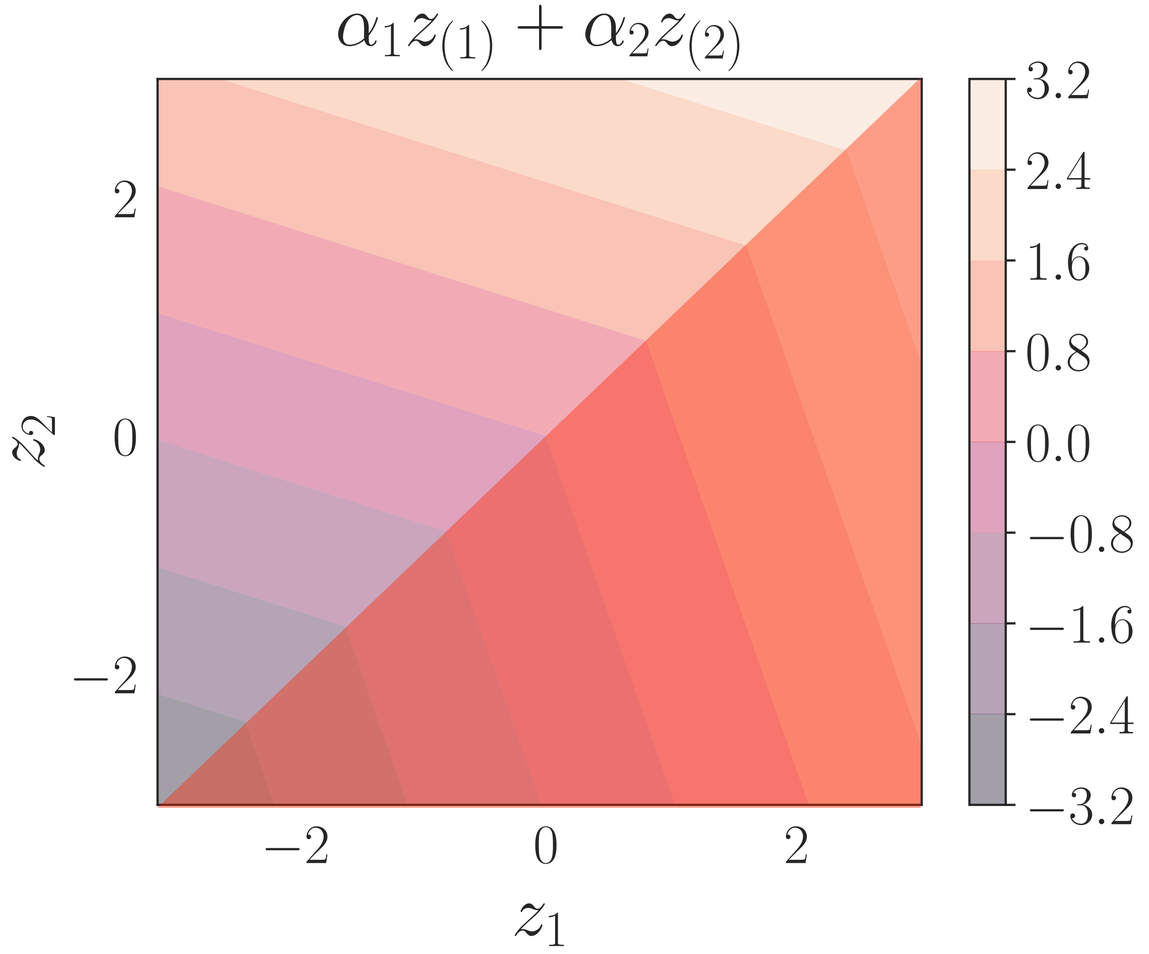
Objective is Piecewise Linear

$$f(z_1, z_2) = 0.3z_{(1)} + 0.7z_{(2)}$$
$$= 0.3z_1 + 0.7z_2$$



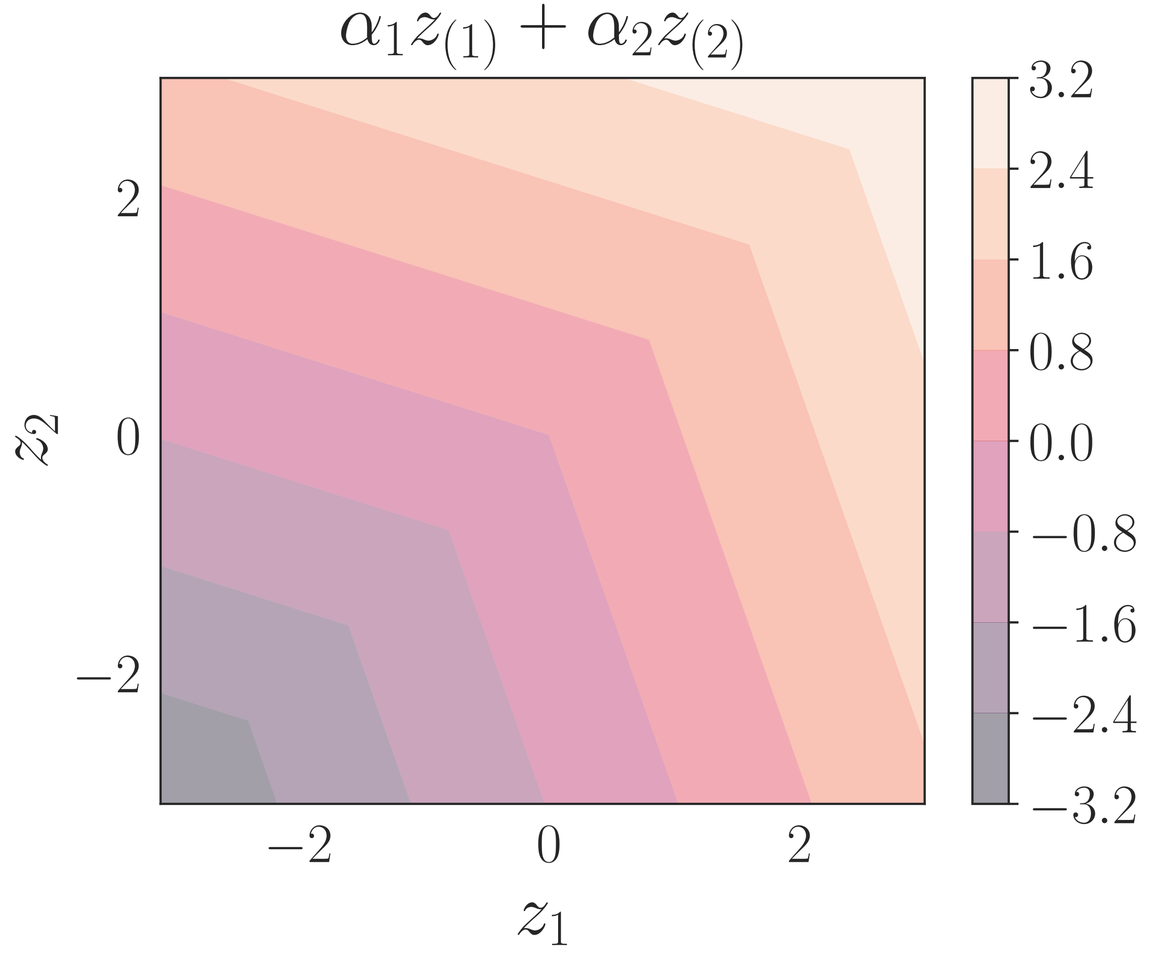
Objective is Piecewise Linear

$$f(z_1, z_2) = 0.3z_{(1)} + 0.7z_{(2)}$$
$$= 0.7z_1 + 0.3z_2$$



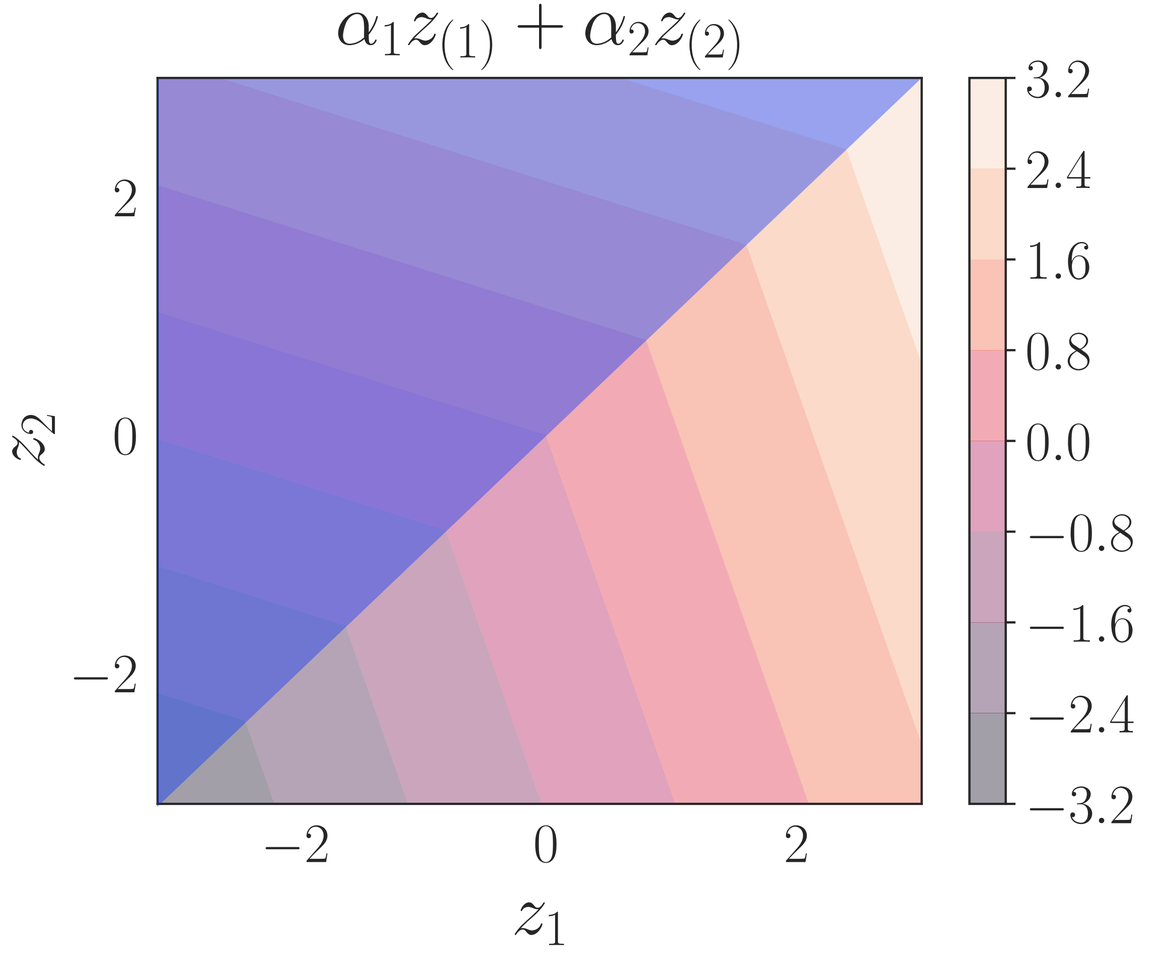
Objective is Piecewise Linear

$$f(z_1, z_2) = \max_{\pi} 0.3z_{\pi(1)} + 0.7z_{\pi(2)}$$



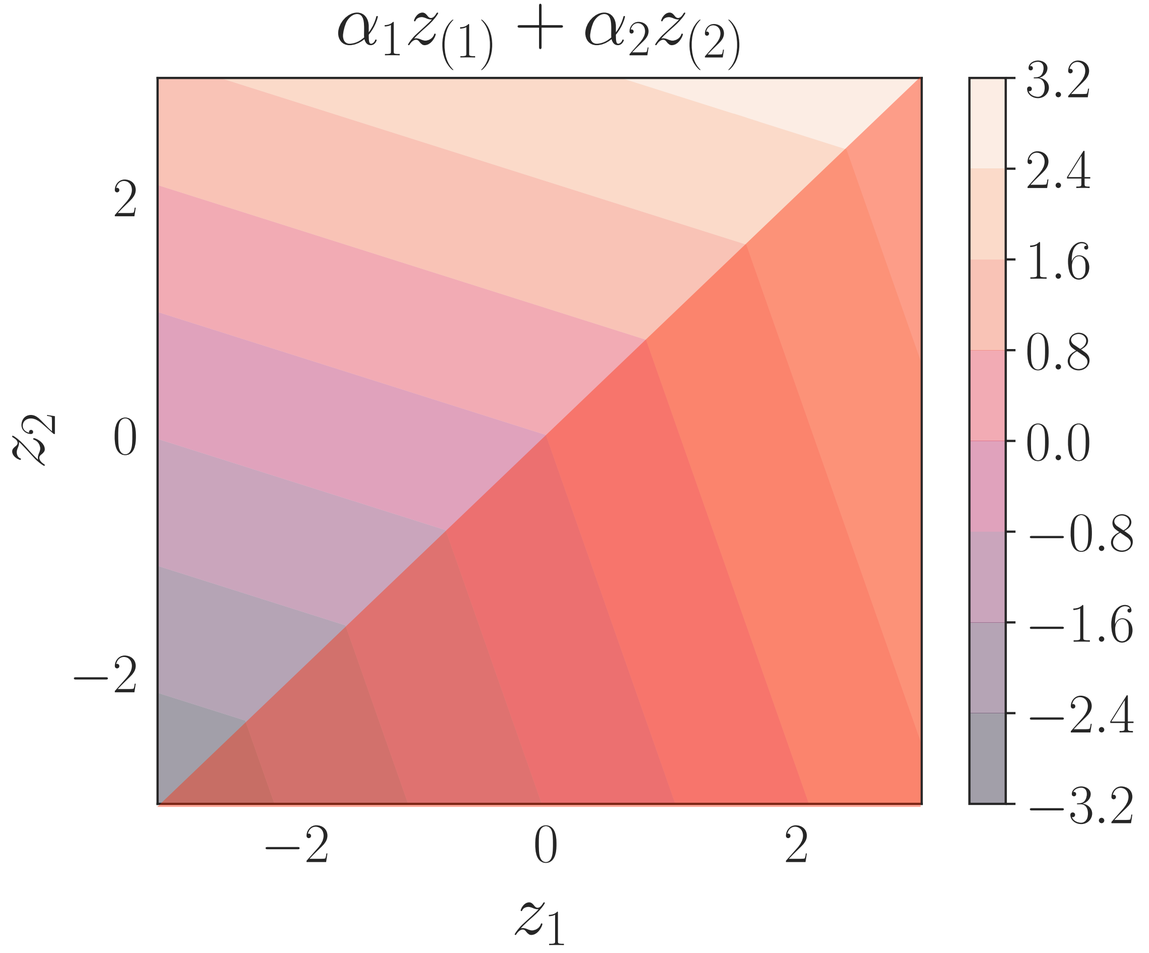
Objective is Piecewise Linear

$$f(z_1, z_2) = \max_{\pi} 0.3 z_{\pi(1)} + 0.7 z_{\pi(2)}$$
$$= 0.3 z_{\pi^*(1)} + 0.7 z_{\pi^*(2)}$$
$$\pi^*(1) = 1$$
$$\pi^*(2) = 2$$



Objective is Piecewise Linear

$$f(z_1, z_2) = \max_{\pi} 0.3z_{\pi(1)} + 0.7z_{\pi(2)}$$
$$= 0.3z_{\pi^*(1)} + 0.7z_{\pi^*(2)}$$
$$\pi^*(1) = 2$$
$$\pi^*(2) = 1$$



Optimization Properties

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- Convex if the ℓ_i's are.

Optimization Properties

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- Convex if the ℓ_i's are.
- Non-smooth.

Optimization Properties

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- Convex if the ℓ_i's are.
- Non-smooth.
- Gradient: $\sum_{i=1}^{n} \alpha_i \nabla \ell_{(i)}(w)$.

Optimization Properties

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- Convex if the ℓ_i's are.
- Non-smooth.
- ~~Gradient~~ Subdifferential: $\operatorname{conv}\left\{ \sum_{i=1}^{n} \alpha_i \nabla \ell_{\pi(i)}(w) : \pi \text{ achieves max} \right\}$.

Algorithms

- **Brute Force:** (batch) gradient descent: $w_{t+1} \leftarrow w_t - \eta_t \sum_{i=1}^{n} \alpha_i \nabla \ell_{(i)}(w_t)$, where η_t is a learning rate.

Algorithms

- **Brute Force:** (batch) gradient descent: $w_{t+1} \leftarrow w_t - \eta_t \sum_{i=1}^{n} \alpha_i \nabla \ell_{(i)}(w_t)$, where η_t is a learning rate.
- **Problem:** $O(n \log n)$ per-iteration time complexity. Need to know α_i weights for unbiased mini-batch estimate of gradient.

Algorithms

- **Brute Force:** (batch) gradient descent: $w_{t+1} \leftarrow w_t - \eta_t \sum_{i=1}^{n} \alpha_i \nabla \ell_{(i)}(w_t)$, where η_t is a learning rate.
- **Problem:** $O(n \log n)$ per-iteration time complexity. Need to know α_i weights for unbiased mini-batch estimate of gradient.
- **Contributions:** stochastic optimization algorithms for ORM objective with $O(1)$ gradient calls per-iteration and extremely fast convergence empirically!

Stochastic Gradient Descent (SGD)

- **Initialization:** Set $w_0 = 0 \in \mathbb{R}^d$.
- **Update:**
  - Sample i_t uniformly.
  - Compute $g_t = \nabla \ell_{i_t}(w_t)$.
  - Set $w_{t+1} = w_t - \eta_t g_t$.

Stochastic Average Gradient Accelerated (SAGA)

- **Initialization:** Set $w_0 = 0 \in \mathbb{R}^d$, store table of gradients $g^{(i)} = \nabla \ell_i(w_0)$ for $i = 1,\ldots,n$, and compute $\bar{g}_0 = \sum_{i=1}^{n} \frac{1}{n} g^{(i)}$.
- **Update:**
  - Sample i_t uniformly.
  - Compute $v_t = \nabla \ell_{i_t}(w_t) - g^{(i_t)} + \bar{g}_t$.
  - Set $w_{t+1} = w_t - \eta_t v_t$.

Stochastic Average Gradient Accelerated (SAGA)

- **Initialization:** Set $w_0 = 0 \in \mathbb{R}^d$, store table of gradients $g^{(i)} = \nabla \ell_i(w_0)$ for $i = 1,\ldots,n$, and compute $\bar{g}_0 = \sum_{i=1}^{n} \frac{1}{n} g^{(i)}$.
- **Update:**
  - Sample i_t uniformly
  - Compute $v_t = \nabla \ell_{i_t}(w_t) - g^{(i_t)} + \bar{g}_t$.
  - Set $w_{t+1} = w_t - \eta_t v_t$.

v_t is an unbiased estimate of the gradient!

$$\mathbb{E}_{i_t|w_t}\left[v_t\right] = \mathbb{E}\left[\nabla \ell_{i_t}(w_t)\right] + \mathbb{E}\left[-g^{(i_t)} + \bar{g}_t\right] = \sum_{i=1}^{n} \frac{1}{n} \nabla \ell_i(w_t)$$

Stochastic Average Gradient Accelerated (SAGA)

- **Initialization:** Set $w_0 = 0 \in \mathbb{R}^d$, store table of gradients $g^{(i)} = \nabla \ell_i(w_0)$ for $i = 1,\ldots,n$, and compute $\bar{g}_0 = \sum_{i=1}^{n} \frac{1}{n} g^{(i)}$.
- **Update:**
  - Sample i_t uniformly.
  - Compute $v_t = \nabla \ell_{i_t}(w_t) - g^{(i_t)} + \bar{g}_t$.
  - Set $w_{t+1} = w_t - \eta_t v_t$.
  - Sample j_t uniformly and update $\bar{g}_{t+1} = \bar{g}_t - \frac{1}{n} g^{(j_t)} + \frac{1}{n} \nabla \ell_{j_t}(w_{t+1})$ and $g^{(j_t)} = \nabla \ell_{j_i}(w_{t+1})$.

Ordered-SAGA (O-SAGA)

- **Initialization:** Set $w_0 = 0 \in \mathbb{R}^d$, store table of losses $\ell^{(i)} = \ell_i(w_0)$ and gradients $g^{(i)} = \nabla \ell_i(w_0)$ for $i = 1,\ldots,n$. Compute $\pi =$ argsort of $\ell^{(i)}$'s and store $\bar{g}_0 = \sum_{i=1}^{n} \alpha_i g^{(\pi(i))}$.
- **Update:**
  - Sample i_t uniformly.
  - Compute $v_t = n\alpha_{i_t} \nabla \ell_{\pi(i_t)}(w_t) - n\alpha_{i_t} g^{(\pi(i_t))} + \bar{g}_t$.
  - Set $w_{t+1} = w_t - \eta_t v_t$.
  - Sample j_t uniformly and update $\ell^{(j_t)} = \ell_{j_t}(w_{t+1})$, $g^{(j_t)} = \nabla \ell_{j_t}(w_{t+1})$, sorting vector π and $\bar{g}_{t+1}$ simultaneously.

O-SAGA Update Direction is Biased

$$\mathbb{E}_{i_t|w_t}\left[v_t\right] = \mathbb{E}\left[n\alpha_{i_t}\nabla\ell_{\pi(i_t)}(w_t) - n\alpha_{i_t}g^{(\pi(i_t))} + \bar{g}_t\right]$$

$$= \sum_{i=1}^{n}\alpha_i\nabla\ell_{\pi(i)}(w_t) - \sum_{i=1}^{n}\alpha_i g^{(\pi(i))} + \bar{g}_t(w_t)$$

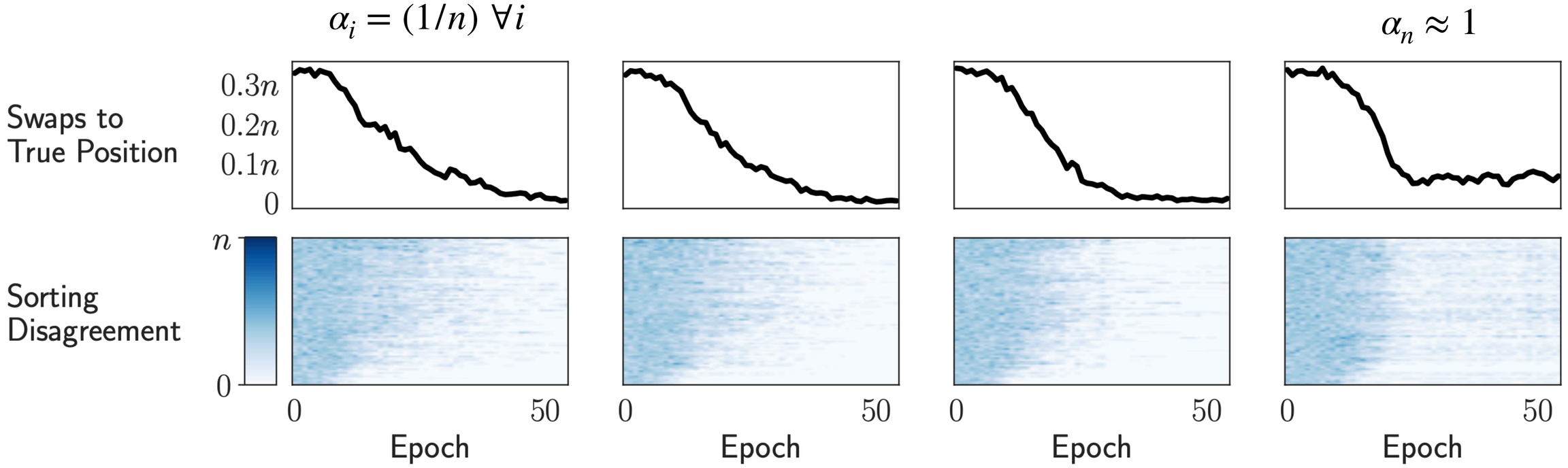
$$= \sum_{i=1}^{n}\alpha_i\nabla\ell_{\pi(i)}(w_t)$$ Sorting from running table (could be stale).

$$\neq \sum_{i=1}^{n}\alpha_i\nabla\ell_{(i)}(w_t)$$ Sorting from $\ell_1(w_t), \ldots, \ell_n(w_t)$ (true gradient).

Bias Decreases with Iteration Number

Error between π and true argsort stabilizes quickly,

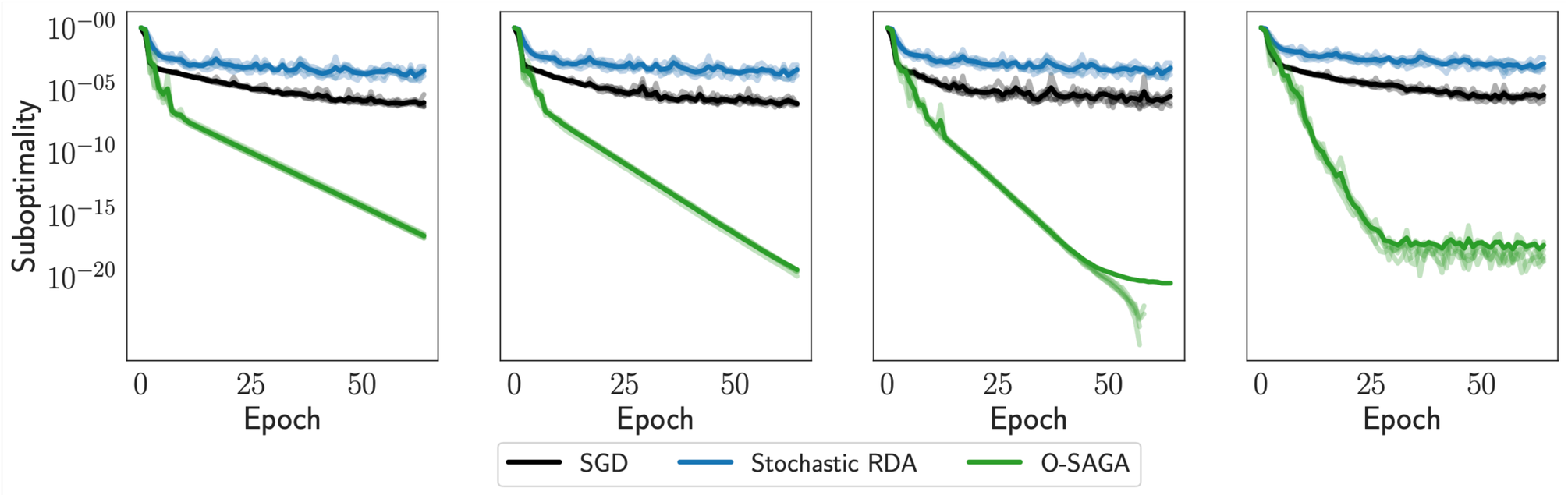
so using the "wrong" argsort still works.



O-SAGA Converges Linearly and Dominates Baselines

$$\text{Suboptimality} = \frac{\text{current value} - \text{min value}}{\text{initial value} - \text{min value}}$$

$\alpha_i = (1/n)\ \forall i$ $\alpha_n \approx 1$



Summary

We present a stochastic algorithm to optimize non-smooth L-statistics of the empirical loss distribution, that

- finds an exact minimizer,
- has bias proportional to suboptimality,
- makes $O(1)$ gradient calls per update, and
- dominates out-of-the-box convex optimizers on synthetic and real data.

Thank you!

Appendix

Notation

- $\ell(w) = \big(\ell_1(w), \ldots, \ell_n(w)\big)$.
- $\alpha_i\big(\ell(w)\big) = j$ such that $\ell_i(w)$ gets multiplied by α_j in $\sum_{i=1}^{n} \alpha_i \ell_{(i)}(w)$, i.e.

$$\sum_{i=1}^{n} \alpha_i \ell_{(i)}(w) = \sum_{i=1}^{n} \alpha_i\big(\ell(w)\big)\, \ell_i(w)$$

- $\alpha\big(\ell(w)\big) = \Big(\alpha_1\big(\ell(w)\big), \ldots, \alpha_n\big(\ell(w)\big)\Big)$.

ORM Optimization

- **Attempt 3:** minibatch stochastic regularized dual averaging (SRDA): out-of-the-box convex optimization method.
- **Problem:** Non-vanishing bias for estimating $\bar{g}_t$, unlike ERM.

ORM Optimization

- **Attempt 2:** minibatch stochastic subgradient method (SGD): $w_{t+1} \leftarrow w_t - \eta_t g_t$, where $g_t = \sum_{j=1}^{m} \alpha'_j \nabla \ell_{(j)}(w_t)$ and $\alpha'_j = \left(j/m\right)^r - ((j-1)/m)^r$.
- **Problem:** Non-vanishing bias for estimating $\bar{g}_t$, unlike ERM.

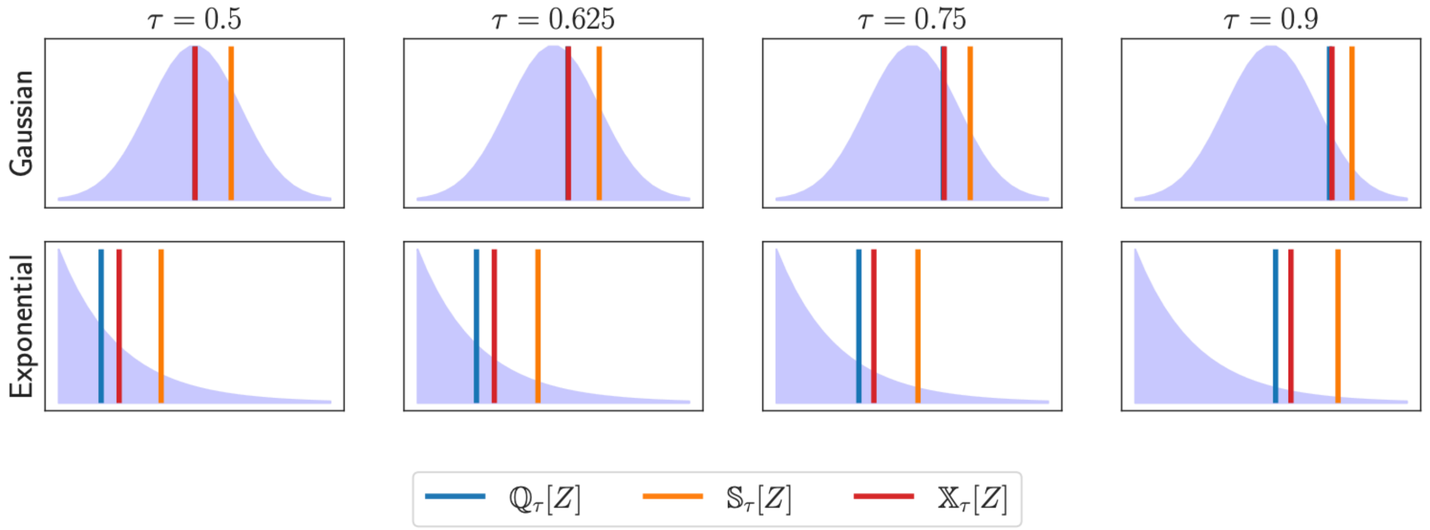
Background

- Let Z be an r.v. with CDF F and *quantile function* $F^{-1}(t) = \inf\{z : F(z) > t\}$.
- Let $a : [0,1] \to \mathbb{R}$ be a *weight function*.
- Population *L-functional* or *spectral risk measure*:

$$\mathbb{L}[Z] = \mathbb{L}[F] = \int F^{-1}(t) \cdot a(t)\ dt$$

Background

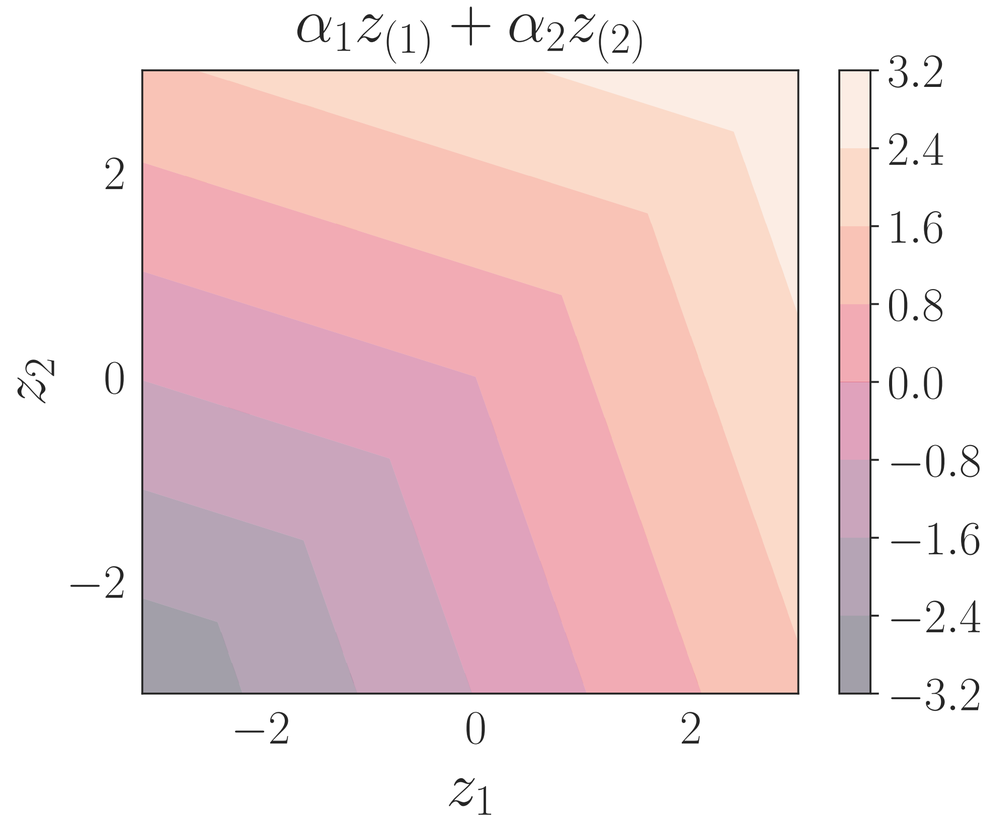
- τ-**Quantile:** $\mathbb{Q}_\tau[Z] = F^{-1}(\tau) \leftrightarrow a(t) =$ point mass at τ.
- τ-**Superquantile:** $\mathbb{S}_\tau[Z] = \mathbb{E}\left[Z \mid Z > \mathbb{Q}_\tau[Z]\right] \leftrightarrow a(t) = 1/(1-\tau) \cdot I_{[\tau,1]}(t)$.
- τ-**Extremile:** $\mathbb{X}_\tau[Z] = \mathbb{E}\left[\max\{Z_1, \ldots, Z_r\}\right] \leftrightarrow a(t) = t^r$ with $r = -(1/\log_2(\tau))$.

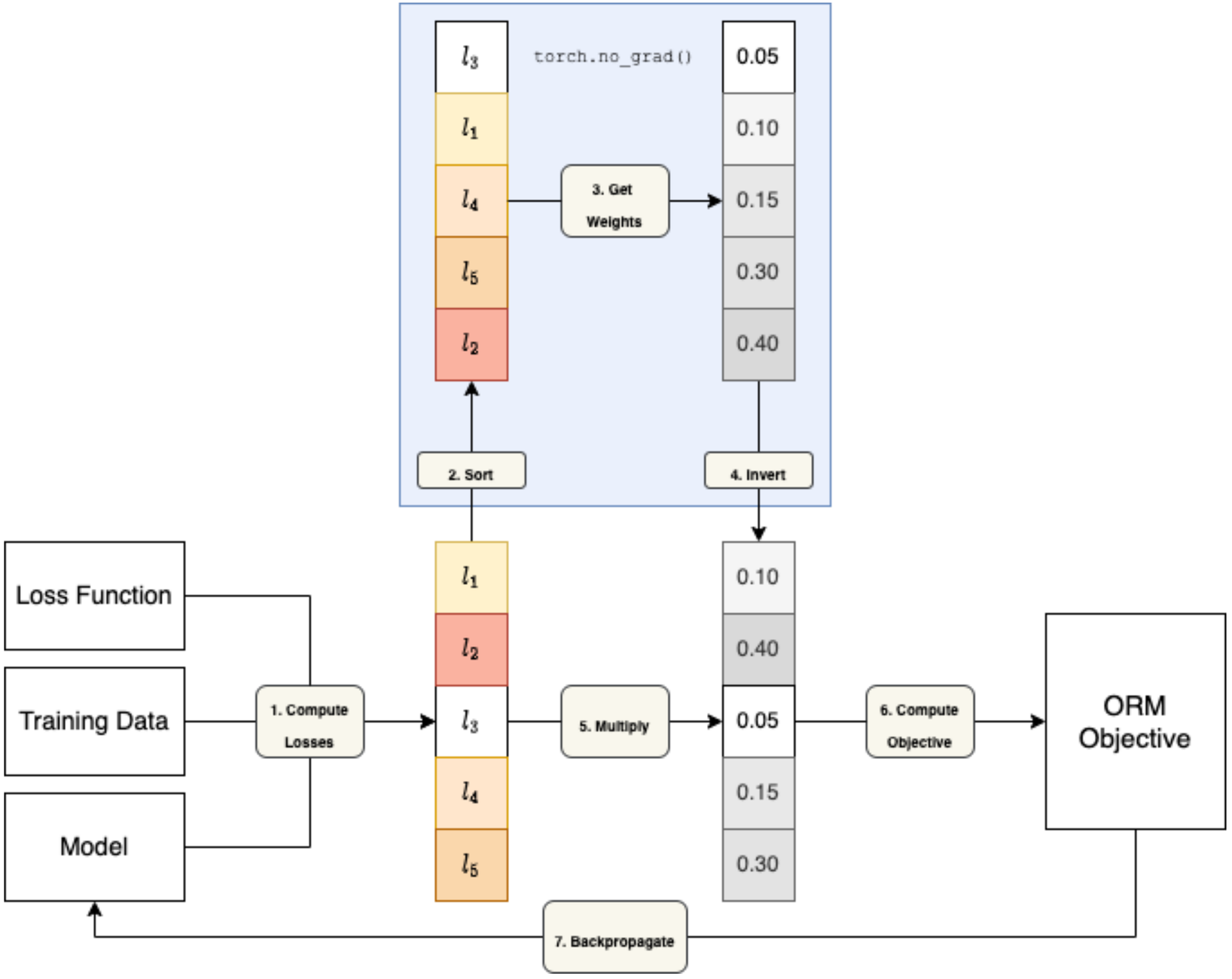


Background

- Consider sample $z_1, \ldots, z_n$ with empirical distribution F_n and *order statistics* $z_{(1)} \leq \ldots \leq z_{(n)}$.

$$
\begin{aligned}
\mathbb{L}[F_n] &= \int_0^1 F_n^{-1}(t) \cdot a(t)\ dt \\
&= \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} z_{(i)} \cdot a(t)\ dt \\
&= \sum_{i=1}^{n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} a(t)\ dt \right) z_{(i)} \\
&= \sum_{i=1}^{n} \alpha_i z_{(i)}
\end{aligned}
$$




torch.no_grad()
l_3
l_1
l_4
l_5
l_2
3. Get Weights
0.05
0.10
0.15
0.30
0.40
2. Sort
4. Invert
Loss Function
Training Data
Model
1. Compute Losses
l_1
l_2
l_3
l_4
l_5
5. Multiply
0.10
0.40
0.05
0.15
0.30
6. Compute Objective
ORM Objective
7. Backpropagate

ORM Objective

$$\sum_{i=1}^{n} \alpha_i \ell_{(i)}(w)$$

- Setting $\alpha_i = (1/n)$ recovers average, $\alpha_n = 1$ recovers maximum.
- Class of *superquantiles* $\alpha_i = 1/\big(n(1-\tau)\big)\, I_{i \geq \tau n}$ for $\tau \in [0,1)$.
- Class of *extremiles* $\alpha_i = (i/n)^r - ((i-1)/n)^r$ for $r \in [1,\infty)$.

Connecting Views

- Picture with combining densities and stuff

Setting: Population View

- Z is a random variable with iid copies $Z_1, \ldots, Z_n$ (e.g. training examples).
- Losses $\ell_i(w) = l(w, Z_i)$ are realizations of $l(w, Z) \sim F$.

Stochastic Optimization with Expectation:

$$\min_w \mathbb{E}\left[l(w, Z)\right]$$

Setting: Population View

- Z is a random variable with iid copies $Z_1, \ldots, Z_n$ (e.g. training examples).
- Losses $\ell_i(w) = l(w, Z_i)$ are realizations of $l(w, Z) \sim F$.

Stochastic Optimization with Tail Risk Measure:

$$\min_w \mathbb{E}\left[l(w, Z)\right] \longrightarrow \min_w \mathbb{T}\left[l(w, Z)\right]$$

Connecting Views

$$\sum_{i=1}^{n} x_i \cdot a\left(F_n(x_i)\right) \qquad \approx \qquad \int x \cdot a(F(x))\ dx$$

Connecting Views

$$\sum_{i=1}^{n} x_i \cdot a\left(F_n(x_i)\right) \qquad \approx \qquad \int x \cdot a(F(x))\ dx$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(F_n(x_{(i)})\right)$$

Connecting Views

$$\sum_{i=1}^{n} x_i \cdot a\left(F_n(x_i)\right) \qquad \approx \qquad \int x \cdot a(F(x))\ dx$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(F_n(x_{(i)})\right)$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(\frac{i}{n}\right)$$

Connecting Views

$$\sum_{i=1}^{n} x_i \cdot a\left(F_n(x_i)\right) \qquad \approx \qquad \int x \cdot a(F(x))\ dx$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(F_n(x_{(i)})\right)$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(\frac{i}{n}\right)$$

$$\sum_{i=1}^{n} x_{(i)} \cdot \alpha_i$$

Connecting Views

$$\sum_{i=1}^{n} x_i \cdot a\left(F_n(x_i)\right) \quad \approx \quad \int x \cdot a(F(x))\ dx$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(F_n(x_{(i)})\right)$$

$$\sum_{i=1}^{n} x_{(i)} \cdot a\left(\frac{i}{n}\right)$$

$$\sum_{i=1}^{n} x_{(i)} \cdot \alpha_i \quad \approx \quad \int_0^1 F^{-1}(t) \cdot a(t)\ dt$$

L-Statistics & Spectral Risk Measures

- LHS is an *L-statistic,* or linear combination of order statistics.
- RHS is a *spectral risk measure*, or weighted integral of a quantile function.

$$\sum_{i=1}^{n} x_{(i)} \cdot \alpha_i \approx \int_0^1 F^{-1}(t) \cdot a(t)\ dt$$

L-Statistics & Spectral Risk Measures

- LHS is an *L-statistic,* or linear combination of order statistics.
- RHS is a *spectral risk measure*, or weighted integral of a quantile function.

$$\sum_{i=1}^{n} \ell_{(i)}(w) \cdot \alpha_i = \sum_{i=1}^{n} x_{(i)} \cdot \alpha_i \approx \int_0^1 F^{-1}(t) \cdot a(t)\ dt$$

L-Statistics & Spectral Risk Measures

- LHS is an *L-statistic,* or linear combination of order statistics.
- RHS is a *spectral risk measure*, or weighted integral of a quantile function.

$$\sum_{i=1}^{n} \ell_{(i)}(w) \cdot \alpha_i = \sum_{i=1}^{n} x_{(i)} \cdot \alpha_i \approx \int_0^1 F^{-1}(t) \cdot a(t)\ dt = \mathbb{T}\left[l(w, Z)\right] = \mathbb{T}[F]$$

Statistical View of Objective

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- LHS is an *L-statistic*, or linear combination of order statistics.

Statistical View of Objective

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

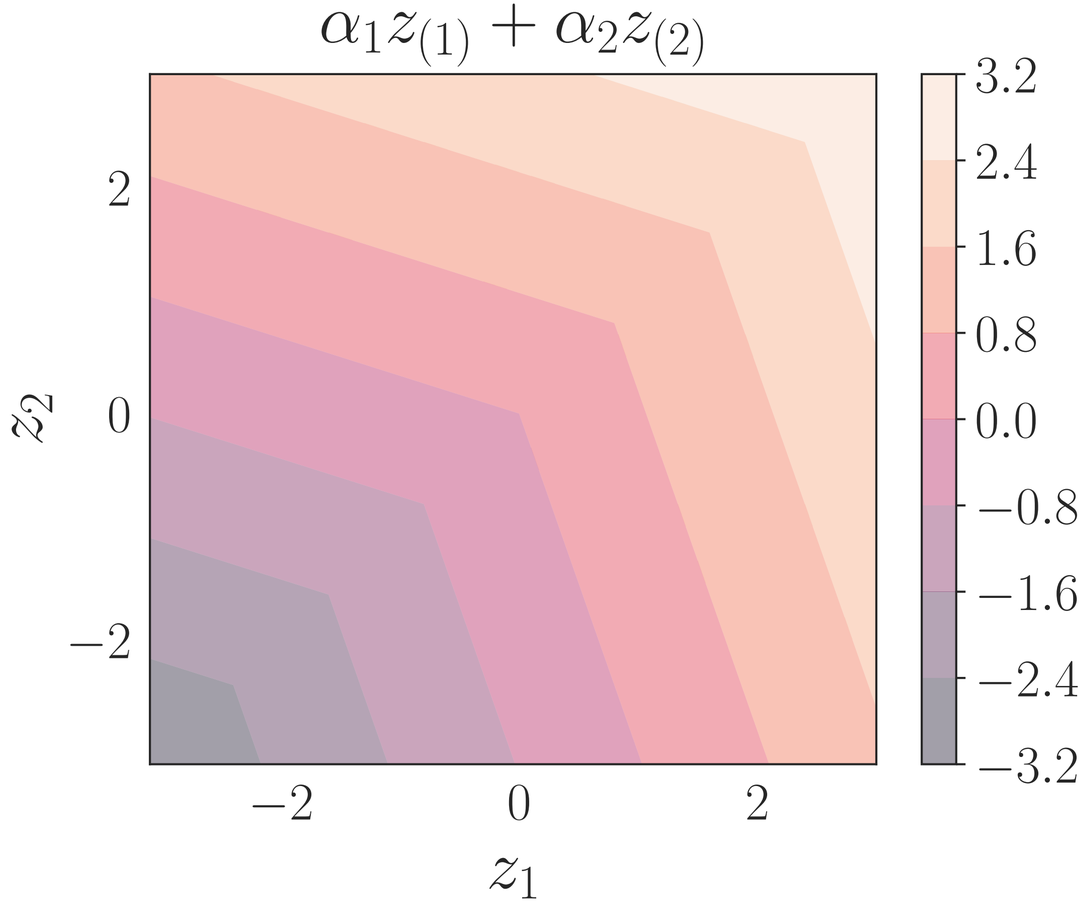
- LHS is an *L-statistic*, or linear combination of order statistics.
- Order statistic $\ell_{(i)}(w)$ is the (i/n)-th quantile of the empirical distribution of losses.

Optimization View of Objective

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- Convex if the ℓ_i's are convex.
- Non-smooth.
- Gradient:

$$\sum_{i=1}^{n} \alpha_i \nabla \ell_{(i)}(w)$$

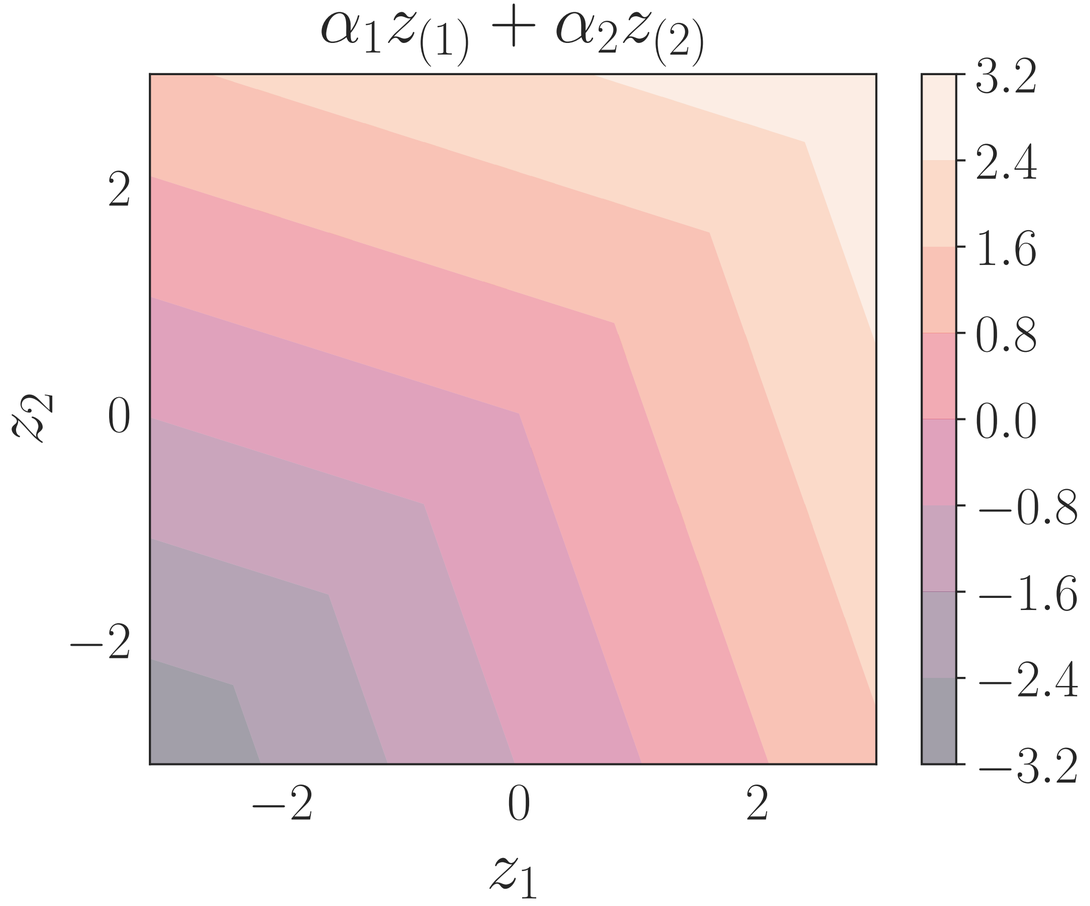


Optimization View of Objective

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- Convex if the ℓ_i's are convex.
- Non-smooth.
- ~~Gradient~~ Subdifferential:

$$\operatorname{conv}\left\{ \sum_{i=1}^{n} \alpha_i \nabla \ell_{\pi(i)}(w) : \pi \text{ achieves max} \right\}$$



Statistical View of Objective

$$\sum_{i=1}^{n} \alpha_i \cdot \ell_{(i)}(w) = \max_{\pi} \sum_{i=1}^{n} \alpha_i \cdot \ell_{\pi(i)}(w)$$

- LHS is an *L-statistic*, or linear combination of order statistics.
- Order statistic $\ell_{(i)}(w)$ is the (i/n)-th quantile of the empirical distribution of losses.
- LHS also a *spectral risk measure*, or weighted average of the (empirical) quantile function.