Stochastic Optimization for Spectral Risk Measures

Ronak Mehta **Vincent Roulet**[1] **Krishna Pillutla**[1] **Lang Liu** **Zaid Harchaoui**
University of Washington

Abstract

Spectral risk objectives – also called L-risks – allow for learning systems to interpolate between optimizing average-case performance (as in empirical risk minimization) and worst-case performance on a task. We develop LSVRG, a stochastic algorithm to optimize these quantities by characterizing their subdifferential and addressing challenges such as biasedness of subgradient estimates and non-smoothness of the objective. We show theoretically and experimentally that out-of-the-box approaches such as stochastic subgradient and dual averaging can be hindered by bias, whereas our approach exhibits linear convergence.

1 INTRODUCTION

A cornerstone of machine learning is the *empirical risk minimization (ERM)* problem, written

$$\min_{w\in\mathbb{R}^d}\left[\mathcal{R}(w) := \frac{1}{n}\sum_{i=1}^n \ell_i(w)\right], \tag{1}$$

where $\ell_i(w)$ quantifies loss on training example i using a model with weights $w \in \mathbb{R}^d$. The objective (1) represents an often unquestioned modeling choice: to summarize $\ell_1(w), ..., \ell_n(w)$, the empirical sample of losses, using its average. At first glance, this is a natural summary, inheriting both the statistical convenience of the sample mean (Shalev-Shwartz and Ben-David, 2014) and the wide arsenal of optimization algorithms designed specifically for finite sum objectives (Le Roux et al., 2012; Defazio et al., 2014; Johnson and Zhang, 2013; Reddi et al., 2016). However, as modern learning systems are deployed in critical domain applications such as energy planning (Guigues and Sagastizábal, 2013), materials engineering (Yeh, 2006), and financial regulation (He et al., 2022), safe and reliable performance in "worst-case" scenarios is paramount.

This imperative can be modeled by alternate *risk measures* (statistical functionals of the loss distribution), particularly those that encapsulate the behavior of the distribution's upper tail. We investigate objectives of the form

$$\min_{w\in\mathbb{R}^d}\left[\mathcal{R}_\sigma(w) := \sum_{i=1}^n \sigma_i \ell_{(i)}(w)\right], \tag{2}$$

where $\ell_{(1)}(w) \le ... \le \ell_{(n)}(w)$ are the *order statistics* of the losses, and $0 \le \sigma_1 \le \cdots \le \sigma_n \le 1$ is a sequence of non-decreasing weights satisfying $\sum_{i=1}^n \sigma_i = 1$, called the *spectrum* of $\mathcal{R}_\sigma$.

The expression (2) is called an L-*estimator* (Shorack, 2017; Maurer et al., 2021) for a generic linear combination of order statistics and an L-*risk* when the ordered elements are losses incurred on a training set (Maurer et al., 2021; Khim et al., 2020). The σ_i's allow the practitioner to interpolate between the average-case ($\sigma_i = 1/n\ \forall i$) and worst-case ($\sigma_n = 1$) performance on the training set. Such objectives have garnered a flurry of recent interest in machine learning (Fan et al., 2017; Williamson and Menon, 2019; Khim et al., 2020; Maurer et al., 2021; Holland and Mehdi Haress, 2022; Leqi et al., 2019; Lee et al., 2020; Kawaguchi and Lu, 2020).

Despite their increasing adoption, however, optimization approaches have relied on using the full-batch or stochastic subgradient method out-of-the-box (Fan et al., 2017; Kawaguchi and Lu, 2020; Laguel et al., 2020; Levy et al., 2020), both enduring considerable limitations. The per-iteration complexity of full-batch methods is $O(n)$ function/gradient evaluations and $O(n \log n)$ elementary operations (as we discuss in Prop. 2). For stochastic[2] variants, unbiased estimates of any subgradient, while needing only $O(1)$ gradient evaluations, still need $O(n)$ function calls and $O(n \log n)$ elementary operations, yielding the same per-iteration complexity as the full-batch method in automatic differentiation frameworks. A number of methods abandon convergence to the minimal L-risk altogether and resort to $O(1)$-time stochastic subgradient updates, but are biased (Kawaguchi and Lu, 2020; Levy et al., 2020).

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

[1] Now at Google Research

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

Given the relevance of stochastic training algorithms to machine learning, the question remains whether there exist optimization algorithms that converge to the minimum L-risk while needing only $O(1)$ gradient calls per iteration. In Sec. 2, we show the consistency of the empirical L-risks for their population counterparts. In Sec. 3, we characterize the subdifferential and continuity properties of L-risks as a function of the underlying losses and quantify the bias of current stochastic approaches. We propose LSVRG, an algorithm that converges linearly to a smoothed approximation of the L-risk requiring $O(1)$ function/gradient evaluations and $O(\log n)$ elementary operations per iteration. Finally, we demonstrate superior convergence of LSVRG experimentally on the non-smooth objective via numerical evaluations in Sec. 4, with concluding remarks in Sec. 5.

Related Work Risk measures have been studied extensively in quantitative finance (Artzner et al., 1999; Föllmer and Schied, 2002; Rockafellar and Uryasev, 2013; Acerbi and Tasche, 2002; Pflug and Ruszczyński, 2005; Kuhn et al., 2019), convex analysis (Rockafellar and Royset, 2014; Ben-Tal and Teboulle, 2007), and distributionally robust learning (Sarykalin et al., 2008; Guigues and Sagastizábal, 2013; Fan et al., 2017; Hu et al., 2018; Lee and Raginsky, 2018; Duchi and Namkoong, 2019; Laguel et al., 2020; Chen and Paschalidis, 2020; Li et al., 2021). We refer to He et al. (2022) for a review of the axiomatic theory of risk measures and Shapiro et al. (2014, Chap. 6) for applications to optimization.

A number of recent works study L-risks, with a focus on statistical properties. The works Khim et al. (2020) and Maurer et al. (2021) provide classical statistical learning bounds for L-risk objectives and the latter focuses on unsupervised tasks like clustering. Holland and Mehdi Haress (2022) present a derivative-free learning procedure for general L-risk problems in the fully stochastic/streaming setting. A particular risk measure called the *superquantile* or *conditional value-at-risk (CVaR)*, has recently received careful attention in the learning setting (Curi et al., 2020; Levy et al., 2020; Laguel et al., 2020, 2021). Other risk measures include cumulative prospect theory (CPT) measures and optimized certainty equivalent (OCE) measures (Leqi et al., 2019; Lee et al., 2020).

Fan et al. (2017) and Kawaguchi and Lu (2020) study batch and stochastic optimization algorithms respectively for the "average top-k" loss, which is exactly equivalent to the superquantile. We instead focus on developing incremental algorithms, akin to those for ERM (Mairal, 2014; Le Roux et al., 2012; Defazio et al., 2014; Shalev-Shwartz and Zhang, 2013; Johnson and Zhang, 2013), which apply to all L-risks. We aim to find algorithms that operate on non-smooth objectives, a fixed training set, and require only a constant number of function value and gradient computations per iterate.

2 SPECTRAL RISK MEASURES

In this section, we relate the empirical quantity (2) to its population counterpart, justifying its use as an estimator for n sufficiently large. To achieve this, we will write L-risks as functionals of an empirical cumulative distribution function (CDF), and show that it consistently estimates the value of the same functional applied to a population CDF.

Notation Let $\{D_1, \dots, D_n\}$ be an i.i.d. sample from a distribution $\mathbb{P}$ over a sample space $\mathcal{D}$. Let $\ell : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$ be a loss function consuming model weights $w \in \mathbb{R}^d$ and $\mathcal{D}$-valued training example D (e.g., a feature-label pair). We denote the training loss as $\ell_i(w) := \ell(w, D_i)$ for short. Let $Z_i := \ell(w, D_i)$ for $i \in \{1, \dots, n\}$. It follows that $\{Z_1, \dots, Z_n\}$ is a real-valued i.i.d. sample whose CDF is denoted by F, and the L-risk (2) reads

$$\mathcal{R}_\sigma(w) = \sum_{i=1}^{n} \sigma_i \ell_{(i)}(w) = \sum_{i=1}^{n} \sigma_i Z_{(i)}, \tag{3}$$

where $Z_{(1)} \leq \dots \leq Z_{(n)}$ are order statistics of $\{Z_i\}_{i=1}^n$.

We describe subsequent results as if $\{Z_i\}_{i=1}^n$ are arbitrary real-valued random variables drawn i.i.d. from CDF F, keeping in mind that in our case, these refer to losses on data instances D_i under parameter vector w.

Spectral Risk Measures We rewrite the L-risk (3) as a functional of the CDF known as a *spectral risk measure* (Acerbi and Tasche, 2002). To do this, let $F_n(z) := \frac{1}{n}\sum_{i=1}^n 1_{(-\infty, z]}(Z_i)$ denote the (random) empirical CDF of the sample and define the empirical *quantile function* (or inverse CDF) as $F_n^{-1}(t) := \inf\{z : F_n(z) \geq t\}$ for $t \in (0, 1)$. The population quantile function is defined similarly as $F^{-1}(t) := \inf\{z : F(z) \geq t\}$. The empirical quantile function can be written in terms of the order statistics as $F_n^{-1}(t) = Z_{(\lceil nt \rceil)}$, as seen in Fig. 1 (top left). Notice in particular that when $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right)$, we have that $F_n^{-1}(t) = Z_{(i)}$, where end-points are chosen to make F_n^{-1} left continuous.

The spectrum σ of an L-risk is typically defined as a discretization of a probability density s on $(0, 1)$, such that $\sigma_i = \int_{(i-1)/n}^{i/n} s(t)\, \mathrm{d}t$, so that it need not be redefined for every n. Examples of spectra for various risk measures are shown in Fig. 1 (bottom), in which the value of σ_i is equal to the area of the shaded region immediately under it. The associated formulae are in Tab. 1. The superquantile with parameter $q \in (0, 1)$ has enjoyed much attention in quantitative finance and more recently, machine learning (Laguel et al., 2021), the *extremile* with parameter $r \geq 1$ has been introduced by (Daouia et al., 2019) as an alternative risk measure, and the *exponential spectral risk measure (ESRM)* with parameter $\rho > 0$ is used in futures clearinghouse margin requirements (Cotter and Dowd, 2006).

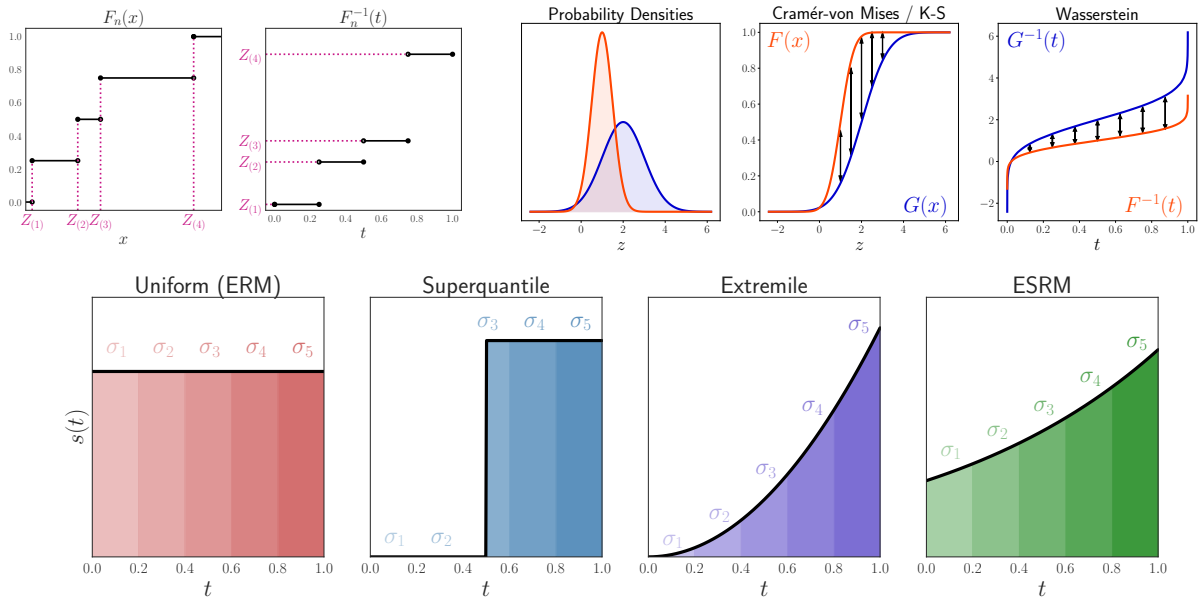

Figure 1: **Top Left:** Empirical CDF F_n and quantile function F_n^{-1} of $Z_1, \cdots, Z_4$. **Top Right:** Comparison of two distribution in CDFs (F and G) as well as quantile functions (F^{-1} and G^{-1}). **Bottom:** Continuous spectra $s(t)$ and their discretization $(\sigma_1, \ldots, \sigma_5)$ for various risk measures.

Given both the construction of s and F_n^{-1} we can rewrite the L-risk (3) as

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

where $\mathbb{L}_s[G] := \int_0^1 s(t) G^{-1}(t)\,\mathrm{d}t$ is called a spectral risk measure with spectrum s applied to CDF G. It stands to reason that $\mathbb{L}_s[F_n]$ converges to $\mathbb{L}_s[F]$ in an appropriate sense. This convergence is governed by the Wasserstein distance between the empirical and population distribution, which we briefly recall here.

Wasserstein Distances For two probability distributions μ and ν on $\mathbb{R}$, the 1-Wasserstein distance $W_1(\mu, \nu)$ between μ and ν is defined by

$$W_1(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}} |x - y|\,\mathrm{d}\gamma(x, y),$$

where $\Pi(\mu, \nu)$ is the set of couplings (or joint distributions) with marginals being μ and ν. It is a metric on the space of probability distributions on $\mathbb{R}$. If F and G are the CDFs associated with μ and ν, respectively, it is known (e.g., Bobkov and Ledoux, 2019, Thm. 2.10) that $W_1(\mu, \nu)$ quantifies the disagreement in either the CDF or quantile functions, i.e.,

$$\begin{aligned}W_1(\mu, \nu) &= \int_0^1 \left| F^{-1}(t) - G^{-1}(t) \right| \mathrm{d}t \\ &= \int_{-\infty}^{+\infty} |F(z) - G(z)|\,\mathrm{d}z. \end{aligned} \tag{4}$$

In contrast, other statistical divergences such as the Cramer von Mises criterion $\int_{-\infty}^{\infty} |F_n(z) - F(z)|\,\mathrm{d}F(z)$ and the Kolmogorov-Smirnoff statistic $\sup_{z \in \mathbb{R}} |F_n(z) - F(z)|$ only measure the disagreement in CDFs, as illustrated in Fig. 1 (top right). The relation (4) is used to prove the upcoming Prop. 3, along with the consistency result below.

Proposition 1. *Assume that* $\mathbb{E}|Z|^p < \infty$ *for some* $p > 2$ *and that* $\|s\|_\infty := \sup_{t \in (0,1)} |s(t)| < \infty$. *Then,*

$$\mathbb{E}\left|\mathbb{L}_s[F_n] - \mathbb{L}_s[F]\right|^2 \leq \frac{2\|s\|_\infty^2 \left(\frac{p}{p-2}\right)^2 \mathbb{E}\left[|Z|^p\right]^{\frac{2}{p}}}{n}.$$

Proof Sketch. By boundedness of s and (4),

$$\begin{aligned}&\mathbb{E}\left|\mathbb{L}_s[F_n] - \mathbb{L}_s[F]\right|^2 \\ &= \mathbb{E}\left| \int_0^1 s(t) \cdot \left(F_n^{-1}(t) - F^{-1}(t)\right) \mathrm{d}t \right|^2 \\ &\leq \|s\|_\infty^2 \cdot \mathbb{E}\left(\int_{-\infty}^{+\infty} |F_n(z) - F(z)|\,\mathrm{d}z \right)^2.\end{aligned}$$

Apply the triangle inequality on $L^2(\mathbb{P})$ to obtain

$$\begin{aligned}&\sqrt{\mathbb{E}\left(\int_{-\infty}^{+\infty}|F_n(z)-F(z)|\,\mathrm{d}z\right)^2}\\&\leq\int_{-\infty}^{+\infty}\sqrt{\mathbb{E}\,|F_n(z)-F(z)|^2}\,\mathrm{d}z\\&=n^{-1/2}\int_{-\infty}^{+\infty}\sqrt{F(z)(1-F(z))}\,\mathrm{d}z,\end{aligned}$$

where the last step uses that for any $z\in\mathbb{R}$, we have $nF_n(z)\sim\mathrm{Binom}\,(n,F(z))$ and compute its variance. The remainder of the proof uses elementary concentration inequalities to bound $\int_{-\infty}^{+\infty}\sqrt{F(z)(1-F(z))}\,\mathrm{d}z$ (see Appx. A). □

Prop. 1 operates in general conditions that are of particular importance in optimization. To put this in context, a number of works provide non-asymptotic uniform learning bounds on spectral (and related) risks (Maurer et al., 2021; Khim et al., 2020; Lee et al., 2020). However, these approaches require boundedness of the random variable of interest, which eliminates any potential application to heavy-tailed losses. Asymptotic approaches proceed by assuming Lipschitz continuity of the spectrum s (Shao, 1989), the trimming of s (i.e. $s(t)=0$ for all $t\in[0,\alpha)\cup(1-\alpha,1]$ with $0<\alpha\leq 1$) (Shorack, 2017; Shao, 1989), or bounded derivatives of the population quantile function F^{-1} (Xiang, 1995). The q-superquantile does not even have a continuous spectrum, whereas the spectrum of the r-extremile is not Lipschitz for $1\leq r<2$. Because s must be non-decreasing to achieve convexity (as we discuss in the upcoming Prop. 2), trimming the upper tail of s is not reflective of practice. Finally, because losses such as the square loss or logistic loss can grow to infinity, the derivative $F^{-1}(t)$ as $t\to\infty$ cannot be assumed to be bounded. Prop. 1 only requires that the population losses satisfy a moment condition and holds without trimming or assumptions of boundedness or Lipschitz continuity on the spectrum. Other recent works employ concentration of the empirical measure in Wasserstein distance to give concentration inequalities for spectral risk measures under sub-Gaussian conditions and moment conditions similar to ours (Prashanth and Bhat, 2022; Bhat and Prashanth, 2019; Pandey et al., 2019).

3 OPTIMIZATION ALGORITHMS

We now consider the optimization of the regularized empirical L-risk objective, for $\mu>0$,

$$\mathcal{R}_\sigma(w)+\frac{\mu}{2}\|w\|_2^2\quad\text{for }\mathcal{R}_\sigma(w)=\sum_{i=1}^n\sigma_i\ell_{(i)}(w).\qquad(5)$$

with $0\leq\sigma_1\leq\ldots\sigma_n\leq 1$, $\sum_{i=1}^n\sigma_i=1$ and ℓ_i convex.

Risk	$s(t)$	$\mathbb{L}_s[F]$
Uniform	1	$\mathbb{E}[Z]$
q-Superquantile	$\frac{1_{[q,1]}(t)}{1-q}$	$\mathbb{E}[Z\|Z\geq F^{-1}(q)]$
r-Extremile	rt^{r-1}	$\mathbb{E}[\max_{k=1,\ldots r}Z_k]$
ρ-ESRM	$\frac{\rho e^{-\rho}e^{\rho t}}{1-e^{-\rho}}$	N/A

Table 1: Common spectral risk measures, with spectra $s(t)$, interpretation of the L-statistics $\mathbb{L}_s[F]$ for F the CDF of Z.

Convexity and Subdifferential As in ERM, the function $\mathcal{R}_\sigma$ is convex as long as each ℓ_i is convex, as we see next. Let ∂f denote the subdifferential of a convex function f and $aS_1+bS_2=\{as_1+bs_2:s_1\in S_1,\,s_2\in S_2\}$ denote the Minkowski sum of sets S_1,S_2 with weights $a,b\in\mathbb{R}$.

Proposition 2. *If $\ell_1,\ldots,\ell_n$ are convex, the function $\mathcal{R}_\sigma$ is also convex, with subdifferential*

$$\partial\mathcal{R}_\sigma(w)=\mathrm{conv}\left(\bigcup_{\pi\in\mathrm{argsort}(\ell(w))}\sum_{i=1}^n\sigma_i\partial\ell_{\pi(i)}(w)\right),$$

where $\mathrm{argsort}\,(\ell(w))=\left\{\pi:\ell_{\pi(1)}(w)\leq\ldots\leq\ell_{\pi(n)}(w)\right\}$. *Moreover, if each ℓ_i is G-Lipschitz continuous, $\mathcal{R}_\sigma$ is also G-Lipschitz continuous.*

Convexity crucially relies on σ_i's being non-decreasing. If each ℓ_i is differentiable, the function $\mathcal{R}_\sigma$ is differentiable almost everywhere, as $\mathrm{argsort}\,(\ell(w))$ is a singleton at almost all $w\in\mathbb{R}^d$. The objective can be non-differentiable at vectors $w\in\mathbb{R}^d$ leading up to ties in the losses such as $\ell_i(w)=\ell_j(w)$ for $i\neq j$.

Computing Subgradients Prop. 2 also gives us a simple recipe to retrieve some $g\in\partial\mathcal{R}_\sigma(w)$ with a differentiable programming framework like JAX (Frostig et al., 2018) or PyTorch (Paszke et al., 2019): (i) compute the losses $\ell_i(w)$, (ii) sort the losses to get $\ell_{\pi(1)}(w),\ldots,\ell_{\pi(n)}(w)$, (iii) compute the weighted sum of the sorted losses $\sum_i\sigma_i\ell_{\pi(i)}(w)$, and (iv) access $g=\sum_i\sigma_i\nabla\ell_{\pi(i)}(w)$ at the sorting given by π using automatic differentiation. We can write this in PyTorch as:

```
l = compute_losses(w)
l_ord = torch.sort(l)[0]
risk = torch.dot(sigmas, l_ord)
g = torch.autograd.grad(risk, w)[0]
```

The dependence of the sorting permutation π on w is not recorded in the computation graph. Multiple options for π occur with probability zero if the losses are continuous random variables, though if they do, we select one arbitrarily.

Stochastic Subgradient Method (SGD) A baseline approach is the stochastic subgradient method displayed in

Algorithm 1 Stochastic Subgradient Method (SGD)

Require: Number of iterates T, minibatch size m, learning rate sequence $(\eta^{(t)})_{t=1}^T$, spectrum s, oracles $(\ell_i)_{i=1}^n$ and $(\nabla \ell_i)_{i=1}^n$, regularization $\mu > 0$.

1: Initialize $w^{(0)} = 0 \in \mathbb{R}^d$.
2: Compute $\hat{\sigma}_1, ..., \hat{\sigma}_m$, where $\hat{\sigma_j} := \int_{(j-1)/m}^{j/m} s(t)\,\mathrm{d}t$.
3: **for** $t = 0, ..., T-1$ **do**
4: Sample without replacement $(i_1, ..., i_m) \subseteq [n]$.
5: Select $\pi \in \operatorname{argsort}\left(\ell_{i_1}(w^{(t)}), ..., \ell_{i_m}(w^{(t)})\right)$.
6: Set $v_m^{(t)} = \sum_{j=1}^m \hat{\sigma}_j \nabla \ell_{i_{\pi(j)}}\left(w^{(t)}\right)$.
7: Set $w^{(t+1)} = (1 - \eta^{(t)}\mu) w^{(t)} - \eta^{(t)} v_m^{(t)}$.
8: **return** $\bar{w}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} w^{(t)}$.

Algorithm 1; we refer to this as (minibatch) SGD for convenience. Given a minibatch size m, the method discretizes the spectrum s into m bins (line 2) instead of n (as in objective (5)). We then sample m indices $\{i_1, ..., i_m\}$ randomly sampled from $\{1, ..., n\}$ (line 4). We retrieve a sorting permutation $\pi : [m] \to [m]$ satisfying $\ell_{i_{\pi(1)}} \leq \ldots \leq \ell_{i_{\pi(m)}}$ (line 5) and use it to compute the update direction $v_m^{(t)}$ (line 6). While the per-iteration cost is m gradient evaluations and $O(md)$ time complexity, Algorithm 1 can fail to minimize the true objective $\mathcal{R}_\sigma$ for non-uniform s due to the bias of the minibatch estimate. For instance, at the extreme $m = 1$, notice that $\hat{\sigma}_1 = \hat{\sigma}_m = 1$ and the subgradient estimate corresponds to $\nabla \ell_i(w)$ for some i. This is an unbiased gradient estimate of the ERM objective rather than $\mathcal{R}_\sigma$, reducing the algorithm to standard SGD. For non-uniform s, the bias can only be fully avoided at $m = n$, recovering the full batch subgradient method.

SGD Analysis Let $u(t) := 1_{(0,1)}(t)$ be the uniform density on $(0, 1)$, which is also the spectrum of the expected value. We have the following convergence guarantee.

Proposition 3. *If the losses $\ell_1, ..., \ell_n$ are G-Lipschitz continuous, differentiable, and convex, the output $\bar{w}^{(T)}$ of Alg. 1 with $\eta^{(t)} = \frac{1}{\mu(t+1)}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}_{\sigma,\mu}\left(\bar{w}^{(T)}\right)\right] - \mathcal{R}_{\sigma,\mu}(w^*) \leq \underbrace{2\sqrt{2} C_s B_\mu \sqrt{\frac{n-m}{m(n-1)}}}_{\text{bias term}} + \underbrace{\frac{2G^2(1+\log T)}{\mu T}}_{\text{optimization term}}.$$

for $w^* = \arg\min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu}(w)$, $C_s = \sup_{t \in (0,1)} |s(t) - u(t)|$, and $B_\mu = \sup_{w : \|w\|_2 \leq G/\mu} \max_{i=1,...,n} |\ell_i(w)| < \infty$. The expectation is taken over the sampling of each minibatch.

In Prop. 3, notice that the bias term can be reduced by decreasing C_s (by pushing s closer to uniformity, hence ERM), decreasing B_μ (by increasing the regularization parameter μ), or decreasing $(n - m)/(mn)$ (by increasing the minibatch size). The optimization term is standard for SGD on convex, Lipschitz objectives with strongly convex regularizers.

Algorithm 2 LSVRG

Require: Number of iterations T, loss functions $(\ell_i)_{i=1}^n$ and their gradient oracles, initial point $w^{(0)}$, learning rate η, sorting update frequency N, spectrum $(\sigma_i)_{i=1}^n$, probability of checkpointing q^*, regularization μ

for $t = 0, ..., T-1$ **do**
  if $t \mod N = 0$ **then** ▷ Update weights
    Select $\pi \in \operatorname{argsort}\left(\ell_1(w^{(t)}), \ldots, \ell_n(w^{(t)})\right)$.
    Update $\lambda^{(t)} = (\sigma_{\pi^{-1}(i)})_{i=1}^n$.
  else
    $\lambda^{(t)} = \lambda^{(t-1)}$.
  Sample $q_t \sim \mathrm{Unif}([0, 1])$.
  if $t \mod N = 0$ or $q_t \leq q^*$ **then** ▷ Checkpoint
    Set $\bar{w}^{(t)} = w^{(t)}$.
    $\bar{g}^{(t)} = \sum_{i=1}^n \lambda_i^{(t)} \nabla \ell_i(\bar{w}^{(t)})$.
  else
    $\bar{w}^{(t)} = \bar{w}^{(t-1)}$ and $\bar{g}^{(t)} = \bar{g}^{(t-1)}$.
  Sample $i_t \sim \mathrm{Unif}([n])$.
  $v^{(t)} = n\lambda_{i_t}^{(t)} \nabla \ell_{i_t}(w^{(t)}) - n\lambda_{i_t}^{(t)} \nabla \ell_{i_t}(\bar{w}^{(t)}) + \bar{g}^{(t)}$.
  $w^{(t+1)} = (1 - \eta\mu) w^{(t)} - \eta v^{(t)}$.
return $w^{(T)}$

Proof Sketch. Given a minibatch $i_1, \ldots, i_m$, let $\ell_{i_{(1)}}(w) \leq \ldots \leq \ell_{i_{(m)}}(w)$ be the order statistics of the losses. Define $\mathcal{R}_{\hat{\sigma}}(w) := \sum_{j=1}^m \hat{\sigma}_j \ell_{i_{(j)}}(w)$. Consider the surrogate objective $\bar{\mathcal{R}}_{\hat{\sigma}}(w) := \mathbb{E}\left[\mathcal{R}_{\hat{\sigma}}(w) \mid w\right]$, where the expectation is taken over the randomness in the minibatch indices $i_1, \ldots, i_m$. We observe that the update directions $v_m^{(t)}$ of Algorithm 1 are *unbiased* estimates for a subgradient in $\partial \bar{\mathcal{R}}_{\hat{\sigma}}(w^{(t)})$. For $\bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) = \bar{\mathcal{R}}_{\hat{\sigma}}(w) + \frac{\mu}{2}\|w\|_2^2$, after enough iterations, we have

$$\bar{\mathcal{R}}_{\hat{\sigma},\mu}(\bar{w}^{(T)}) \approx \min_{w \in \mathbb{R}^d} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w)$$

with error quantified by the optimization term. Letting $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_2 \leq G/\mu\}$, we also show that

$$\bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) - \min_{w' \in \mathcal{W}} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w') \approx \mathcal{R}_{\sigma,\mu}(w) - \min_{w' \in \mathcal{W}} \mathcal{R}_{\sigma,\mu}(w')$$

for any $w \in \mathcal{W}$, quantified by the bias term. After showing that the minimizers of $\mathcal{R}_{\sigma,\mu}$ and $\bar{\mathcal{R}}_{\hat{\sigma},\mu}$ over $\mathbb{R}^d$ as well as $w^{(T)}$ are contained in $\mathcal{W}$, we sum the two errors to give the final result. □

LSVRG Algorithm To circumvent the per-iteration cost of full batch algorithms, we consider adapting the SVRG method (Johnson and Zhang, 2013) for ERM to account for the ordering of the losses, leading to the LSVRG algorithm presented in Alg. 2. Overall, the algorithm consists of considering the objective $\sum_{i=1}^n \sigma_i \ell_{(i)}(w) + \mu\|w\|_2^2/2$ as a

weighted average $\frac{1}{n}\sum_{i=1}^{n}(n\sigma_{\pi^{-1}(i)}\ell_i(w)+\mu\|w\|_2^2/2)$ for $\pi \in \operatorname{argsort}(\ell(w))$ and to run epochs of a q-SVRG (Hofmann et al., 2015) algorithm on an objective of the form $\frac{1}{n}\sum_{i=1}^{n}(n\sigma_{\hat{\pi}^{-1}(i)}\ell_i(w)+\mu\|w\|_2^2/2)$ for $\hat{\pi}$ the ordering of losses computed at some regular checkpoints.

Concretely, with frequency N starting with the first iterate, we compute (i) the n losses at the current iterate to define a vector of weights $\lambda^{(t)}$ associated to the empirical ordered statistics at that point in lines 3 and 4, and (ii) store the current iterate as a checkpoint $\bar{w}^{(t)}$ together with the average gradients of the losses $\bar{g}^{(t)}$ at that checkpoint in lines 9 and 10. In addition, with probability q^* at each iteration we update the checkpoint $\bar{w}^{(t)}$ and the associated average gradients $\bar{g}^{(t)}$ as per rule of line 8, without updating the weights $\lambda^{(t)}$. The main iteration of the algorithm in lines 14 and 15 is a variance-reduced gradient step akin to SVRG on an objective of the form $\frac{1}{n}\sum_{i=1}^{n}(n\bar{\lambda}_i\ell_i(w)+\mu\|w\|_2^2/2)$ where $\bar{\lambda}=\lambda^{(t)}$ are the current weights.

LSVRG Analysis To account for the non-differentiability of the sorting operation in the convergence analysis, we analyze a variant of the LSVRG algorithm that operates on the smooth approximation $h_{\nu\Omega}$ of the empirical L-risk $h(l)=\sum_{i=1}^{n}\sigma_i l_{(i)}$ for $l\in\mathbb{R}^n$, defined using a strongly convex function Ω as (Nesterov, 2005; Beck and Teboulle, 2012)

$$h_{\nu\Omega}(l) := \max_{\lambda\in\mathcal{P}(\sigma)} \left\{l^\top\lambda - \nu\Omega(\lambda)\right\},$$

where $\mathcal{P}(\sigma)=\{\lambda=\Pi\sigma : \Pi\mathbf{1}=\mathbf{1}, \Pi^\top\mathbf{1}=\mathbf{1}, \Pi\in[0,1]^{n\times n}\}$ is the permutahedron generated by σ. The original L-risk is obtained as $\nu\to 0$ since $h(l)=\max_{\lambda\in\mathcal{P}(\sigma)} l^\top\lambda$; this follows from the σ_i's being non-decreasing. The implementation of the smooth approximation of the empirical L-statistic and its gradient is given by solving an isotonic regression problem at a cost of $O(n\log n)$ elementary computations; see Appx. D.

The resulting smooth surrogate of (5) is

$$\mathcal{R}_{\sigma,\mu,\nu\Omega}(w) = h_{\nu\Omega}\big(\ell(w)\big) + \frac{\mu}{2}\|w\|^2, \tag{6}$$

for $\ell(w)=(\ell_1(w),\dots,\ell_n(w))$. The smoothed version of LSVRG we analyze computes the weights in line 4 as $\lambda^{(t)}=\nabla h_{\nu\Omega}(\ell(w^{(t-1)}))$. Note that this update recovers the original one in Algorithm 2 as $\nu\to 0$ when the losses $\ell_i(w^{(t)})$ are unique. Under appropriate smoothness assumptions and choice of the smoothing parameter ν, this variant of LSVRG converges linearly to the minimizer of the smoothed objective.

Theorem 4. *Consider the smooth objective (6) where each ℓ_i is convex, G-Lipschitz continuous and L-smooth, and $\Omega(\lambda)=\|\lambda-\mathbf{1}/n\|_2^2/2$. Consider the sequence $(w^{(t)})$ generated by the smoothed variant of LSVRG with inputs $\nu\geq 4nG^2/\mu$, $N=4(n+8\kappa)$, $\eta=2/((n+8\kappa)\mu)$ where $\kappa=n\sigma_n L/\mu+1$ is a condition number. We have that $w^{(t)}$ converges to $w^*=\arg\min_{w\in\mathbb{R}^d}\mathcal{R}_{\sigma,\mu,\nu\Omega}(w)$ as*

$$\mathbb{E}\|w^{(kN)}-w^*\| \leq (1/2)^k\,\|w^{(0)}-w^*\|$$

for $k\in\mathbb{N}$. Consequently, LSVRG can produce a point $\hat{w}$ satisfying $\left(\mathbb{E}\,\|\hat{w}-w^*\|\right)^2\leq\epsilon$ in

$$T \leq C(n+\kappa)\log\left(\|w^{(0)}-w^*\|_2^2/\epsilon\right)$$

gradient evaluations, where C is an absolute constant.

Proof Sketch. Consider

$$\Phi(w,\lambda) := \sum_{i=1}^{n}\lambda_i\ell_i(w) + \frac{\mu}{2}\|w\|^2 - \nu\Omega(\lambda),$$

so that $\mathcal{R}_{\sigma,\mu,\nu\Omega}(w)=\max_{\lambda\in\mathcal{P}(\sigma)}\Phi(w,\lambda)$. We interpret Algorithm 2 as trying to find the unique saddle point (w^*,λ^*) of Φ by alternating the updates $\lambda^{(k)}=\arg\max_{\lambda\in\mathcal{P}(\sigma)}\Phi(w^{(k)},\lambda)$ and $w^{(k+1)}\approx w_*^{(k+1)}:=\arg\min_w\Phi(w,\lambda^{(k)})$ using N steps of q-SVRG. An error analysis of the latter yields

$$\mathbb{E}_k\|w^{(k+1)}-w_*^{(k+1)}\| \leq \frac{1}{5}\|w^{(k)}-w_*^{(k+1)}\|,$$

where $\mathbb{E}_k$ denotes an expectation conditioned on the sigma-algebra generated by $w^{(k)}$. Smoothness and strong convexity/concavity of Φ gives

$$\|w_*^{(k+1)}-w^*\| \leq \frac{\sqrt{n}G}{\mu}\|\lambda^{(k)}-\lambda^*\| \leq \frac{nG^2}{\mu\nu}\|w^{(k)}-w^*\|.$$

Putting these together with the triangle inequality and $nG^2/(\mu\nu)\leq 1/4$ completes the proof. □

The approximation error induced by the smooth approximation can be controlled by the smoothing coefficient ν. For any non-negative, strongly convex, decomposable ($\Omega(\lambda)=\sum_{i=1}^{n}\omega(\lambda_i)$) function Ω we have $0\leq\mathcal{R}_{\sigma,\mu}(w)-\mathcal{R}_{\sigma,\mu,\nu\Omega}(w)\leq\nu\Omega(\sigma)$. The quantity $\Omega(\sigma)$ can then itself be bounded in terms of a divergence of s to the uniform distribution. In particular, using a centered negative entropy as Ω, we have $\Omega(\sigma)\leq\mathrm{KL}(s\|u)$, Kullback-Leibler divergence from s to u. On the other hand, using a centered squared Euclidean norm as in Thm. 4, we get $\Omega(\sigma)\leq\chi^2(s\|u)/n$, the χ^2-divergence. See Appx. D for details. In summary, if a point $\hat{w}$ is an $\varepsilon/2$-accurate minimizer of the smoothed objective, i.e., $\mathcal{R}_{\sigma,\mu,\nu\Omega}(\hat{w})-\min_{w\in\mathbb{R}^d}\mathcal{R}_{\sigma,\mu,\nu\Omega}(w)\leq\varepsilon$, then it is a $\varepsilon/2+\nu\chi^2(s\|u)/n$-approximate one on the original non-smooth objective when choosing Ω as in Thm. 4. This smoothing error vanishes when considering $s=u$, as in ERM.

Combining the smoothing error with the requirement $\nu>O(nG^2/\mu)$ of Thm. 4, we get an end-to-end bound on the original non-smooth objective when $\varepsilon>G^2\chi^2(s\|u)/\mu$.

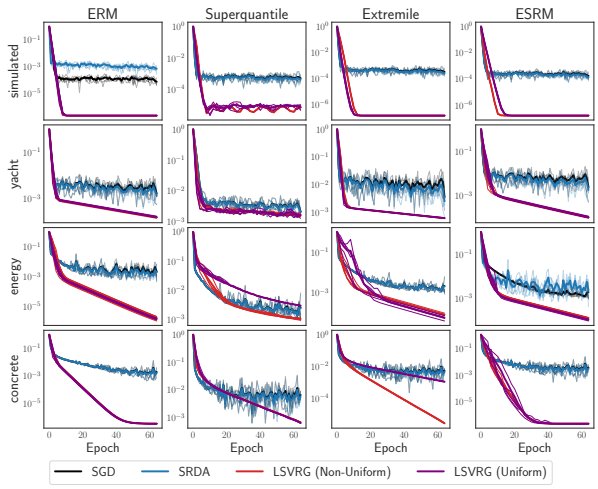

Figure 2: The suboptimality gap (Equation (7)) for various optimization algorithms on the mean, superquantile, extremile, and ESRM risk measures. The x-axis shows the number of effective passes through the data. Five seed trajectories are plotted translucently for every algorithm.

However, as we show empirically in Appx. G, smoothing has a minimal impact on the empirical behavior of LSVRG. While the non-smoothness of the empirical L-risk is an obstacle for the theoretical convergence of LSVRG, this non-smoothness may not impact the empirical behavior. Indeed, if the minimizer of the objective has distinct loss values, then the objective is locally smooth around the minimizer.

Time Complexity In practice, we consider simply taking $N = n$ and $q^* = 0$ to simplify the hyperparameter choices and reduce the overall time complexity. In that case, the time complexity of LSVRG is $O(d)$ per iteration with 2 gradient evaluations, which is identical to the number of gradient calls of the biased subgradient method with batch size $m = 2$. LSVRG also requires n gradient evaluations and sorting at the start of an epoch, contributing an additional $O(nd + n \log n)$ elementary operations. This per-epoch complexity is nearly identical to vanilla SVRG in the ERM case. LSVRG, like vanilla SVRG, also requires an additional storage of $O(d)$ to store $\bar{g}^{(t)} \in \partial \mathcal{R}_{\sigma,\mu}(\bar{w}^{(t)})$ as compared to the stochastic subgradient method. Run times are evaluated experimentally in Appx. G.

4 EXPERIMENTAL RESULTS

We compare the performance of minibatch SGD and LSVRG on benchmark datasets and study their bias and variance properties in a number of supervised and unsupervised learning tasks. Experimental details can be found in Appx. F, with additional experiments with varied hyperparameters can be found in Appx. G.

4.1 Regression

We consider 4 regression datasets:

- `simulated`: a synthetic task of predicting observations generated from a noisy linear model.
- `yacht`: prediction of the residuary resistance of a sailing yacht based on its physical attributes (Tsanas and Xifara, 2012).
- `energy`: prediction of the cooling load of a building based on its physical attributes (Baressi Segota et al., 2020).
- `concrete`: prediction of the compressive strength of a concrete type based on its physical and chemical attributes (Yeh, 2006).

We use the squared loss under a linear model and aim to minimize the regularized objective (5) where the spectra s are obtained from the empirical mean, superquantile ($q = 0.5$), extremile ($r = 2$), and ESRM ($\rho = 1$) of the losses. Both training curves and test losses for other values of (q, r, ρ) are shown in Appx. G, which follow similar trends.

In addition to minibatch SGD, we consider another biased method, *stochastic regularized dual averaging (SRDA)* (Xiao, 2009), both with a batch size of 64. We compare them with LSVRG, by plotting in Fig. 2 the suboptimality, defined as

$$\text{suboptimality gap}_t := \frac{\mathcal{R}_\sigma(w^{(t)}) - \mathcal{R}_\sigma(w^*)}{\mathcal{R}_\sigma(w^{(0)}) - \mathcal{R}_\sigma(w^*)} \tag{7}$$

We find that LSVRG (without smoothing) exhibits empirical linear convergence for the ERM, extremile, and ESRM. It often vastly outperforms SGD and SRDA, which exhibit sublinear convergence. On the superquantile, LSVRG exhibits the same sublinear convergence as SGD, suggesting that the discontinuity of the spectrum can yield additional challenges in optimization. Overall, LSVRG is the best or close to the best algorithm across all tasks.

LSVRG relies on the hypothesis that the sorted order of losses stabilize as iterates $w^{(t)}$ get close to the optimum. We see from Fig. 3 that there is a clear phase change after which disagreements between the true and estimated ordering are visually unnoticeable. The exception to this is the superquantile, where the sorting does not stabilize within 64 epochs. This corroborates the apparent hardness of optimizing the superquantile in Fig. 2.

4.2 Classification

Image Classification The `iWildCam` challenge dataset (Beery et al., 2020) contains natural images from wilderness sites with distribution shifts arising from diverse camera angles, backgrounds, and relative animal frequencies. We take a subsample of $n = 20,000$ data points from classes with at least 100 examples after removing the "background

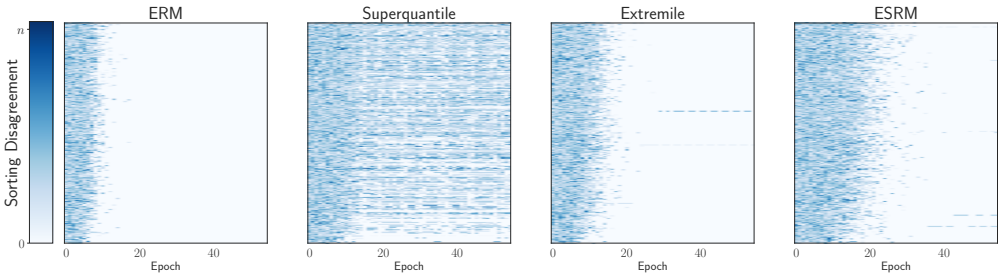

Figure 3: Sorting sensitivity for `simulated` dataset ($n = 800$) along epochs (x-axis) of LSVRG applied each spectral risk objective. Each heatmap shows the vector of disagreements between sorting permutations π at each epoch of Algorithm 2.

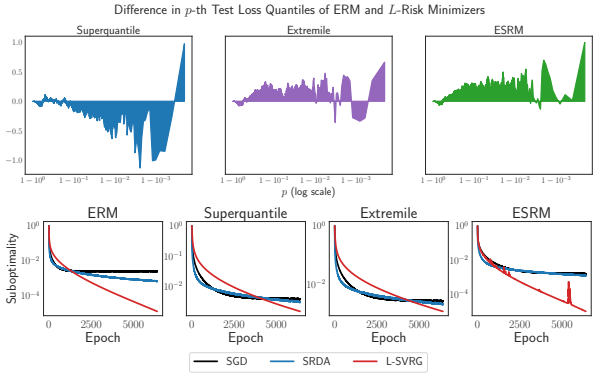

Figure 4: **Top:** Differences in the p-th quantile of the loss distribution for the ERM solution and the L-Risk solution on the `iWildCam` test set. **Bottom:** Training curves for SGD, SRDA, and L-SVRG for different L-Risks on `iWildCam` training set.

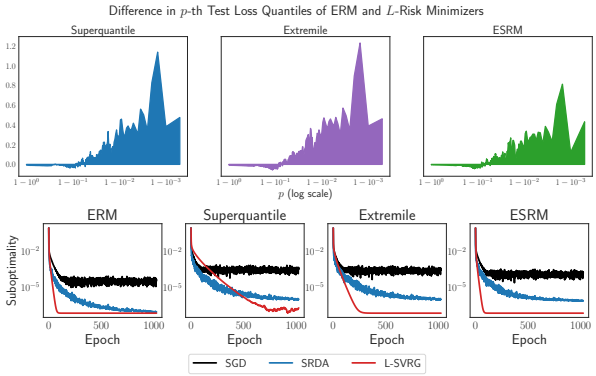

Figure 5: **Top:** Differences in the p-th quantile of the loss distribution for the ERM solution and the L-Risk solution on the `emotion` test set. **Bottom:** Training curves for SGD, SRDA, and L-SVRG for different L-Risks on `emotion` training set.

image" class. For each data point, we compute the penultimate layer of a ResNet50 neural network that is pre-trained on ImageNet (see Appx. F for further details). The resulting vectors are reduced in dimension by PCA, after which the convex optimization problem considered is multinomial logistic regression with the reduced vectors as inputs. As in regression, the training curves in the bottom row of Fig. 4 indicate that SGD and SRDA fail to converge due to bias and variance. Letting $\hat{w}_{\text{ERM}}$ be the approximate solution of ERM, whereas $\hat{w}_{\text{LRM}}$ is the approximate solution of an L-Risk minimization problem other than ERM, the top row plots the following against p:

$$\frac{\ell_{(\lceil np \rceil)}\left(\hat{w}_{\text{ERM}}\right) - \ell_{(\lceil np \rceil)}\left(\hat{w}_{\text{LRM}}\right)}{\frac{1}{n}\sum_{i=1}^{n} \ell_i\left(\hat{w}_{\text{ERM}}\right)},$$

that is, the difference in the p-th quantile of the test loss of $\hat{w}_{\text{ERM}}$ and the p-th quantile of the test loss of $\hat{w}_{\text{LRM}}$, normalized by the mean test loss of ERM. Because logistic loss measures the negative logarithm of the probability that the model assigns to the correct label, tail events for this loss amount to a model exhibiting high confidence for a set of incorrect labels. The median test loss ($p = 0.5$) is similar between the L-risk minimizers and standard ERM. However, for $p > 0.5$, the ERM solution can make predictions with much higher losses. Comparing various L-risks, we find that the superquantile controls tail error at very high quantiles ($p > 0.95$), but generally underperforms for the rest of the loss distribution. The extremile and ESRM on the other hand, have generally better performance than ERM throughout the loss distribution. We also plot the quantile differences for the regression tasks in Appx. G.

Crucially, we find that the large n regime exacerbates the bias issues when the epoch length is set to n. We instead use a smaller epoch length of 100, and plot suboptimality against the number of gradient evaluations in Fig. 4 to ensure a fair comparison. Each epoch is defined as the number of gradient evaluations in SGD or SRDA, which is $100m = 6,400$.

Text Classification The `emotion` dataset (Saravia et al., 2018) contains English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise, which comprise six classes for classification. In this example, we split the $n = 16,000$ training examples into two random

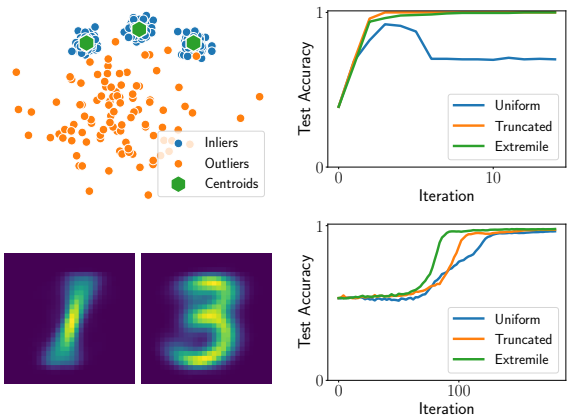

Figure 6: Robust clustering with L-statistics in the presence of outliers. **Top**: synthetic data, **Bottom**: MNIST digits.

subsets of size $8,000$. The first is used to fine-tune each layer of a BERT neural network using the ERM objective on multinomial logistic loss (a non-convex optimization problem) for 2 epochs. The backbone of this network is then used as a fixed embedding function applied to the remaining subset of $8,000$ points. After dimension reduction by PCA, a linear model is trained using the L-Risk objective to map to the final predictions. The reason for fine-tuning is that we found that using the embeddings from the pre-trained BERT model result in instability in training if the classes are not well-separated by the embeddings. After improving the embeddings, a large epoch length of $1,000$ could be used for LSVRG. The results are plotted in Fig. 5. Similar to the image classification setting, we find that on the test examples, the ERM minimizer incurs large losses in the upper tail. Here, the superquantile, extremile, and ESRM all exhibit similar, stable behavior. Both examples demonstrate two ways to adapt LSVRG, designed specifically for convex objectives, to the non-convex regime.

4.3 Clustering

We also explore an unsupervised clustering approach from Maurer et al. (2021) on synthetic data and real data. We seek to cluster n points $x_1, \dots, x_n$ into k clusters with centers $C = (c_1, \dots, c_k)$ by minimizing a weighted average of the distances of each point to its closest center, i.e., problems of the form

$$\min_{C \in \mathbb{R}^{d \times k}} \sum_{i=1}^{n} \sigma_i \ell_{(i)}(C), \text{ for}$$

$$\ell_i(C) = \min_{\substack{z_i \in \{0,1\}^k \\ z_i^\top \mathbf{1} = \mathbf{1}}} \sum_{j=1}^{k} z_{ij} \|x_i - c_j\|_2^2.$$

Taking $\sigma_i = 1/n$, we retrieve the usual objective minimized by k-means. Maurer et al. (2021) propose to take σ_i non-uniform to mitigate the effect of outliers in the data. Specifically, we consider $\sigma_i = \int_{(i-1)/n}^{i/n} s(t)\mathrm{d}t$ for a truncated spectrum $s_q(t) = \mathbf{1}_{[0,q]}(t)/q$ or a risk-seeking version of the extremile, $s_r(t) = r(1-t)^r$. In addition, Maurer et al. (2021) optimize the clustering objective by alternating k-means iterations and sorting the resulting losses. Instead, we apply minibatch SGD from Algorithm 1 with a constant stepsize found by grid search and a batch size of 64.

Synthetic data. We generate a dataset of three Gaussian clouds of 100 points each and an additional set of 100 outliers (top left in Fig. 6). We compare the accuracy of clustering 300 new inlier points with different spectra. We observe in Fig. 6 (top right) that the minibatch estimates of the subgradient of the truncated or extremile spectra are sufficient to reach a perfect 100% accuracy, while vanilla k-means with its uniform spectrum leads to poor performance due to outliers.

MNIST data. We consider distinguishing between the digits 1 and 3 from the MNIST dataset (LeCun et al., 1998) by clustering the images of a training set composed of 1000 samples of 1 and 3 each and additional 125 outliers for each other digit. We test the clustering procedure on the images of 1 and 3 digits from the MNIST test set. We see from Fig. 6 (bottom right) that minibatch SGD with a batch size of 256 achieves 97.3% for the truncated spectrum and 97.8% for the extremile spectrum versus 96.3% for the uniform spectrum. Even in terms of convergence speed for different spectra, we observe that extremile $\succ$ truncated $\succ$ uniform. Finally, Fig. 6 (bottom left) shows us that the centers computed with the extremile spectrum are clear representatives of these digits while taking a uniform spectrum leads to more blurry representatives, as shown in Appx. F.

5 CONCLUSION

L-risks, based on spectral risk measures, span an entire spectrum of learning objectives such as the ERM objective, the superquantile-based objective, and other distributionally robust ones. We presented LSVRG, a stochastic optimization algorithm for minimizing L-risks, and analyzed its convergence properties alongside biased minibatch SGD. Establishing the regular subdifferential in the non-convex setting and studying the robustness properties of L-risk minimizers are interesting venues for future work.

Acknowledgements. This work was supported by NSF DMS-2023166, CCF-2019844, DMS-2052239, DMS-2134012, DMS-2133244, NIH, CIFAR-LMB, and faculty research awards. Part of this work was done while Zaid Harchaoui was visiting the Simons Institute for the Theory of Computing, and while Krishna Pillutla and Vincent Roulet were at the University of Washington.

References

C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

P. Artzner, F. Delbaen, E. Jean-Marc, and D. Heath. Coherent Measures of Risk. *Mathematical Finance*, 9:203 – 228, 07 1999.

F. Bach. *Learning Theory from First Principles*. The MIT Press, 2023.

S. Baressi Segota, N. Andelic, J. Kudlacek, and R. Cep. Artificial neural network for predicting values of residuary resistance per unit weight of displacement. *Journal of Maritime & Transportation Science*, 57, 2020.

A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

S. Beery, E. Cole, and A. Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17:449–476, 2007.

M. J. Best, N. Chakravarti, and V. A. Ubhaya. Minimizing Separable Convex Functions Subject to Simple Chain Constraints. *SIAM Journal on Optimization*, 10(3):658–672, 2000.

S. P. Bhat and L. A. Prashanth. Concentration of risk measures: A Wasserstein distance approach. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

M. Blondel, O. Teboul, Q. Berthet, and J. Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959, 2020.

S. G. Bobkov and M. Ledoux. One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances. *Memoirs of the American Mathematical Society*, 2019.

R. Chen and I. C. Paschalidis. Distributionally Robust Learning. *Foundations and Trends® in Optimization*, 4(1-2): 1–243, 2020.

J. Cotter and K. Dowd. Extreme Spectral Risk Measures: An Application to Futures Clearinghouse Margin Requirements. *Journal of Banking & Finance*, 30(12):3469–3485, 2006.

S. Curi, K. Y. Levy, S. Jegelka, and A. Krause. Adaptive Sampling for Stochastic Risk-Averse Learning. In *Neural Information Processing Systems*, volume 33, 2020.

A. Daouia, I. Gijbels, and G. Stupfler. Extremiles: A New Perspective on Asymmetric Least Squares. *Journal of the American Statistical Association*, 114(527):1366–1381, 2019.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *Neural Information Processing Systems*, volume 27, 2014.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

J. C. Duchi and H. Namkoong. Variance-based Regularization with Convex Objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

Y. Fan, S. Lyu, Y. Ying, and B. Hu. Learning with Average Top-k Loss. In *Neural Information Processing Systems*, volume 30, 2017.

H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance Stochastics*, 6(4):429–447, 2002.

R. Frostig, M. J. Johnson, and C. Leary. Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, 4(9), 2018.

V. Guigues and C. A. Sagastizábal. Risk-averse feasible policies for large-scale multistage stochastic linear programs. *Mathematical Programming*, 138(1-2):167–198, 2013.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

X. D. He, S. Kou, and X. Peng. Risk Measures: Robustness, Elicitability, and Backtesting. *Annual Review of Statistics and Its Application*, 9(1), 2022.

A. Henzi, A. Mösching, and L. Dümbgen. Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Methodology and computing in applied probability*, pages 1–13, 2022.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer, 1993.

T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance Reduced Stochastic Gradient Descent with Neighbors. *Neural Information Processing Systems*, 28, 2015.

M. J. Holland and E. Mehdi Haress. Spectral risk-based learning using unbounded losses. In *International Con-*

ference on Artificial Intelligence and Statistics, volume 151, pages 1871–1886, 2022.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037, 2018.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Neural Information Processing Systems*, volume 26, 2013.

K. Kawaguchi and H. Lu. Ordered SGD: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pages 669–679, 2020.

J. Khim, L. Leqi, A. Prasad, and P. Ravikumar. Uniform Convergence of Rank-weighted Learning. In *International Conference on Machine Learning*, volume 119, pages 5254–5263, 2020.

D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

Y. Laguel, J. Malick, and Z. Harchaoui. First-Order Optimization for Superquantile-Based Supervised Learning. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 09 2020.

Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. Superquantiles at Work: Machine Learning Applications and Efficient Subgradient Computation. *Set-Valued and Variational Analysis*, 2021.

N. Le Roux, M. Schmidt, and F. Bach. A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets. In *Neural Information Processing Systems*, volume 25, 2012.

Y. LeCun, C. Cortes, and B. Christopher. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 1998.

J. Lee and M. Raginsky. Minimax Statistical Learning with Wasserstein distances. In *Neural Information Processing Systems*, volume 31, pages 2687–2696, 2018.

J. Lee, S. Park, and J. Shin. Learning Bounds for Risk-sensitive Learning. In *Neural Information Processing Systems*, volume 33, pages 13867–13879, 2020.

L. Leqi, A. Prasad, and P. K. Ravikumar. On Human-Aligned Risk Minimization. In *Neural Information Processing Systems*, volume 32, 2019.

D. Levy, Y. Carmon, J. Duchi, and A. Sidford. Large-Scale Methods for Distributionally Robust Optimization. In *Neural Information Processing Systems*, volume 33, 2020.

T. Li, A. Beirami, M. Sanjabi, and V. Smith. Tilted Empirical Risk Minimization. In *International Conference on Learning Representations*, 2021.

C. H. Lim and S. J. Wright. Efficient Bregman Projections onto the Permutahedron and Related Polytopes. In *International Conference on Artificial Intelligence and Statistics*, pages 1205–1213, 2016.

J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal on Optimization*, 25, 02 2014.

A. Maurer, D. A. Parletta, A. Paudice, and M. Pontil. Robust Unsupervised Learning via L-statistic Minimization. In *International Conference on Machine Learning*, pages 7524–7533, 2021.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

A. K. Pandey, L. A. Prashanth, and S. P. Bhat. Estimation of spectral risk measures. In *AAAI Conference on Artificial Intelligence*, 2019.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Neural Information Processing Systems*, volume 32, pages 8024–8035. Curran Associates Inc., 2019.

G. C. Pflug and A. Ruszczyński. Measuring Risk for Income Streams. *Computational Optimization and Applications*, 32(1):161–178, 2005.

L. A. Prashanth and S. P. Bhat. A Wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23(238):1–61, 2022.

S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for smooth nonconvex optimization. In *IEEE 55th Conference on Decision and Control (CDC)*, pages 1971–1977, 2016.

R. T. Rockafellar and J. O. Royset. Random variables, monotone relations, and convex analysis. *Mathematical Programming*, 148(1–2):297–331, 2014.

R. T. Rockafellar and S. Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18:33–53, 2013.

E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

S. Sarykalin, G. Serraino, and S. Uryasev. Value-at-Risk vs. Conditional Value-at-Risk in Risk Management and Optimization. In *State-of-the-art decision-making tools in the information-intensive age*, pages 270–294. INFORMS, 2008.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

S. Shalev-Shwartz and T. Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

J. Shao. Functional calculus and asymptotic theory for statistical analysis. *Statistics & Probability Letters*, 8(5): 397–405, 1989.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming - Modeling and Theory, Second Edition*, volume 16. SIAM, 2014.

G. Shorack. *Probability for Statisticians*. Springer Texts in Statistics, 2017.

A. Tsanas and A. Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49: 560–567, 2012.

R. Williamson and A. Menon. Fairness Risk Measures. In *International Conference on Machine Learning*, pages 6786–6797, 2019.

X. Xiang. A note on the bias of L-estimators and a bias reduction procedure. *Statistics & Probability Letters*, 23 (2):123–127, 1995.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv Preprint*, 2017.

L. Xiao. Dual Averaging Method for Regularized Stochastic Learning and Online Optimization. In *Neural Information Processing Systems*, volume 22, 2009.

I. Yeh. Analysis of Strength of Concrete Using Design of Experiments and Neural Networks. *Journal of Materials in Civil Engineering*, 18, 2006.

Appendix

In the appendices, we give the proofs of consistency (Prop. 1) in Appx. A and the variational properties of the objective (Prop. 2) in Appx. B. Appx. C contains the analysis of bias SGD (Prop. 3). Appx. E contains the analysis of LSVRG (Thm. 4), with necessary background in Appx. D. We then give describe the experimental setup in detail (Appx. F) and give some additional numerical results (Appx. G).

Table of Contents

A CONSISTENCY OF THE EMPIRICAL SPECTRAL RISK

We first recall the setting of Prop. 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a common probability space, upon which we consider an i.i.d. sample $\{Z_1, \ldots, Z_n\}$ with each $Z_i : \Omega \to \mathbb{R}$ being $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable, where $\mathcal{B}(\mathbb{R})$ denotes the Borel sets on the real line. Each shares a common cumulative distribution function (CDF) F and quantile function F^{-1} given by

$$F(z) := \mathbb{P}\left[Z_1^{-1}\left((-\infty, z]\right)\right] \text{ and } F^{-1}(t) := \inf\left\{z : F(z) \geq t\right\}.$$

Similarly, define the empirical CDF and quantile functions by

$$F_n(z;\omega) := \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,z]}\left(Z_i(\omega)\right) \text{ and } F_n^{-1}(t;\omega) := \inf\left\{z : F_n(z;\omega) \geq t\right\}.$$

Construct the random variables $F_n(z) : \omega \mapsto F_n(z;\omega)$ and $F_n^{-1}(t) : \omega \mapsto F_n^{-1}(t;\omega)$. Here, $z \in \mathbb{R}$ and $t \in (0,1)$, and the infimum is always attained (Bobkov and Ledoux, 2019, Page 83). We can ensure measurability of $F_n^{-1}(t)$ by taking the infimum only over $z \in \mathbb{Q}$. All expected values will be taken with respect to $(\Omega, \mathcal{F}, \mathbb{P})$ and will be denoted by $\mathbb{E}$. For s a probability density function (PDF) on $(0,1)$, the L-functional $\mathbb{L}_s$ with spectrum s is defined as

$$\mathbb{L}_s[F] := \int_0^1 s(t) \cdot F^{-1}(t)\,\mathrm{d}t. \tag{8}$$

We first establish that (8) is well-defined, using a well-known result of quantile functions.

Proposition 5. *(Bobkov and Ledoux, 2019, Proposition A.1) Let Z be a random variable and let F be its cumulative distribution function. If U is a random variable distributed uniformly in $(0,1)$, then the random variable $F^{-1}(U)$ has F as its distribution function. In particular,*

$$\mathbb{E}\,|Z|^p = \int_0^1 \left|F^{-1}(t)\right|^p \, dt$$

when the left hand side is finite.

Lemma 6. *Let s be bounded, and $\mathbb{E}\,|Z_1| < \infty$. Then $|\mathbb{L}_s[F]| < \infty$.*

Proof. Let $\|s\|_\infty := \sup_{t\in(0,1)} |s(t)| < \infty$. Write

$$|\mathbb{L}_s[F]| = \left|\int_0^1 s(t) \cdot F^{-1}(t)\,\mathrm{d}t\right| \leq \|s\|_\infty \cdot \int_0^1 \left|F^{-1}(t)\right| \,\mathrm{d}t \overset{\text{Prop. 5}}{\leq} \|s\|_\infty \,\mathbb{E}\,|Z_1| < \infty.$$

□

We restate Prop. 1 below.

Proposition 1. *Assume that $\mathbb{E}\,|Z|^p < \infty$ for some $p > 2$ and that $\|s\|_\infty := \sup_{t\in(0,1)} |s(t)| < \infty$. Then,*

$$\mathbb{E}\left|\mathbb{L}_s\left[F_n\right] - \mathbb{L}_s\left[F\right]\right|^2 \leq \frac{2\,\|s\|_\infty^2 \left(\frac{p}{p-2}\right)^2 \mathbb{E}\left[|Z|^p\right]^{\frac{2}{p}}}{n}.$$

The proof is summarized by the following steps.

1. By boundedness of the spectrum, we have that $\mathbb{E}\left|\mathbb{L}_s\left[F_n\right] - \mathbb{L}_s\left[F\right]\right|^2 \leq \|s\|_\infty^2 \cdot \mathbb{E}\left[\left(\int_0^1 \left|F_n^{-1}(t) - F^{-1}(t)\right| \,\mathrm{d}t\right)^2\right]$.
2. Using the triangle inequality on $L^2(\mathbb{P})$ and relationships between quantile functions and CDFs, we relate $\sqrt{\mathbb{E}\left[\left(\int_0^1 \left|F_n^{-1}(t) - F^{-1}(t)\right| \,\mathrm{d}t\right)^2\right]}$ to the quantity $\frac{1}{\sqrt{n}}\int_{-\infty}^{+\infty} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z$.
3. We then use elementary concentration inequalities to bound $\int_{-\infty}^{+\infty} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z$ by $\sqrt{2}\frac{p}{p-2}\mathbb{E}\left[|Z|^p\right]^{1/p}$.

The following theorem details how the L^1 distance between the quantile functions of two probability distributions is equal to the L^1 distance between the corresponding CDFs.

Theorem 7 (Theorem 2.10 of Bobkov and Ledoux (2019))**.** *Let μ, ν be two probability distributions on $\mathbb{R}$ with associated CDF F and G, respectively, with quantile functions $F^{-1}(t) := \inf\{z \in \mathbb{R} : F(z) \geq t\}$ and $G^{-1} := \inf\{z \in \mathbb{R} : G(z) \geq t\}$. Given that μ and ν have finite first moment, i.e., $\int |z| \, d\mu(z) < \infty$ and $\int |z| \, d\nu(z) < \infty$, we have that*

$$W_1(\mu, \nu) = \int_{-\infty}^{\infty} |F(z) - G(z)| \, dz = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right| \, dt,$$

where both the left and right hand sides are finite.

Next, we ensure that the L^1 distance between F_n^{-1} and F_n is a square-integrable random variable.

Lemma 8. *Assume that $\mathbb{E}\left|Z_1\right|^2 < \infty$. Then, the random variable $V_n(\omega) := \int_0^1 \left|F_n^{-1}(t;\omega) - F^{-1}(t)\right| \, dt$ is well-defined, and $\mathbb{E}[V_n^2] < \infty$.*

Proof. For any particular realization $\omega \in \Omega$, write

$$\begin{aligned}
|V_n(\omega)|^2 &= \left| \int_0^1 \left|F_n^{-1}(t;\omega) - F^{-1}(t)\right| \, \mathrm{d}t \right|^2 \\
&\leq \int_0^1 \left|F_n^{-1}(t;\omega) - F^{-1}(t)\right|^2 \, \mathrm{d}t && \text{Jensen's inequality} \\
&\leq 2\int_0^1 \left|F_n^{-1}(t;\omega)\right|^2 \, \mathrm{d}t + 2\int_0^1 \left|F^{-1}(t)\right|^2 \, \mathrm{d}t \\
&= \frac{2}{n}\sum_{i=1}^n |Z_i(\omega)|^2 + 2\mathbb{E}\left|Z_1\right|^2 && \text{Prop. 5.}
\end{aligned}$$

Then, $\mathbb{E}[V_n^2] \leq 4\mathbb{E}\left|Z_1\right|^2$, completing the proof. □

The next lemma applies the above theorem to bound the expected distance between empirical and population quantile functions in terms of the population CDF, expanding upon remarks made on page 20 of Bobkov and Ledoux (2019).

Lemma 9. *Assume that $\mathbb{E}\left|Z_1\right|^2 < \infty$. Then,*

$$\sqrt{\mathbb{E}\left[\left(\int_0^1 \left|F_n^{-1}(t) - F^{-1}(t)\right| \, dt\right)^2\right]} \leq \frac{1}{\sqrt{n}} \int_{-\infty}^{+\infty} \sqrt{F(z)(1 - F(z))} \, dz,$$

where the right hand side is permitted to be infinite.

Proof. By Lem. 8 we have that

$$\mathbb{E}\left[\left(\int_0^1 \left|F_n^{-1}(t) - F^{-1}(t)\right| \, \mathrm{d}t\right)^2\right] = \mathbb{E}\left[V_n^2\right] < \infty, \tag{9}$$

so that the left hand side is well-defined and finite. By Thm. 7, we also have that $\int_0^1 \left|F_n^{-1}(t;\omega) - F^{-1}(t)\right| \, \mathrm{d}t = \int_{-\infty}^{\infty} |F_n(z;\omega) - F(z)| \, \mathrm{d}z$, indicating with (9) that the random variable

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

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

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

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

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

completing the proof. □

The final lemma bounds the right hand side of Lem. 9.

Lemma 10. *Consider a random variable Z with c.d.f. F. If Z satisfies $\mathbb{E}\left[|Z|^p\right] < \infty$ for $p > 2$, then*

$$\int_{-\infty}^{+\infty} \sqrt{F(z)(1 - F(z))}\, dz \leq \sqrt{2}\left(\frac{p}{p-2}\right)\mathbb{E}\left[|Z|^p\right]^{\frac{1}{p}}.$$

Proof. By definition, $\int_{-\infty}^{\infty} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z = \lim_{a\to+\infty} \int_{-a}^{a} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z$. Denote $c = \mathbb{E}\left[|Z|^p\right]^{1/p}$. For any constant $a \geq c > 0$, we have

$$\begin{aligned}\int_{-a}^{a} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z &= \int_{-a}^{0} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z + \int_{0}^{a} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z \\ &\leq \int_{-a}^{0} \sqrt{F(z)}\,\mathrm{d}z + \int_{0}^{a} \sqrt{(1-F(z))}\,\mathrm{d}z \\ &= \int_{-a}^{0} \sqrt{\mathbb{P}(Z \leq z)}\,\mathrm{d}z + \int_{0}^{a} \sqrt{\mathbb{P}(Z > z)}\,\mathrm{d}z \\ &= \int_{0}^{a} \sqrt{\mathbb{P}(Z \leq -z)} + \sqrt{\mathbb{P}(Z > z)}\,\mathrm{d}z.\end{aligned}$$

Then, use that for any $a, b \geq 0$,

$$(\sqrt{a} + \sqrt{b})^2 = a + b + 2\sqrt{ab} \leq 2(a+b) \implies \sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}.$$

Using this, and that $z \geq 0$, we have

$$\begin{aligned}\sqrt{\mathbb{P}(Z \leq -z)} + \sqrt{\mathbb{P}(Z > z)} &\leq \sqrt{2(\mathbb{P}(Z \leq -z) + \mathbb{P}(Z > z))} \\ &= \sqrt{2(\mathbb{P}(|Z| > z) + \mathbb{P}(Z = -z))} \\ &\leq \sqrt{2(\mathbb{P}(|Z| > z) + \mathbb{P}(|Z| = z))} \\ &= \sqrt{2\mathbb{P}(|Z| \geq z)}.\end{aligned}$$

Combining with the first display, we have that

$$\begin{aligned}
\int_{-a}^{a} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z &\le \int_0^a \sqrt{\mathbb{P}(Z \le -z)} + \sqrt{\mathbb{P}(Z > z)}\,\mathrm{d}z \\
&\le \sqrt{2}\int_0^a \sqrt{\mathbb{P}(|Z| \ge z)}\,\mathrm{d}z \\
&\le \sqrt{2}\int_0^a \sqrt{\min\left\{1, \frac{c^p}{z^p}\right\}}\,\mathrm{d}z && \text{Markov's inequality} \\
&= \sqrt{2}\left(c + c^{p/2}\int_c^a z^{-p/2}\,\mathrm{d}z\right).
\end{aligned}$$

Computing the integral yields

$$\int_c^a z^{-p/2}\,\mathrm{d}z = \frac{a^{1-p/2} - c^{1-p/2}}{1-p/2}.$$

Because $1 - p/2 < 0$, we have that $\lim_{a\to\infty}\int_c^a z^{-p/2}\,\mathrm{d}z = \frac{c^{1-p/2}}{p/2-1}$. Combining the steps above, we obtain

$$\begin{aligned}
\int_{-\infty}^{\infty} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z &= \lim_{a\to\infty}\int_{-a}^{a} \sqrt{F(z)(1-F(z))}\,\mathrm{d}z \\
&\le \lim_{a\to\infty} \sqrt{2}\left(c + c^{p/2}\int_c^a z^{-p/2}\,\mathrm{d}z\right) \\
&= \sqrt{2}c\left(1 + \frac{1}{p/2-1}\right) \\
&= \sqrt{2}\frac{pc}{p-2}.
\end{aligned}$$

Resubstituting $c = \mathbb{E}\left[|Z|^p\right]^{1/p}$ completes the proof. □

We now have the tools to prove Prop. 1.

Proof of Prop. 1. For a particular realization $Z_1(\omega), ..., Z_n(\omega)$, we have that

$$\begin{aligned}
\left|\mathbb{L}_s\left[F_n(\cdot;\omega)\right] - \mathbb{L}_s\left[F\right]\right| &= \left|\int_0^1 s(t)\cdot F_n^{-1}(t;\omega)\,\mathrm{d}t - \int_0^1 s(t)\cdot F^{-1}(t)\,\mathrm{d}t\right| \\
&= \left|\int_0^1 s(t)\cdot\left(F_n^{-1}(t;\omega) - F^{-1}(t)\right)\,\mathrm{d}t\right| \\
&\le \sup_{t\in(0,1)}|s(t)|\cdot\int_0^1 \left|F_n^{-1}(t;\omega) - F^{-1}(t)\right|\,\mathrm{d}t \\
&= \|s\|_\infty\cdot\int_0^1 \left|F_n^{-1}(t;\omega) - F^{-1}(t)\right|\,\mathrm{d}t.
\end{aligned}$$

We then take the square and expectation.

$$\begin{aligned}
\mathbb{E}\left|\mathbb{L}_s\left[F_n\right] - \mathbb{L}_s\left[F\right]\right|^2 &\le \|s\|_\infty^2\cdot\mathbb{E}\left[\left(\int_0^1 \left|F_n^{-1}(t) - F^{-1}(t)\right|\,\mathrm{d}t\right)^2\right] \\
&\le \frac{\|s\|_\infty^2}{n}\cdot\left(\int_{-\infty}^{+\infty}\sqrt{F(z)(1-F(z))}\,\mathrm{d}z\right)^2 && \text{Lem. 9} \\
&\le \frac{2\|s\|_\infty^2}{n}\left(\frac{p}{p-2}\right)^2\mathbb{E}\left[|Z|^p\right]^{\frac{2}{p}}. && \text{Lem. 10}
\end{aligned}$$

□

B CONVEXITY AND SUBDIFFERENTIAL PROPERTIES

Recall the expression of the empirical L-statistics

$$\mathcal{R}_\sigma(w) := \sum_{i=1}^n \sigma_i \ell_{(i)}(w). \tag{10}$$

where $0 \leq \sigma_1 \leq \cdots \leq \sigma_n$, $\sum_{i=1}^n \sigma_i = 1$, each $\ell_i : \mathbb{R}^d \to \mathbb{R}$ is a function representing performance of model weights w on training instance i, and for a vector $l \in \mathbb{R}^n$, we denote $l_{(1)} \leq \ldots \leq l_{(n)}$ its ordered coefficients. We recall Prop. 2 and present its proof.

Proposition 2. *If $\ell_1, \ldots, \ell_n$ are convex, the function $\mathcal{R}_\sigma$ is also convex, with subdifferential*

$$\partial \mathcal{R}_\sigma(w) = \operatorname{conv}\left(\bigcup_{\pi \in \operatorname{argsort}(\ell(w))} \sum_{i=1}^n \sigma_i \partial \ell_{\pi(i)}(w) \right),$$

where $\operatorname{argsort}(\ell(w)) = \left\{\pi : \ell_{\pi(1)}(w) \leq \ldots \leq \ell_{\pi(n)}(w)\right\}$. *Moreover, if each ℓ_i is G-Lipschitz continuous, $\mathcal{R}_\sigma$ is also G-Lipschitz continuous.*

Proof. Since the coefficients $\sigma = (\sigma_1, \ldots, \sigma_n)$ are non-decreasing, the function $\mathcal{R}_\sigma$ can be written as the maximum over all possible permutations of the losses, i.e.,

$$\mathcal{R}_\sigma(w) = \max_{\pi \in \Pi_n} \sum_{i=1}^n \sigma_i \ell_{\pi(i)}(w) = \max_{\pi \in \Pi_n} \sum_{i=1}^n \sigma_{\pi^{-1}(i)} \ell_i(w),$$

where Π_n is the set of permutations of $\{1, \ldots, n\}$. For any $\pi \in \Pi_n$, $w \mapsto \sum_{i=1}^n \sigma_{\pi^{-1}(i)} \ell_i(w)$ is a convex combination of convex functions, hence it is convex. Since the pointwise maximum of convex functions is convex, $\mathcal{R}_\sigma$ is convex.

The pointwise maximum $f = \max_{j=1,\ldots,N} f_j$ of N convex functions $\{f_j\}_{j=1}^N$ has a subdifferential defined by $\partial f(x) = \operatorname{conv} \bigcup_{j \in \arg\max\{f_j(x)\}} \partial f_j(x)$ where $\operatorname{conv}(A)$ denotes the convex hull of a set A (Hiriart-Urruty and Lemaréchal, 1993, Lemma 4.4.1). Letting $N = n!$, consider the finite set of convex functions $\{f_\pi : \pi \in \Pi_n\}$ with each $f_\pi : w \mapsto \sum_{i=1}^n \sigma_i \ell_{\pi(i)}(w)$. The subdifferential of f_π is $\partial f_\pi(w) = \sum_{i=1}^n \sigma_i \partial \ell_{\pi(i)}(w)$, where the sum is to be understood as a Minkowski sum of sets (Hiriart-Urruty and Lemaréchal, 1993, Lemma 4.4.1). Hence, the subdifferential of $\mathcal{R}_\sigma(w)$ is defined by

$$\partial \mathcal{R}_\sigma(w) = \operatorname{conv}\left(\bigcup_{\pi \in \arg\max f_\pi(w)} \partial f_\pi(w) \right) = \operatorname{conv}\left(\bigcup_{\pi \in \operatorname{argsort}(\ell(w))} \sum_{i=1}^n \sigma_i \partial \ell_{\pi(i)}(w) \right),$$

where we used that $\operatorname{argsort}(\ell(w)) = \arg\max_\pi \sum_{i=1}^n \sigma_i \ell_{\pi(i)}(w)$ when $\sigma_1 \leq \cdots \leq \sigma_n$.

Finally if all ℓ_i are G-Lipschitz continuous, i.e., have G-bounded subgradients, then for any permutation π of $\{1, \ldots, n\}$, any $g \in \sum_{i=1}^n \sigma_i \partial \ell_{\pi(i)}(w)$ is bounded by G as a convex combination of G-bounded vectors. Hence any $g \in \partial \mathcal{R}_\sigma(w)$ is bounded by G as a convex combination of G-bounded vectors. The function $\mathcal{R}_\sigma$ is then convex with subgradients bounded by G, hence it is G-Lipschitz continuous.

□

C BIASED SGD CONVERGENCE ANALYSIS

Recall the regularized L-risk considered

$$\min_{w \in \mathbb{R}^d} \mathcal{R}_\sigma(w) + \frac{\mu}{2}\|w\|_2^2 \quad \text{for } \mathcal{R}_\sigma(w) = \sum_{i=1}^n \sigma_i \ell_{(i)}(w), \tag{11}$$

where $\mu > 0$, $0 \leq \sigma_1 \leq \cdots \leq \sigma_n$, $\sum_{i=1}^n \sigma_i = 1$, each $\ell_i : \mathbb{R}^d \to \mathbb{R}$ is a function representing performance of model weights w on training instance i and $\ell_{(1)}(w) \leq \ldots \leq \ell_{(n)}(w)$. In the following, we consider G-Lipschtiz continuous convex losses. The aim of this section is to prove the following proposition.

Proposition 3. *If the losses $\ell_1, ..., \ell_n$ are G-Lipschitz continuous, differentiable, and convex, the output $\bar{w}^{(T)}$ of Alg. 1 with $\eta^{(t)} = \frac{1}{\mu(t+1)}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}_{\sigma,\mu}\left(\bar{w}^{(T)}\right)\right] - \mathcal{R}_{\sigma,\mu}(w^*) \leq \underbrace{2\sqrt{2}C_s B_\mu \sqrt{\frac{n-m}{m(n-1)}}}_{\text{bias term}} + \underbrace{\frac{2G^2(1+\log T)}{\mu T}}_{\text{optimization term}}.$$

for $w^* = \arg\min_{w\in\mathbb{R}^d} \mathcal{R}_{\sigma,\mu}(w)$, $C_s = \sup_{t\in(0,1)} |s(t) - u(t)|$, and $B_\mu = \sup_{w:\|w\|_2 \leq G/\mu} \max_{i=1,\ldots,n} |\ell_i(w)| < \infty$. The expectation is taken over the sampling of each minibatch.

The proof will proceed in three parts, which comprise the next three subsections. The final subsection proves the main result.

1. The convexity and G-Lipschitz continuity of the losses is used to analyze the convergence of Alg. 1 on a surrogate objective for which there is no bias.
2. We then establish a uniform bias bound between the surrogate function and the original function over a set $\mathcal{W} \subseteq \mathbb{R}^d$.
3. We then relate the suboptimality gap of the surrogate objective to the suboptimality of the original objective by using the bias bound and establishing that the iterates and minimizers are contained in $\mathcal{W}$.

C.1 SGD Analysis for Convex, Lipschitz Loss and Strongly Convex Regularizer

We present first a generic convergence result for stochastic subgradient algorithms such as Alg. 1 applied to regularized non-smooth functions in Lem. 11, which is a minor adaptation of Bach (2023, Theorem 5.5).

Lemma 11. *Consider a G-Lipschitz continuous, convex function $f : \mathbb{R}^d \to \mathbb{R}$ and a regularization $\mu\|\cdot\|_2^2$ for $\mu > 0$ defining an objective of the form $f_\mu(w) = f(w) + \mu\|w\|_2^2/2$. Given a an initial point $w^{(0)} = 0 \in \mathbb{R}^d$ consider iterates of the form, for $t \geq 0$,*

$$w^{(t+1)} = w^{(t)} - \eta^{(t)}(v^{(t)} + \mu w^{(t)}),$$

for $v^{(t)}$ a random vector whose distribution depends only on $w^{(t)}$ satisfying $\mathbb{E}\left[v^{(t)} \mid w^{(t)}\right] \in \partial f(w^{(t)})$ and $\left\|v^{(t)}\right\|_2 \leq G$, and $\eta^{(t)} > 0$. Consider outputting after T iterations the estimate $\bar{w}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} w^{(t)}$. Provided that $\eta^{(t)} = \frac{1}{\mu(t+1)}$, this estimate satisfies

$$\mathbb{E}[f_\mu(\bar{w}^{(T)})] - f_\mu(w^*) \leq \frac{2G^2(1+\log T)}{\mu T},$$

for $w^* = \arg\min_{w\in\mathbb{R}^d} f_\mu(w)$, where the expectation is taken over the sequence $w^{(1)}, \ldots, w^{(T-1)}$.

Proof. Write the expansion

$$\left\|w^{(t+1)} - w^*\right\|_2^2 = \left\|w^{(t)} - w^*\right\|_2^2 - 2\eta^{(t)}\left\langle v^{(t)} + \mu w^{(t)}, w^{(t)} - w^*\right\rangle + (\eta^{(t)})^2\left\|v^{(t)} + \mu w^{(t)}\right\|_2^2. \tag{12}$$

Because $\left\|w^{(0)}\right\|_2 = \|0\|_2 \leq G/\mu$, $\|v^{(t)}\| \leq G$ and $w^{(t+1)}$ can be expressed as a convex combination of $w^{(t)}$ and $-v^{(t)}/\mu$, that is,

$$w^{(t+1)} = (1 - \eta^{(t)}\mu)w^{(t)} + \eta^{(t)}\mu\left(-\frac{1}{\mu}v^{(t)}\right),$$

we can conclude by induction that $\left\|w^{(t)}\right\|_2 \leq G/\mu$ for all $t = 0, \ldots, T-1$ when $\eta^{(t)}\mu \leq 1$, which is satisfied for our choice of $\eta^{(t)}$. Thus, $(\eta^{(t)})^2\left\|v^{(t)} + \mu w^{(t)}\right\|_2^2 \leq (\eta^{(t)})^2 \cdot 4G^2$. Taking the conditional expectation of $v^{(t)}$ given $w^{(t)}$ of (12) yields

$$\mathbb{E}\left[\left\|w^{(t+1)} - w^*\right\|_2^2 \mid w^{(t)}\right] \leq \left\|w^{(t)} - w^*\right\|_2^2 - 2\eta^{(t)}\left\langle \mathbb{E}\left[v^{(t)} \mid w^{(t)}\right] + \mu w^{(t)}, w^{(t)} - w^*\right\rangle + (\eta^{(t)})^2 4G^2. \tag{13}$$

Because $\mathbb{E}\left[v^{(t)} \mid w^{(t)}\right] \in \partial f(w^{(t)})$, we have that $\mathbb{E}\left[v^{(t)} \mid w^{(t)}\right] + \mu w^{(t)} \in \partial f_\mu(w^{(t)})$, and by the μ-strong convexity of f_μ, we have

$$-\left\langle \mathbb{E}\left[v^{(t)} \mid w^{(t)}\right] + \mu w^{(t)}, w^{(t)} - w^* \right\rangle \leq -\left(f_\mu(w^{(t)}) - f_\mu(w^*)\right) - \frac{\mu}{2}\left\|w^{(t)} - w^*\right\|_2^2,$$

which, substituted into (13) gives

$$\begin{aligned}
\mathbb{E}\left[\left\|w^{(t+1)} - w^*\right\|_2^2 \mid w^{(t)}\right] &\leq \left(1 - \eta^{(t)}\mu\right)\left\|w^{(t)} - w^*\right\|_2^2 - 2\eta^{(t)}\left(f_\mu(w^{(t)}) - f_\mu(w^*)\right) + (\eta^{(t)})^2 4G^2 \\
\implies f_\mu(w^{(t)}) - f_\mu(w^*) &\leq \frac{1}{2}\left(\left(\frac{1}{\eta^{(t)}} - \mu\right)\left\|w^{(t)} - w^*\right\|_2^2 - \frac{1}{\eta^{(t)}}\mathbb{E}\left[\left\|w^{(t+1)} - w^*\right\|_2^2 \mid w^{(t)}\right]\right) + 2\eta^{(t)}G^2 \\
&= \frac{1}{2}\left(\mu t\left\|w^{(t)} - w^*\right\|_2^2 - \mu(t+1)\mathbb{E}\left[\left\|w^{(t+1)} - w^*\right\|_2^2 \mid w^{(t)}\right]\right) + \frac{2G^2}{\mu(t+1)}.
\end{aligned}$$

Take the expectation over the entire sequence $w^{(0)}, \ldots, w^{(t)}$, sum over $t = 0, \ldots, T-1$, and divide by T to get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} f_\mu(w^{(t)})\right] - f_\mu(w^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{2G^2}{\mu(t+1)} \leq \frac{2G^2(1+\log T)}{\mu T},$$

which combined with $f_\mu(\bar{w}^{(T)}) \leq \frac{1}{T}\sum_{t=0}^{T-1} f_\mu(w^{(t)})$ completes the proof. □

C.2 Bias Control

In this section, we control the bias term appearing in the convergence analysis. The following lemmas consider a set of real numbers, representing losses at a single $w \in \mathbb{R}^d$. Let $x_1, \ldots, x_n \in \mathbb{R}$ be call the *full batch*, and let $X_1, \ldots X_m$ be a random sample selected uniformly *without* replacement from $\{x_1, \ldots, x_n\}$, called the *minibatch*. Let

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(x_i) \text{ and } F_{n,m}(x) := \frac{1}{m}\sum_{j=1}^{m} 1_{(-\infty,x]}(X_j)$$

be the empirical CDFs, and let

$$F_n^{-1}(t) := \inf\{x : F_n(x) \geq t\} \text{ and } F_{n,m}^{-1}(t) := \inf\{x : F_{n,m}(x) \geq t\}.$$

be the empirical quantile functions of the full batch and minibatch respectively. Similarly, let

$$\mu_n := \sum_{i=1}^{n}\delta_{x_i} \text{ and } \mu_{n,m} = \sum_{j=1}^{m}\delta_{X_j}$$

be the empirical measures of the full batch and minibatch, respectively, with δ_x indicating a Dirac point mass at x. Let $u(t) := 1_{(0,1)}(t)$ be the uniform spectrum. Note that $F_{n,m}$ and $\mu_{n,m}$ are random, since they depend on the random sample $X_1, \ldots, X_m$.

The first result shows that the minibatch L-risk is an unbiased estimator of the full batch L-risk for the case of the uniform spectrum.

Lemma 12. *We have that for any $n \in \mathbb{N}$, $m \leq n$,*

$$\mathbb{E}\left[\mathbb{L}_u[F_{n,m}]\right] = \mathbb{L}_u[F_n],$$

where the expectation is taken over the sampling of $X_1, \ldots, X_m$ without replacement.

Proof.

$$\mathbb{E}\left[\mathbb{L}_u[F_{n,m}]\right] = \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} X_{(j)}\right] = \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} X_j\right] = \frac{1}{\binom{n}{m}}\sum_{i_1<\ldots<i_m}\frac{1}{m}\sum_{j=1}^{m} x_{i_j} = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{L}_u[F_n].$$

□

Recall the definition of the 1-Wasserstein distance between probability measures μ and ν over $\mathbb{R}$ with finite first moment, given by

$$W_1(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}} |x - y| \, \mathrm{d}\gamma(x, y),$$

where $\Pi(\mu, \nu)$ is the set of couplings (or joint distributions) with marginals being μ and ν.

Lemma 13. *For any $n \in \mathbb{N}, m \leq n$ the 1-Wasserstein distance between the empirical measures satisy*

$$\mathbb{E}_{\mu_n} \left[W_1(\mu_{n,m}, \mu_n) \right] \leq \sqrt{2 \frac{n-m}{m(n-1)}} \max_{i=1,\ldots,n} |x_i| \,,$$

where $\mathbb{E}_{\mu_n}[\cdot]$ denotes the expected value according to the sampling of $X_1, \ldots, X_m$ without replacement from $\{x_1, \ldots, x_n\}$.

Proof. Note that both μ_n and any realization of $\mu_{n,m}$ have finite first moment. Then, we have that

$$\begin{aligned}
\mathbb{E}_{\mu_n} \left[W_1(\mu_{n,m}, \mu_n) \right] &= \mathbb{E}_{\mu_n} \left[\int_{\mathbb{R}} |F_{n,m}(x) - F_n(x)| \, \mathrm{d}x \right] && \text{Thm. 7} \\
&= \int_{\mathbb{R}} \mathbb{E}_{\mu_n} \left[|F_{n,m}(x) - F_n(x)| \right] \mathrm{d}x \\
&\leq \int_{\mathbb{R}} \sqrt{\mathbb{E}_{\mu_n} \left[(F_{n,m}(x) - F_n(x))^2 \right]} \, \mathrm{d}x && \text{Jensen's inequality} \\
&= \int_{\mathbb{R}} \sqrt{\mathrm{Var}\left[F_{n,m}(x) \right]} \, \mathrm{d}x,
\end{aligned}$$

where the last line follows because $F_n(x)$ is the expected value of $F_{n,m}(x) = \frac{1}{m} \sum_{j=1}^{m} 1_{(-\infty, x]}(X_j)$ (even when sampling without replacement). Then, notice that due to the finite population sample correction when sampling without replacement, we have that

$$\mathrm{Var}\left[F_{n,m}(x) \right] = \frac{n-m}{m(n-1)} F_n(x)(1 - F_n(x)),$$

and so

$$\mathbb{E}_{\mu_n} \left[W_1(\mu_{n,m}, \mu_n) \right] \leq \sqrt{\frac{n-m}{m(n-1)}} \int_{\mathbb{R}} \sqrt{F_n(x)(1 - F_n(x))} \, \mathrm{d}x.$$

Next, because μ_n is a categorical distribution over a finite set, it has p-th moment for any $p > 0$. So, we may apply Lem. 10 to write

$$\int_{\mathbb{R}} \sqrt{F_n(x)(1 - F_n(x))} \leq \sqrt{2} \frac{p}{p-2} \left(\frac{1}{n} \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \overset{p \to \infty}{\to} \sqrt{2} \max_{i=1,\ldots,n} |x_i| \,,$$

completing the proof. □

Next, we can give the bias bound for an L-functional applied to $\mu_{n,m}$ and μ_n.

Lemma 14. *Let $\mathbb{L}_s$ be an L-functional with spectrum s, and let u be the uniform spectrum. Then,*

$$\left| \mathbb{E}\left[\mathbb{L}_s[F_{n,m}] \right] - \mathbb{L}_s[F_n] \right| \leq \sqrt{2 \frac{n-m}{m(n-1)}} \left\| s - u \right\|_\infty \max_{i=1,\ldots,n} |x_i| \,,$$

where the expectation is taken over the sampling of $X_1, \ldots, X_m$ without replacement, and $\|s - u\|_\infty = \sup_{t \in (0,1)} |s(t) - u(t)|$.

Proof. Write

$$\begin{aligned}
|\mathbb{E}\left[\mathbb{L}_s[F_{n,m}]\right] - \mathbb{L}_s[F_n]| &= |\mathbb{E}\left[\mathbb{L}_s[F_{n,m}]\right] - \mathbb{L}_s[F_n] - (\mathbb{E}\left[\mathbb{L}_u[F_{n,m}]\right] - \mathbb{L}_u[F_n])| && \text{Lem. 12} \\
&= \left|\mathbb{E}\left[\int_0^1 (s(t)-u(t)) \cdot \left(F_{n,m}^{-1}(t) - F_n^{-1}(t)\,\mathrm{d}t\right)\right]\right| \\
&\leq \mathbb{E}\left[\left|\int_0^1 (s(t)-u(t)) \cdot \left(F_{n,m}^{-1}(t) - F_n^{-1}(t)\,\mathrm{d}t\right)\right|\right] && \text{Jensen's inequality} \\
&\leq \|s-u\|_\infty \mathbb{E}\left[\left\|F_{n,m}^{-1} - F_n^{-1}\right\|_1\right] && \text{Hölder's inequality} \\
&= \|s-u\|_\infty \mathbb{E}\left[W_1\left(\mu_{n,m}, \mu_n\right)\right]. && \text{Thm. 7}
\end{aligned}$$

Applying Lem. 13 achieves the desired result. □

Next, consider the situation in which $x_1 = \ell_1(w), \ldots, x_n = \ell_n(w)$, i.e., the loss functions for each data point evaluated at w, with $X_j = \ell_{i_j}(w)$ for the randomly samplied minibatch $(i_1, \ldots, i_m)$. By defining

$$F_n(x; w) := \frac{1}{n}\sum_{i=1}^n 1\left(\ell_i(w) \leq x\right) \text{ and } F_{n,m}(x; w) := \frac{1}{m}\sum_{j=1}^m 1\left(\ell_{i_j}(w) \leq x\right),$$

we have that

$$\mathcal{R}_\sigma(w) = \mathbb{L}_s[F_n(\cdot; w)] \text{ and } \mathcal{R}_{\hat{\sigma}}(w) := \mathbb{L}_s[F_{n,m}(\cdot; w)].$$

This gives the following corollary.

Corollary 15. *Given a minibatch $(i_1, \ldots, i_m)$ sampled uniformly randomly without replacement from $[n]$, define $\mathcal{R}_{\hat{\sigma}}(w) := \sum_{j=1}^m \hat{\sigma}_j \ell_{i_{(j)}}(w)$, where $\ell_{i_{(1)}}(w) \leq \ldots \leq \ell_{i_{(m)}}(w)$ are the minibatch order statistics. It holds that*

$$\sup_{w:\|w\|_2 \leq G/\mu} |\mathbb{E}\left[\mathcal{R}_{\hat{\sigma}}(w) \mid w\right] - \mathcal{R}_\sigma(w)| \leq \sqrt{2}\,\|s-u\|_\infty \sqrt{\frac{n-m}{m(n-1)}} B_\mu,$$

where $B_\mu = \sup_{w:\|w\| \leq G/\mu, i \in [n]} |\ell_i(w)| < \infty$.

Proof. For any $w \in \mathbb{R}^d$, we have that

$$\begin{aligned}
|\mathbb{E}\left[\mathcal{R}_{\hat{\sigma}}(w) \mid w\right] - \mathcal{R}_\sigma(w)| &= |\mathbb{E}\left[\mathbb{L}_s[F_{n,m}(\cdot; w)]\right] - \mathbb{L}_s[F_n(\cdot; w)]| \\
&\leq \sqrt{2}\,\|s-u\|_\infty \sqrt{\frac{n-m}{m(n-1)}} \max_{i=1,\ldots,n} |\ell_i(w)| && \text{Lem. 14.}
\end{aligned}$$

Take the supremum for $\{w : \|w\|_2 \leq G/\mu\}$ on both sides. Because each ℓ_i is continuous, so is $\max_{i=1,\ldots,n} |\ell_i(w)|$. Because the supremum is taken over a compact set (the ball of radius G/μ), it is finite. □

C.3 Proof of Main Result

Proposition 3. *If the losses $\ell_1, ..., \ell_n$ are G-Lipschitz continuous, differentiable, and convex, the output $\bar{w}^{(T)}$ of Alg. 1 with $\eta^{(t)} = \frac{1}{\mu(t+1)}$ satisfies*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_{\sigma,\mu}\left(\bar{w}^{(T)}\right)\right] - \mathcal{R}_{\sigma,\mu}(w^*) \\
&\leq \underbrace{2\sqrt{2} C_s B_\mu \sqrt{\frac{n-m}{m(n-1)}}}_{\textit{bias term}} + \underbrace{\frac{2G^2(1+\log T)}{\mu T}}_{\textit{optimization term}}.
\end{aligned}$$

for $w^* = \arg\min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu}(w)$, $C_s = \sup_{t \in (0,1)} |s(t) - u(t)|$, and $B_\mu = \sup_{w:\|w\|_2 \leq G/\mu} \max_{i=1,\ldots,n} |\ell_i(w)| < \infty$. The expectation is taken over the sampling of each minibatch.

Proof. Define the surrogate function

$$\bar{\mathcal{R}}_{\hat{\sigma}}(w) := \mathbb{E}\left[\mathcal{R}_{\hat{\sigma}}(w) \mid w\right] = \mathbb{E}\left[\sum_{j=1}^{m} \hat{\sigma}_j \ell_{i_{(j)}}(w) \mid w\right],$$

where the expectation is taken over sampling the minibatch $(i_1, \ldots, i_m)$. The expectation is over a discrete distribution on a finite set, so it is a well-defined function of w. We now establish the properties of $\bar{\mathcal{R}}_{\hat{\sigma}}$ required to apply the generic analysis of the stochastic subgradient method given in Lem. 11 with $f = \bar{\mathcal{R}}_{\hat{\sigma}}$ and $v^{(t)} = \sum_{j=1}^{m} \hat{\sigma}_j \nabla \ell_{i_{(j)}}(w)$. This choice of $\bar{\mathcal{R}}_{\hat{\sigma}}$ is clearly convex, and because $v^{(t)}$ is a subgradient of $\mathcal{R}_{\hat{\sigma}}$, $\mathbb{E}\left[v^{(t)} \mid w^{(t)}\right]$ is a subgradient of $\bar{\mathcal{R}}_{\hat{\sigma}}$. Given the G-Lipschitzness of each ℓ_i, we have that $\mathcal{R}_{\hat{\sigma}}$ is also G-Lipschitz by Prop. 2, and thus so is $\bar{\mathcal{R}}_{\hat{\sigma}}$. Then, letting $\bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) := \bar{\mathcal{R}}_{\hat{\sigma}}(w) + \frac{\mu}{2}\|w\|_2^2$, applying Lem. 11 gives

$$\mathbb{E}\left[\bar{\mathcal{R}}_{\hat{\sigma},\mu}\left(\bar{w}^{(T)}\right)\right] - \bar{\mathcal{R}}_{\hat{\sigma},\mu}\left(\bar{w}^*\right) \leq \frac{2G^2(1+\log T)}{\mu T}, \tag{14}$$

where $\bar{w}^* = \arg\min_{w \in \mathbb{R}^d} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w)$. We must now pass this result regarding $\bar{\mathcal{R}}_{\hat{\sigma},\mu}$ to a similar one regarding $\mathcal{R}_{\sigma,\mu}$. Define $\mathcal{W} := \{w \in \mathbb{R}^d : \|w\|_2 \leq G/\mu\}$. We first establish that $w^*, \bar{w}^* \in \mathcal{W}$, so that

$$\min_{w \in \mathbb{R}^d} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) = \min_{w \in \mathcal{W}} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) \text{ and } \min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu}(w) = \min_{w \in \mathcal{W}} \mathcal{R}_{\sigma,\mu}(w). \tag{15}$$

The subdifferentials of $\bar{\mathcal{R}}_{\hat{\sigma},\mu}$ and $\mathcal{R}_{\sigma,\mu}$ are given by

$$\partial\bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) = \partial\bar{\mathcal{R}}_{\hat{\sigma}}(w) + \mu w = \{g + \mu w : g \in \partial\bar{\mathcal{R}}_{\hat{\sigma}}(w)\},$$
$$\partial\mathcal{R}_{\sigma,\mu}(w) = \partial\mathcal{R}_{\sigma}(w) + \mu w = \{g + \mu w : g \in \partial\mathcal{R}_{\sigma}(w)\}.$$

Then, by optimality,

$$g^* + \mu w^* = 0 \text{ and } \bar{g} + \mu \bar{w}^* = 0$$

for some $g^* \in \partial\mathcal{R}_{\sigma}(w^*)$ and $\bar{g} \in \partial\bar{\mathcal{R}}_{\hat{\sigma}}(\bar{w}^*)$, yielding that $w^*, \bar{w}^* \in \mathcal{W}$ because $\|g^*\|_2, \|\bar{g}\|_2 \leq G$. Next, note that $\left\|w^{(t)}\right\|_2 \leq G/\mu$ for any $t = 0, \ldots, T$. To see this, observe that $\left\|w^{(0)}\right\|_2 = \|0\|_2 \leq G/\mu$, and if $\left\|w^{(t)}\right\|_2 \leq G/\mu$, then

$$\left\|w^{(t+1)}\right\|_2 = \left\|(1-\eta^{(t)}\mu)w^{(t)} + \eta^{(t)}\mu\left(-\frac{1}{\mu}v^{(t)}\right)\right\|_2 \leq (1-\eta^{(t)}\mu)\left\|w^{(t)}\right\|_2 + \eta^{(t)}\mu\left\|\frac{1}{\mu}v^{(t)}\right\|_2 \leq \frac{G}{\mu}$$

if $\eta^{(t)}\mu \leq 1$, which is satisfied for $\eta^{(t)} = 1/(\mu t)$. By convexity of $\mathcal{W}$, this means that $\bar{w}^{(T)} \in \mathcal{W}$. Given that $\bar{w}^*, w^*$, and $\bar{w}^{(T)}$ are contained in $\mathcal{W}$, if we can show that $\mathcal{R}_{\sigma}$ and $\bar{\mathcal{R}}_{\hat{\sigma}}$ are close on this set, then optimizing $\bar{\mathcal{R}}_{\hat{\sigma}}$ should also result in a near-optimal value of $\mathcal{R}_{\sigma}$. Assume there existed $\delta > 0$ such that $\sup_{w \in \mathcal{W}} \left|\mathcal{R}_{\sigma}(w) - \bar{\mathcal{R}}_{\hat{\sigma}}(w)\right| = \delta < \infty$. Then, $\mathcal{R}_{\sigma}\left(\bar{w}^{(T)}\right) \leq \bar{\mathcal{R}}_{\hat{\sigma}}\left(\bar{w}^{(T)}\right) + \delta$, and, for any $w \in \mathcal{W}$,

$$\mathcal{R}_{\sigma}(w) \geq \bar{\mathcal{R}}_{\hat{\sigma}}(w) - \delta \implies -\min_{w \in \mathcal{W}}\left(\mathcal{R}_{\sigma}(w) + \frac{\mu}{2}\|w\|_2^2\right) \leq -\min_{w \in \mathcal{W}}\left(\bar{\mathcal{R}}_{\hat{\sigma}}(w) + \frac{\mu}{2}\|w\|_2^2\right) + \delta, \tag{16}$$

giving

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}_{\sigma,\mu}\left(\bar{w}^{(T)}\right)\right] - \min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu}(w) &= \mathbb{E}\left[\mathcal{R}_{\sigma,\mu}\left(\bar{w}^{(T)}\right)\right] - \min_{w \in \mathcal{W}} \mathcal{R}_{\sigma,\mu}(w) && (15)\\
&\leq 2\delta + \mathbb{E}\left[\bar{\mathcal{R}}_{\hat{\sigma},\mu}\left(\bar{w}^{(T)}\right)\right] - \min_{w \in \mathcal{W}} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) && (16)\\
&= 2\delta + \mathbb{E}\left[\bar{\mathcal{R}}_{\hat{\sigma},\mu}\left(\bar{w}^{(T)}\right)\right] - \min_{w \in \mathbb{R}^d} \bar{\mathcal{R}}_{\hat{\sigma},\mu}(w) && (15)\\
&\leq 2\delta + \frac{2G^2(1+\log T)}{\mu T}. && (14)
\end{aligned}$$

Establishing the existence of such a δ and showing $2\delta \leq 2\sqrt{2}C_s B_\mu \sqrt{\frac{n-m}{m(n-1)}}$ completes the proof. This is accomplished by Cor. 15, which gives

$$2\delta = 2\sup_{w \in \mathcal{W}}\left|\bar{\mathcal{R}}_{\hat{\sigma}}(w) - \mathcal{R}_{\sigma}(w)\right| = 2\sup_{w : \|w\|_2 \leq G/\mu}\left|\mathbb{E}\left[\mathcal{R}_{\hat{\sigma}}(w) \mid w\right] - \mathcal{R}_{\sigma}(w)\right| \leq 2\sqrt{2}\|s-u\|_\infty \sqrt{\frac{n-m}{m(n-1)}} B_\mu,$$

the desired result. □

D SMOOTHING THE EMPIRICAL SPECTRAL RISK MEASURE

Recall that we consider objectives of the form (ignoring the regularization part)

$$\mathcal{R}_\sigma(w) = \sum_{i=1}^n \sigma_i \ell_{(i)}(w), \tag{17}$$

where σ_i are the weights associated with the discretization of a spectral risk, i.e., $\sigma_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} s(t)\mathrm{d}t$ for $s : (0,1) \to [0, +\infty)$ non-decreasing and such that $\int_0^1 s(t)\mathrm{d}t = 1$.

We can rewrite problems (17) as minimizing a composition

$$\mathcal{R}_\sigma(w) = h(\ell(w)) \text{ with } h(l) = \sum_{i=1}^n \sigma_i l_{(i)} \text{ and } \ell(w) = (\ell_1(w), \dots, \ell_n(w)).$$

Since the coefficients σ_i are not decreasing, the outer function h can be expressed as

$$h(l) = \max_{\lambda \in \mathcal{P}(\sigma)} \lambda^\top l$$

where $\mathcal{P}(\sigma) = \{\lambda = \Pi\sigma : \Pi\mathbf{1} = \mathbf{1}, \Pi^\top \mathbf{1} = \mathbf{1}, \Pi \in [0,1]^{n\times n}\}$ is the permutahedron associated with the weights σ.

The function h is non-differentiable at points l with ties. However, smooth approximations of h can be defined by means of a strongly convex regularizer as presented by Nesterov (2005). For Ω strongly convex w.r.t. some norm $\|\cdot\|$ and $\nu \geq 0$, we consider a smooth approximation of h defined by

$$h_{\nu\Omega}(l) := \max_{\lambda \in \mathcal{P}(\sigma)} \left\{l^\top \lambda - \nu\Omega(\lambda)\right\}, \text{ with } \nabla h_{\nu\Omega}(l) = \operatorname*{arg\,max}_{\lambda \in \mathcal{P}(\sigma)} \left\{l^\top \lambda - \nu\Omega(\lambda)\right\}.$$

By standard convex duality arguments, the smooth approximation of h can also be written as the inf-convolution of h with the convex conjugate of $\nu\Omega$, i.e,

$$h_{\nu\Omega}(l) = \min_{z \in \mathbb{R}^n} \left\{h(z) + \nu\Omega^\star((l-z)/\nu)\right\}, \ \nabla h_{\nu\Omega}(l) = \nabla\Omega^\star((l-z^*)/\nu) \text{ for } z^* = \operatorname*{arg\,min}_{z \in \mathbb{R}^n} \left\{h(z) + \nu\Omega^\star((l-z)/\nu)\right\}. \tag{18}$$

We consider the surrogate objective defined by

$$\mathcal{R}_{\sigma,\nu\Omega}(w) := h_{\nu\Omega}(\ell(w)), \text{with } h_{\nu\Omega}(l) := \max_{\lambda \in \mathcal{P}(\sigma)} \left\{l^\top \lambda - \nu\Omega(\lambda)\right\} \text{ and } \ell(w) = (\ell_1(w), \dots, \ell_n(w)). \tag{19}$$

Note that for any $\nu \geq 0$, if the losses ℓ_i are convex, then the surrogate objective $\mathcal{R}_{\sigma,\nu\Omega}$ is also convex.

In the following, we recall the smoothness properties of the smooth approximation in Lem. 16, we present the approximation incurred by the smoothing in terms of the spectrum in Lem. 17. We give the implementation of the gradient evaluation of the smooth approximation for appropriate choices of regularizers in Sec. D.2.

D.1 Smoothing Properties and Approximation Bounds

We recall below the smoothness properties of $h_{\nu\Omega}(l)$, see, e.g., (Nesterov, 2005; Beck and Teboulle, 2012) for detailed proofs.

Lemma 16 (Smoothing properties). *For any Ω and $\nu > 0$, the smoothed $h_{\nu\Omega}$ is $\|\sigma\|_p$-Lipschitz continuous w.r.t. $\|\cdot\|_p$ for any $p \in \{1, \dots\} \cup \{+\infty\}$. For any Ω that is 1-strongly convex w.r.t. $\|\cdot\|$, the smoothed $h_{\nu\Omega}$ is $1/\nu$ smooth w.r.t. to the dual norm $\|\cdot\|_*$, i.e., for any $l, l' \in \mathbb{R}^n$, $\|\nabla h_{\nu\Omega}(l) - \nabla h_{\nu\Omega}(l')\| \leq \|l - l'\|_*/\nu$.*

Usual examples are $\Omega = \|\cdot\|_2^2$ or $\Omega = H : \lambda \mapsto \sum_{i=1}^n \lambda_i \ln \lambda_i$, for which we have access to $h_{\nu\Omega}$ by isotonic regression, e.g., (Lim and Wright, 2016; Blondel et al., 2020). In the following, we consider such functions centered around their minimizers in $\mathcal{P}(\sigma)$ to get tighter approximation bounds. Namely, we define $u_n = \mathbf{1}/n \in \mathcal{P}(\sigma)$ and we consider

$$\Omega_1(\lambda) = D_H(\lambda; u_n) := H(\lambda) - H(u_n) - \nabla H(u_n)^\top(\lambda - u_n) = \sum_{i=1}^n \lambda_i \log(n\lambda_i), \quad \text{and} \tag{20}$$

$$\Omega_2(\lambda) = \frac{1}{2}\|\lambda - u_n\|_2^2. \tag{21}$$

We have that Ω_1 is 1-strongly convex w.r.t. $\|\cdot\|_1$ and Ω_2 is 1-strongly convex w.r.t. $\|\cdot\|_2$.

We can then consider optimizing the surrogate objective for $\Omega \in \{\Omega_1, \Omega_2\}$, defined by $\mathcal{R}_{\sigma,\nu\Omega}(w) := h_{\nu\Omega}(\ell(w))$. Lem. 17 details the approximation done by considering the smoothed version of the objective.

Lemma 17 (Approximation bounds)**.** *For any strongly convex function Ω invariant by permutation and such that $\inf_{\lambda\in\mathcal{P}(\sigma)} \Omega(\lambda) \geq 0$, we have that for any $\nu \geq 0$, $l \in \mathbb{R}^n$,*

$$0 \leq h(l) - h_{\nu\Omega}(l) \leq \nu\Omega(\sigma)$$

If, in addition, Ω is decomposable as $\Omega(\lambda) = \sum_{i=1}^n \omega(\lambda_i)$ with ω convex and σ is the discretization of a function s such that $\sigma_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} s(t)\mathrm{d}t$, then

$$\Omega(\sigma) \leq n \int_0^1 \omega\left(\frac{s(t)}{n}\right) \mathrm{d}t.$$

Proof. One one hand, we have that

$$h_{\nu\Omega}(l) = \sup_{\lambda\in\mathcal{P}(\sigma)} \{\lambda^\top l - \nu\Omega(\lambda)\} \leq \sup_{\lambda\in\mathcal{P}(\sigma)} \lambda^\top l = h(l),$$

since $\inf_{\lambda\in\mathcal{P}(\sigma)} \Omega(\lambda) \geq 0$. On the other hand, we have that

$$h_{\nu\Omega}(l) = \sup_{\lambda\in\mathcal{P}(\sigma)} \{\lambda^\top l - \nu\Omega(\lambda)\} \geq \sup_{\lambda\in\mathcal{P}(\sigma)} \{\lambda^\top l\} - \nu \sup_{\lambda\in\mathcal{P}(\sigma)} \Omega(\lambda) = h(l) - \nu \sup_{\lambda\in\mathcal{P}(\sigma)} \Omega(\lambda).$$

Hence, we have $0 \leq h(l) - h_{\nu\Omega}(l) \leq \nu \max_{\lambda\in\mathcal{P}(\sigma)} \Omega(\lambda)$. The maximum of a convex function Ω over a polytope $\mathcal{P}(\sigma)$ is attained at a corner. The corners of the permutahedron $\mathcal{P}(\sigma)$ are permutations of σ. Since Ω is permutation invariant, we have that $\Omega(\sigma) = \Omega(\pi(\sigma))$ for any permutation π. Thus, $\max_{\lambda\in\mathcal{P}(\sigma)} \Omega(\lambda) = \Omega(\sigma)$, completing the proof of the first part. For the second claim, we use Jensen's inequality to get

$$\begin{aligned}\Omega(\sigma) = \sum_{i=1}^n \omega\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} s(t)\mathrm{d}t\right) &= \sum_{i=1}^n \omega\left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} \left(\frac{s(t)}{n}\right) n\,\mathrm{d}t\right) \\ &\leq \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \omega\left(\frac{s(t)}{n}\right) n\,\mathrm{d}t = n\int_0^1 \omega\left(\frac{s(t)}{n}\right)\mathrm{d}t.\end{aligned}$$

□

Corollary 18. *For $\sigma_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} s(t)\mathrm{d}t$ with s a spectrum such that $\int_0^1 s(t)\mathrm{d}t = 1$, and for any $\nu \geq 0$, $l \in \mathbb{R}^n$, we have*

$$\begin{aligned}0 \leq h(l) - h_{\nu\Omega_1}(l) &\leq \nu D_H(\sigma; u_n) \leq \nu \int_0^1 s(t)\ln s(t)\mathrm{d}t := \nu\,\mathrm{KL}(s\|u), \\ 0 \leq h(l) - h_{\nu\Omega_2}(l) &\leq \frac{\nu}{2}\|\sigma - u_n\|_2^2 \leq \frac{\nu}{2n}\int_0^1 (s(t)-1)^2\mathrm{d}t := \frac{\nu}{2n}\chi^2(s\|u),\end{aligned}$$

where $\mathrm{KL}(s\|u)$ and $\chi^2(s\|u)$ denote respectively the Kullback-Leibler divergence and the Chi-square divergence between the spectrum s and the uniform distribution u.

For example, we can derive the bounds for some specific choices of spectra.

1. (Superquantile) For $s(t) = \frac{1}{1-q}\mathbf{1}_{[q,1]}(t)$, with $q \in [0,1]$, we have $\chi^2(s\|u) = \frac{q}{1-q}$ and $\mathrm{KL}(s\|u) = -\ln(1-q)$.
2. (Extremile) For $s(t) = rt^{r-1}$, with $r \geq 1$, we have $\chi^2(s\|u) = \frac{(r-1)^2}{(2r-1)}$ and $\mathrm{KL}(s\|u) = \ln r + \frac{1}{r} - 1$.

The approximations bounds computed for $h_{\nu\Omega}$ and h naturally apply for $\mathcal{R}_{\sigma,\nu\Omega}$ and $\mathcal{R}_\sigma$, that is, for any $w \in \mathbb{R}^d$, we have $0 \leq \mathcal{R}_\sigma(w) - \mathcal{R}_{\sigma,\nu\Omega}(w) \leq \nu\Omega(\sigma)$. This gives the following lemma mentioned in the main text.

Lemma 19. *Consider the regularized objective $\mathcal{R}_{\sigma,\mu}(w) = \mathcal{R}_\sigma(w) + \mu\|w\|_2^2/2$ for $\mathcal{R}_\sigma$ defined as in (17) by non-decreasing non-negative coefficients σ_i summing up to 1 and n functions $(\ell_i)_{i=1}^n$, and consider the smoothed approximation*

Algorithm 3 Pool Adjacent Violators (PAV) Algorithm for ω_1

Inputs: Number of coefficients n, coefficients $(l_i)_{i=1}^n$ and $(s_i)_{i=1}^n$ with $s_i = \ln \sigma_i$
Initialize $P_1 = \{1\}$, $\mathcal{P} = (P_1)$, $v_1 = l_1 - s_1 - \ln n$, $L_1 = l_1$, $M_1 = s_1$, $d = 1$.
for $i = 2, \dots n$ **do**
  Set $P_{d+1} = \{i\}$, $\mathcal{P} \leftarrow (P_1, \dots, P_{d+1})$, $v_{d+1} = l_i - s_i - \ln n$, $L_{d+1} = l_i$, $M_{d+1} = s_i$ and $d = d + 1$,
  while $d \geq 2$ and $v_{d-1} \geq v_d$ **do**
    Set $v_{d-1} \leftarrow \mathrm{LSE}(L_{d-1}, L_d) - \mathrm{LSE}(M_{d-1}, M_d) - \ln n$
    Set $L_{d-1} \leftarrow \mathrm{LSE}(L_{d-1}, L_d)$, $M_{d-1} \leftarrow \mathrm{LSE}(M_{d-1}, M_d)$
    Set $\mathcal{P} \leftarrow (P_1, \dots, P_{d-1} \cup P_d)$
    Set $d \leftarrow d - 1$
Output: $z \in \mathbb{R}^n$ such that $z_i = v_s$ for $i \in P_s$, $s \in \{1, \dots, d\}$.

$\mathcal{R}_{\sigma,\mu,\nu\Omega}(w) = \mathcal{R}_{\sigma,\nu\Omega}(w) + \mu\|w\|_2^2/2$ *for* $\mathcal{R}_{\sigma,\nu\Omega}$ *defined as in* (19) *by* $\nu > 0$ *and a strongly convex function* Ω *invariant by permutation and such that* $\inf_{\lambda \in \mathcal{P}(\sigma)} \Omega(\lambda) \geq 0$.

If $\hat{w} \in \mathbb{R}^d$ *is a* ε*-accurate minimum of the smoothed regularized objective, i.e.,* $\mathcal{R}_{\sigma,\mu,\nu\Omega}(\hat{w}) - \min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu,\nu\Omega}(w) \leq \varepsilon$ *then it is an* $\varepsilon + \nu\Omega(\sigma)$ *accurate minimum of the original regularized objective* $\mathcal{R}_{\sigma,\mu}$*, where upper-bounds of* $\Omega(\sigma)$ *are provided in Lem. 17 and Cor. 18.*

Proof. Denote $w^* = \arg\min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu}(w)$. If $\hat{w} \in \mathbb{R}^d$ satisfies $\mathcal{R}_{\sigma,\mu,\nu\Omega}(\hat{w}) - \min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu,\nu\Omega}(w) \leq \varepsilon$ then

$$\mathcal{R}_{\sigma,\mu}(\hat{w}) - \mathcal{R}_{\sigma,\mu}(w^*) \leq \mathcal{R}_{\sigma,\mu,\nu\Omega}(\hat{w}) - \mathcal{R}_{\sigma,\mu,\nu\Omega}(w^*) + \nu\Omega(\sigma) \leq \varepsilon + \nu\Omega(\sigma)$$

where we used that $0 \leq \mathcal{R}_{\sigma,\mu}(w) - \mathcal{R}_{\sigma,\mu,\nu\Omega}(w) \leq \nu\Omega(\sigma)$ since Lem. 17 holds. □

D.2 Implementation

The implementation of the smoothing is based on considering the primal formulation of the smoothing given in (18) as an isotonic regression problem and by calling a Pool Adjacent Violators (PAV) algorithm to solve it. It has been described in detail by, e.g., Best et al. (2000); Lim and Wright (2016); Blondel et al. (2020); Henzi et al. (2022). For completeness, we detail here the rationale behind the implementation. We then specify here the overall implementation of the gradient oracles of the smooth approximations for the chosen regularizers Ω_1 and Ω_2 defined in (20) and (21) respectively.

Formulation as Isotonic Regression Problem Consider the primal problem (18) defining the smoothing approximation with a decomposable function Ω such that $\Omega(\lambda) = \sum_{i=1}^n \omega(\lambda_i)$, that is

$$h_{\nu\Omega}(l) = \min_{z \in \mathbb{R}^n} \{h(z) + \nu\Omega^\star((l - z)/\nu)\} = \min_{z \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \left[\sigma_i z_{(i)} + \nu\omega^\star((l_i - z_i)/\nu)\right] \right\}$$

As shown by Blondel et al. (2020, Lemma 4), for any scalars l_i, l_j, z_i, z_j such that $l_i \leq l_j$ and $z_i \geq z_j$, we have, using the convexity of ω^* that $\omega^\star(l_i - z_i) + \omega^\star(l_j - z_j) \geq \omega^\star(l_i - z_j) + \omega^\star(l_j - z_i)$. Hence for $z \in \mathbb{R}^n$ to minimize $\nu\Omega^\star((l - z)/\nu) = \sum_{i=1}^n \nu\omega^\star((l_i - z_i)/\nu)$, the coordinates of z must be ordered in the same order as l. Since $h(z) = \sum_{i=1}^n \sigma_i z_{(i)}$ is independent of the ordering of the coordinates of z, we get that, given a permutation τ of $\{1, \dots, n\}$ such that $l_{\tau_1} \leq \dots \leq l_{\tau_n}$, problem (18) is equivalent to

$$h_{\nu\Omega}(l) = \min_{\substack{z \in \mathbb{R}^n \\ z_{\tau_1} \leq \dots \leq z_{\tau_n}}} \sum_{i=1}^n \left(\sigma_i z_{\tau_i} + \nu\omega^\star((l_i - z_i)/\nu)\right).$$

An oracle on the gradient of the smooth approximation is then given by $\arg\max_{\lambda \in \mathcal{P}(\sigma)} \left\{l^\top \lambda - \nu\Omega(\lambda)\right\} = \nabla h_{\nu\Omega}(l)$ with

$$\nabla h_{\nu\Omega}(l) = \nabla\Omega^\star((l - z^*)/\nu) \quad \text{for } z^* = \nu(\mathrm{PAV}_\omega(l_\tau/\nu))_{\tau^{-1}}, \quad \mathrm{PAV}_\omega(l) = \operatorname*{arg\,min}_{\substack{z \in \mathbb{R}^n \\ z_1 \leq \dots \leq z_n}} \sum_{i=1}^n \left(z_i\sigma_i + \omega^\star(l_i - z_i)\right), \quad (22)$$

where PAV is the output of the Pool Adjacent Violators algorithm (Henzi et al., 2022; Lim and Wright, 2016; Best et al., 2000) applied to the given isotonic regression problem.

Algorithm 4 Pool Adjacent Violators (PAV) Algorithm for ω_2

Inputs: Number of coefficients n, coefficients $(l_i)_{i=1}^n$ and $(\sigma_i)_{i=1}^n$
Initialize $P_1 = \{1\}$, $\mathcal{P} = (P_1)$, $v_1 = l_1 + 1/n - \sigma_1$, $C_1 = 1$, $d = 1$.
for $i = 2, \dots n$ **do**
  Set $P_{d+1} = \{i\}$, $\mathcal{P} \leftarrow (P_1, \dots, P_{d+1})$, $v_{d+1} = l_i + 1/n - \sigma_i$, $C_{d+1} = 1$ and $d = d + 1$,
  while $d \geq 2$ and $v_{d-1} \geq v_d$ **do**
    Set $v_{d-1} \leftarrow \frac{C_{d-1}v_{d-1} + C_d v_d}{C_{d-1} + C_d}$
    Set $C_{d-1} \leftarrow C_d + C_{d-1}$
    Set $\mathcal{P} \leftarrow (P_1, \dots, P_{d-1} \cup P_d)$
    Set $d \leftarrow d - 1$
Output: $z \in \mathbb{R}^n$ such that $z_i = v_s$ for $i \in P_s$, $s \in \{1, \dots, d\}$.

Pool Adjacent Violators Algorithm We briefly recall the rationale of the Pool Adjacent Violator algorithm whose implementation for the choices of ω_1 and ω_2 are given in Alg. 3 and Alg. 4 respectively, where we denote $\mathrm{LSE}(s_S) = \ln \sum_{i \in S} \exp(s_i)$.

The Pool Adjacent Violators Algorithm is used to solve problems of the form

$$\min_{\substack{z \in \mathbb{R}^n \\ z_1 \leq \dots \leq z_n}} \sum_{i=1}^n f_i(z_i) \tag{23}$$

for some set of functions $\mathcal{F} = (f_i)_{i=1}^n$, which in our case (22) are given by $f_i(z_i) = z_i\sigma_i + \omega^\star(l_i - z_i)$.

If at the solution z^*, the constraint $z_i^* \leq z_{i+1}^*$ is active, then by definition, $z_i^* = z_{i+1}^*$. More generally if the constraint $z_i^* \leq z_j^*$ is active for $i < j$, then all constraints of the form $z_k^* \leq z_{k+1}^*$ for $k \in \{i, \dots, j-1\}$ are active, i.e., $z_k^* = z_i^*$ for all $k \in \{i, \dots, j\}$. Overall the solution of (23) is characterized by a set of $p \leq n$ coordinates $v_1^*, \dots, v_p^*$ and a partition $\mathcal{P}^* = (P_1^*, \dots, P_p^*)$ of $\{1, \dots, n\}$ into contiguous blocks $P_s^* = \{b_{s-1} + 1, \dots, b_s\}$ for $0 = b_0 < b_1 < \dots < b_p = n$ such that $z_i^* = v_s^*$ if $i \in \{b_{s-1} + 1, \dots, b_s\}$. For any feasible candidate solution z we can define the corresponding partition $\mathcal{P}(z)$ of $\{1, \dots, n\}$ into contiguous blocks of coordinates. Conversely, given a partition $\mathcal{P} = \{P_1, \dots, P_p\}$ of $\{1, \dots, n\}$ into contiguous blocks, we can define a vector $z(\mathcal{P})$ with constant blocks such that $z_i = \bar{z}_{P_s} = \mathrm{Avg}(\mathcal{F}, P_s)$ for $i \in P_s$ where for a set of functions $\mathcal{F} = (f_i)_{i=1}^n$ and a subset $S \subset \{1, \dots, n\}$, we define the function Avg that computes the average solution of the objective of the PAV algorithm on S, i.e.

$$\mathrm{Avg}(\mathcal{F}, S) = \arg\min_{z \in \mathbb{R}} \sum_{i \in S} f_i(z). \tag{24}$$

The principle of the PAV algorithm is to compute the optimal contiguous partition of $\{1, \dots, n\}$ corresponding to the solution of (23) by adding one coordinate of the problem at a time and merging this coordinate with previously computed blocks if the constraints are not satisfied. We refer to, e.g., (Best et al., 2000; Henzi et al., 2022) for a proof of the validity of this strategy. Most importantly, the efficiency of the PAV algorithm relies on having access to a function, which, for $S, T \subseteq \{1, \dots, n\}$, $S \cap T = \emptyset$, is able to compute $\mathrm{Avg}(\mathcal{F}, S \cup T)$ given appropriate stored values (L_s, M_s in Alg. 3 and v_s, C_s in Alg. 4). The algorithms presented in Alg. 3 and Alg. 4 are then based on the computation of $\mathrm{Avg}(\mathcal{F}, S)$ for the functions f_i considered. Namely, denoting $\mathcal{F}_{\omega,l} = (f_{\omega,l,i})_{i=1}^n$ for $f_{\omega,l,i}(z_i) = z_i\sigma_i + \omega^\star(l_i - z_i)$, $s_i = \ln \sigma_i$ and $\mathrm{LSE}(s_S) = \ln \sum_{i \in S} \exp(s_i)$, we have

$$\mathrm{Avg}(\mathcal{F}_{\omega_1,l}, S) = \mathrm{LSE}(l_S) - \mathrm{LSE}(s_S) - \ln n, \quad \mathrm{Avg}(\mathcal{F}_{\omega_2,l}, S) = \frac{1}{|S|} \sum_{i \in S} (z_i + 1/n - \sigma_i),$$

and merging two subsets of coordinates can be done in $O(1)$ time given appropriate stored values as we have

$$\begin{aligned}
\mathrm{Avg}(\mathcal{F}_{\omega_1,l}, S \cup T) &= \mathrm{LSE}(\mathrm{LSE}(l_S), \mathrm{LSE}(l_T)) - \mathrm{LSE}(\mathrm{LSE}(s_S), \mathrm{LSE}(s_T)) - \ln n \\
\mathrm{Avg}(\mathcal{F}_{\omega_2,l}, S \cup T) &= \frac{|S| \,\mathrm{Avg}(\mathcal{F}_{\omega_2,l}, S) + |T| \,\mathrm{Avg}(\mathcal{F}_{\omega_2,l}, T)}{|S| + |T|}.
\end{aligned}$$

Algorithm 5 LSVRG with smoothing

Require: Number of iterations T, loss functions $(\ell_i)_{i=1}^n$ and their gradient oracles, initial point $w^{(0)}$, regularization parameter μ, learning rate η, sorting update frequency N, probability of checkpointing q^*.

1: **for** iterate $t = 0, ..., T-1$ **do**
2: **if** $t \mod N = 0$ **then** ▷ Update weights (generalization of updating the sorting)
3: Update $\lambda^{(t)} = \arg\max_{\lambda \in \mathcal{P}(\sigma)} \left\{ \sum_{i=1}^n \lambda_i \ell_i(w^{(t)}) - \frac{\nu}{2} \|\lambda - u_n\|_2^2 \right\}$ computed using eq. (22) and Alg. 4.
4: **else**
5: $\lambda^{(t)} = \lambda^{(t-1)}$
6: Sample $q_t \sim \text{Unif}([0,1])$
7: **if** $t \mod N = 0$ or $q_t \le q^*$ **then** ▷ Update batch gradient
8: Set $\bar{w}^{(t)} = w^{(t)}$ and $\bar{g}^{(t)} = \sum_{i=1}^n \lambda_i^{(t)} \nabla \ell_i(\bar{w}^{(t)})$.
9: **else**
10: $\bar{w}^{(t)} = \bar{w}^{(t-1)}$ and $\bar{g}^{(t)} = \bar{g}^{(t-1)}$.
11: Sample $i_t \sim \text{Unif}([n])$.
12: $v^{(t)} = n\lambda_{i_t}^{(t)} \nabla \ell_{i_t}(w^{(t)}) - n\lambda_{i_t}^{(t)} \nabla \ell_{i_t}(\bar{w}^{(t)}) + \bar{g}^{(t)}$.
13: $w^{(t+1)} = (1 - \eta\mu) w^{(t)} - \eta v^{(t)}$.
14: **return** $w^{(T)}$

E LSVRG CONVERGENCE ANALYSIS

E.1 Setup for the Convergence Analysis

Consider the optimization problem

$$\min_w \left[\mathcal{R}_{\sigma,\mu,\nu}(w) := h_\nu\big(\ell(w)\big) + \frac{\mu}{2} \|w\|_2^2 \right], \quad \text{where } h_\nu(l) = \max_{\lambda \in \mathcal{P}(\sigma)} \left\{ \lambda^\top l - \frac{\nu}{2} \|\lambda - u_n\|_2^2 \right\} \tag{25}$$

is the L_2-smoothing as defined in Appx. D where $\Omega(\lambda) = \|\lambda - u_n\|^2 / 2$. Here, $\sigma_1 \le \cdots \le \sigma_n$ are given nonnegative weights that sum to 1, $\mathcal{P}(\sigma)$ is the permutahedron of σ, μ is a regularization parameter on the w's, ν is smoothing parameter and $u_n = \mathbf{1}_n / n$ denotes the uniform distribution over n items.

It is convenient to look at the saddle form

$$\Phi_\nu(w, \lambda) := \lambda^\top \ell(w) + \frac{\mu}{2} \|w\|_2^2 - \frac{\nu}{2} \|\lambda - u_n\|_2^2 . \tag{26}$$

Throughout, we make the following assumption:

Assumption 20. *For each $i \in [n]$, $w \mapsto \ell_i(w)$ is convex, G-Lipschitz, and L-smooth.*

We analyze LSVRG with smoothing, as given in Algorithm 5. It only differs from Algorithm 2 presented in the main paper in line 3.

E.2 Convergence Analysis

Algorithm 5 can be interpreted as an algorithm that alternates exactly maximizing over λ in $\Phi_\nu(w, \cdot)$ with w fixed and minimizing $\Phi_\nu(\cdot, \lambda)$ with λ fixed using a particular variant of SVRG known as q-SVRG (Hofmann et al., 2015); see Algorithm 8 for a review of q-SVRG.

Proposition 21. *The iterates $(w_1^{(t)}, \lambda_1^{(t)})$ produced by Algorithm 5 and $(w_2^{(k)}, \lambda_2^{(k)})$ produced by Algorithm 6 with a given starting point $w^{(0)}$, learning rate η, weight update frequency (or inner loop length) N, and number of iterates $T = KN$ where K is the number of epochs of Algorithm 6 satisfy $w_2^{(k)} = w_1^{(kN)}$ and $\lambda_2^{(k)} = \lambda_1^{(kN)}$ for each epoch k.*

Proof. The two algorithms are equivalent iteration for iteration and the proof follows from pattern matching. □

Convergence Analysis We have the following rate when the smoothing parameter $\nu > O(nG^2/\mu)$.

Algorithm 6 LSVRG with smoothing: Rewriting

Require: Number of epochs K, number of SVRG steps N, loss functions $(\ell_i)_{i=1}^n$ and their gradient oracles, initial point $w^{(0)}$, regularization parameter μ, learning rate η, probability of checkpointing q^*, smoothing coefficient ν.

1: **for** epoch $k = 0, ..., K-1$ **do**
2: Compute $\lambda^{(k)} = \arg\max_{\lambda \in \mathcal{P}(\sigma)} \Phi_\nu(w^{(k)}, \lambda)$ using eq. (22) and Alg. 4.
3: Define $\tilde{\ell}_i^{(k)}(w) := n\lambda_i^{(k)}\ell_i(w) + \mu \|w\|_2^2/2$ for $i \in \{1, \dots, n\}$.
4: Compute $w^{(k+1)} = \text{q-SVRG}\left(N, (\tilde{\ell}_i^{(k)})_{i=1}^n, w^{(k)}, \eta, q^*\right)$ using Algorithm 8.
5: **return** $w^{(K)}$

Theorem 1. *Consider problem (25) satisfying Asm. 20. Suppose the smoothing parameter satisfies $\nu \geq 4nG^2/\mu$. The sequence of iterates produced by Algorithm 6 with inputs $N = (n(1 + 8\sigma_n L/\mu) + 8)\log(125/4)$, $\eta = 2/(n(8\sigma_{\max}L + \mu) + 8\mu)$, $q^* = 1/n$, satisfies*

$$\mathbb{E}\|w^{(k)} - w^*\|_2 \leq \left(\frac{1}{2}\right)^k \|w^{(0)} - w^*\|_2 ,$$

where $w^* = \arg\min_{w \in \mathbb{R}^d} \mathcal{R}_{\sigma,\mu,\nu}(w)$.

Consequently, Algorithm 6 (and hence Algorithm 5) can produce a point $\hat{w}$ satisfying $\left(\mathbb{E}\|\hat{w} - w^*\|_2\right)^2 \leq \epsilon$ in

$$T \leq C(n(1 + 8\sigma_n L/\mu) + 8)\log\left(\|w^{(0)} - w^*\|^2/\epsilon\right)$$

gradient evaluations, where C is an absolute constant.

Proof. For each epoch k, Algorithm 6 runs q-SVRG on the function

$$\varphi^{(k)}(w) := \Phi_\nu(w, \lambda^{(k)}) = \frac{1}{n}\sum_{i=1}^n \tilde{\ell}_i^{(k)}(w) \quad \text{where} \quad \tilde{\ell}_i^{(k)}(w) = n\lambda_i^{(k)}\ell_i(w) + \frac{\mu}{2}\|w\|^2 .$$

The aim of this step is to approximate $w_*^{(k+1)} = \arg\min_{w \in \mathbb{R}^d} \Phi_\nu(w, \lambda^{(k)})$ with $w^{(k+1)}$. We start by quantifying this error. Since $\mathcal{P}(\sigma)$ is the permutahedron on σ, we have that

$$\sigma_1 \leq \min\{\lambda_i \,:\, \lambda \in \mathcal{P}(\sigma)\} \leq \max\{\lambda_i \,:\, \lambda \in \mathcal{P}(\sigma)\} \leq \sigma_n.$$

Hence, we have that each $\tilde{\ell}_i^{(k)} = n\lambda_i^{(k)}\ell_i + \mu\|\cdot\|_2^2/2$ is $n\sigma_n L + \mu$-smooth and μ-strongly convex, and its condition number is $\kappa = (n\sigma_n L/\mu + 1)$. Denote the sigma-algebra generated by $w^{(k)}$ as $\mathcal{F}_k$, we have from Thm. 2 that

$$\mathbb{E}\left[\left\|w^{(k+1)} - w_*^{(k+1)}\right\|^2 \middle| \mathcal{F}_k\right] \leq \frac{5}{4}\exp\left(-\frac{N}{8\kappa + n}\right)\left\|w^{(k)} - w_*^{(k+1)}\right\|^2 = \frac{1}{25}\left\|w^{(k)} - w_*^{(k+1)}\right\|^2 .$$

Therefore, Jensen's inequality gives us

$$\mathbb{E}\left[\left\|w^{(k+1)} - w_*^{(k+1)}\right\| \middle| \mathcal{F}_k\right] \leq \frac{1}{5}\left\|w^{(k)} - w_*^{(k+1)}\right\| . \tag{27}$$

Denote $\lambda^* = \arg\max_{\lambda \in \mathcal{P}(\sigma)} \Phi_\nu(w^*, \lambda)$. Since $\Phi_\nu(\cdot, \lambda)$ is strongly convex and $\Phi_\nu(w, \cdot)$ is strongly concave, we have that strong duality holds, i.e., $\min_{w \in \mathbb{R}^d}\max_{\lambda \in \mathcal{P}(\sigma)} \Phi_\nu(w, \lambda) = \max_{\lambda \in \mathcal{P}(\sigma)}\min_{w \in \mathbb{R}^d} \Phi_\nu(w, \lambda)$ (e.g., Hiriart-Urruty and Lemaréchal, 1993, Thm. VII.4.3.1) Therefore, (w^*, λ^*) is the unique saddle point of Φ_ν, so $w^* = \arg\min_{w \in \mathbb{R}^d} \Phi_\nu(w, \lambda^*)$. Together with Lem. 4, this gives us

$$\left\|w_*^{(k+1)} - w^*\right\| \leq \frac{\sqrt{n}G}{\mu}\left\|\lambda^{(k)} - \lambda^*\right\|, \quad \text{and} \quad \left\|\lambda^{(k)} - \lambda^*\right\| \leq \frac{\sqrt{n}G}{\nu}\left\|w^{(k)} - w^*\right\| . \tag{28}$$

From repeated invocations of the triangle inequality, we get,

$$\begin{aligned}
\mathbb{E}\left[\left\|w^{(k+1)}-w^*\right\|\middle|\mathcal{F}_k\right] &\leq \mathbb{E}\left[\left\|w^{(k+1)}-w_*^{(k+1)}\right\|\middle|\mathcal{F}_k\right]+\left\|w_*^{(k+1)}-w^*\right\| \\
&\overset{(27)}{\leq} \frac{1}{5}\left\|w^{(k)}-w_*^{(k+1)}\right\|_2+\left\|w_*^{(k+1)}-w^*\right\|_2 \\
&\leq \frac{1}{5}\left\|w^{(k)}-w^*\right\|_2+\frac{6}{5}\left\|w_*^{(k+1)}-w^*\right\|_2 \\
&\overset{(28)}{\leq} \frac{1}{5}\left\|w^{(k)}-w^*\right\|_2+\frac{6\sqrt{n}G}{5\mu}\left\|\lambda^{(k)}-\lambda^*\right\|_2 \\
&\overset{(28)}{\leq} \left(\frac{1}{5}+\frac{6nG^2}{5\mu\nu}\right)\left\|w^{(k)}-w^*\right\|_2 \\
&\leq \frac{1}{2}\left\|w^{(k)}-w^*\right\|,
\end{aligned}$$

since we assumed ν satisfies $6nG^2/(5\mu\nu) \leq 3/10$. Taking an expecation w.r.t. $\mathcal{F}_k$ and unrolling this completes the proof. □

E.3 LSVRG Variants

Algorithm 7 gives a variant of the LSVRG algorithm that computes checkpoints and the sorting at regular intervals. For simplicity, we visualize this algorithm as running in epochs. As in the usual SVRG algorithm for the ERM setting, we compute the full-batch subgradient at the checkpoint $\bar{w}_k$ at the start of each epoch (line 3). This is used to define the variance-reduced update in line 7. Note also that we consider sampling at each iteration an example i_t distributed as $p_\sigma(i) = \mathbb{P}[i_t = i] = \sigma_i$; this is well-defined since $\sigma_1, ..., \sigma_n$ defines a probability measure over $\{1, \cdots, n\}$.

Algorithm 7 Epoch-based LSVRG with nonuniform sampling

Require: Number of iterates T per epoch, number of epochs K, regularization parameter μ, learning rate η, non-decreasing probability mass function $\sigma = (\sigma_i)_{i=1}^n$, loss functions $(\ell_i)_{i=1}^n$ and their gradient oracles, initial point $\bar{w}_0$.

1: **for** epoch $k = 0, 1, 2, ..., K-1$ **do**
2: Select $\pi_k \in \operatorname{argsort}(\ell(\bar{w}_k))$.
3: $\bar{g}_k = \sum_{i=1}^n \sigma_i \nabla \ell_{\pi_k(i)}(\bar{w}_k)$.
4: $w^{(0)} = \bar{w}_k$.
5: **for** iterate $t = 0, ..., T-1$ **do**
6: Sample $i_t \sim p_\sigma$.
7: $v^{(t)} = \nabla \ell_{\pi_k(i_t)}(w^{(t)}) - \nabla \ell_{\pi_k(i_t)}(\bar{w}_k) + \bar{g}_k$.
8: $w^{(t+1)} = (1-\eta\mu)w^{(t)} - \eta v^{(t)}$.
9: Set $\bar{w}_{k+1} = w^{(T)}$.
10: **return** $\bar{w}_K$.

E.4 q-SVRG Review

Consider the risk-neutral problem

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

The q-SVRG is a variant of SVRG that updates the batch gradient with probability $1/m$ at each step, rather than once every m steps like the usual version of SVRG (Hofmann et al., 2015). See Algorithm 8 for details. It has the following convergence guarantee.

Theorem 2 (Hofmann et al., 2015, Lemma 3)**.** *Suppose each ℓ_i is L-smooth and μ-strongly convex. Then Algorithm 8 with a learning rate $\eta = 2/(8L + n\mu)$ and $q^* = 1/n$ produces a sequence $(w^{(t)})$ that satisfies*

$$\mathbb{E}\left\|w^{(t)}-w^*\right\|^2 \leq \frac{5}{4}\exp\left(-\frac{t}{8\kappa+n}\right)\left\|w^{(0)}-w^*\right\|^2,$$

where $w^* = \arg\min_w f(w)$ and $\kappa = L/\mu$ is the condition number.

Algorithm 8 q-SVRG

Require: Number of iterations T, loss functions $(\ell_i)_{i=1}^n$ and their gradient oracles, initial point $w^{(0)}$, learning rate η, probability of checkpointing q^*.

Set $\bar{w}^{(-1)} = w^{(0)}$ and $\bar{g}^{(-1)} = \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(w^{(0)})$
for iterate $t = 0, ..., T-1$ **do**
  Draw $q_t \sim \text{Unif}([0,1])$
  if $q_t \leq q^*$ **then** ▷ Update the batch gradient
    Set $\bar{w}^{(t)} = w^{(t)}$ and $\bar{g}^{(t)} = \frac{1}{n}\sum_{i=1}^n \nabla \ell_i(w^{(t)})$
  else
    $\bar{w}^{(t)} = \bar{w}^{(t-1)}$ and $\bar{g}^{(t)} = \bar{g}^{(t-1)}$
  Sample $i_t \sim \text{Unif}([n])$
  $v^{(t)} = \nabla \ell_{i_t}(w^{(t)}) - \nabla \ell_{i_t}(\bar{w}^{(t)}) + \bar{g}^{(t)}$
  $w^{(t+1)} = w^{(t)} - \eta v^{(t)}$
return $w^{(T)}$

E.5 Technical Results

Note the following properties of the joint function Φ_ν defined in (26).

Property 3. *The following smoothness properties hold:*

(a) For each $\lambda \in \mathcal{P}(\sigma)$, $\nabla_w \Phi_\nu(\cdot, \lambda)$ is $(L+\mu)$-Lipschitz
(b) For each $w \in \mathbb{R}^d$, $\nabla_\lambda \Phi_\nu(w, \cdot)$ is ν-Lipschitz.
(c) For each $w \in \mathbb{R}^d$, $\nabla_w \Phi_\nu(w, \cdot)$ is $\sqrt{n}G$-Lipschitz.
(d) For each $\lambda \in \mathcal{P}(\sigma)$, $\nabla_\lambda \Phi_\nu(\cdot, \lambda)$ is $\sqrt{n}G$-Lipschitz.

Proof. The result follows from the expressions

$$\nabla_w \Phi_\nu(w, \lambda) = \sum_{i=1}^n \lambda_i \nabla \ell_i(w) + \mu w \quad \text{and} \quad \nabla_\lambda \Phi_\nu(w, \lambda) = \ell(w) - \nu(\lambda - u_n)\,.$$

(a) For any $w, w' \in \mathbb{R}^d$,

$$\begin{aligned}
\|\nabla_w \Phi_\nu(w, \lambda) - \nabla_w \Phi_\nu(w', \lambda)\|_2 &\leq \sum_{i=1}^n \lambda_i \|\nabla \ell_i(w) - \nabla \ell_i(w')\|_2 + \mu \|w - w'\|_2 \\
&\leq \sum_{i=1}^n \lambda_i L \|w - w'\|_2 + \mu \|w - w'\|_2 \\
&\leq (L + \mu) \|w - w'\|_2\,,
\end{aligned}$$

as $\sum_{i=1}^n \lambda_i = 1$ for $\lambda \in \mathcal{P}(\sigma)$.

(b) For any $\lambda, \lambda' \in \mathcal{P}(\sigma)$,

$$\|\nabla_\lambda \Phi_\nu(w, \lambda) - \nabla_\lambda \Phi_\nu(w, \lambda')\|_2 = \|\nu\lambda - \nu\lambda'\|_2 = \nu \|\lambda - \lambda'\|_2\,.$$

(c) For any $\lambda, \lambda' \in \mathcal{P}(\sigma)$,

$$\begin{aligned}
\|\nabla_w \Phi_\nu(w, \lambda) - \nabla_w \Phi_\nu(w, \lambda')\|_2^2 &= \left\| \sum_{i=1}^n (\lambda_i - \lambda_i') \nabla \ell_i(w) \right\|_2^2 \\
&\leq \sum_{i=1}^n \|\nabla \ell_i(w)\|_2^2 \sum_{i=1}^n (\lambda_i - \lambda_i')^2 \\
&\leq nG^2 \|\lambda - \lambda'\|_2^2\,.
\end{aligned}$$

(d) For any $w, w' \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla_\lambda \Phi_\nu(w,\lambda) - \nabla_\lambda \Phi_\nu(w,\lambda')\|_2^2 &= \|\ell(w) - \ell(w')\|_2^2 \\ &= \sum_{i=1}^n (\ell_i(w) - \ell_i(w'))_2^2 \\ &\le \sum_{i=1}^n G\,\|w - w'\|_2^2 \\ &= nG\,\|w - w'\|_2^2 .\end{aligned}$$

□

Lemma 4. *Given closed, convex sets $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}^p$, consider a continuously differentiable function $f : X \times Y \to \mathbb{R}$ such that $f(\cdot, y)$ is μ-strongly convex for all $y \in Y$ and $\nabla_x f(x, \cdot)$ is $L_{x,y}$-Lipschitz for each $x \in X$. Then, the map $x^*(y) = \arg\min_{x \in X} f(x, y)$ is well-defined and is $L_{x,y}/\mu$ Lipschitz.*

Proof. The map $x^*(y)$ is well-defined because $f(\cdot, y)$ is strongly convex and X is closed, convex. Consider two points $y_1, y_2 \in Y$ and let $x_i = x^*(y_i)$ be the corresponding x-values. From the first order optimality conditions of $f(\cdot, y_1)$ and $f(\cdot, y_2)$ respectively, we have

$$\langle \nabla_x f(x_1, y_1), x_2 - x_1 \rangle \ge 0, \quad \text{and} \quad \langle \nabla_x f(x_2, y_2), x_1 - x_2 \rangle \ge 0\,. \tag{29}$$

Using the co-coercivity property $(*)$ of the strong convexity of $f(\cdot, y)$, we have,

$$\begin{aligned}\mu\,\|x_1 - x_2\|^2 &\overset{(*)}{\le} \langle \nabla_x f(x_1, y_2) - \nabla_x f(x_2, y_2), x_1 - x_2 \rangle \\ &\overset{(29)}{\le} \langle \nabla_x f(x_1, y_2), x_1 - x_2 \rangle \\ &\overset{(29)}{\le} \langle \nabla_x f(x_1, y_2) - \nabla_x f(x_1, y_1), x_1 - x_2 \rangle \\ &\le \|\nabla_x f(x_1, y_2) - \nabla_x f(x_1, y_1)\|\ \|x_1 - x_2\| \\ &\le L_{x,y}\,\|y_1 - y_2\|\ \|x_1 - x_2\|\ .\end{aligned}$$

□

F EXPERIMENTAL DETAILS

Appx. F.1 describes the tasks, datasets, and preprocessing steps used in the experiments. Appx. F.2 reviews the objective minimized (including regularization). Appx. F.3 describes the baseline methods compared. Appx. F.4 lists the hyperparameters of each algorithm and describes how they are selected. Appx. F.5 describes the compute environment used to run the experiments.

F.1 Task and Dataset Descriptions

We start by describing the tasks and datasets considered in the experiments as well as their preprocessing steps. For each task, we consider an input $x \in \mathcal{X}$, a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$, and an output space $\mathcal{Y}$. For regression, we have $\mathcal{Y} = \mathbb{R}$ and for classification, we have $\mathcal{Y} = \{1, \ldots, C\}$, where C is the number of classes. We make predictions with a linear model $x \mapsto w^\top \phi(x)$, where $w \in \mathbb{R}^d$ is the parameter vector to be optimized over. We consider the square loss between these predictions and the target y_i:

$$\ell_i(w) = \frac{1}{2}(y_i - w^\top \phi(x_i))^2\,.$$

for regression, and the multinomial logistic loss

$$\ell_i(w) = -\log p_{y_i}(x_i; w), \text{ where } p_{y_i}(x_i; w) := \frac{\exp\left(w_{\cdot y}^\top x_i\right)}{\sum_{y'=1}^C \exp\left(w_{\cdot y'}^\top x_i\right)},\ w \in \mathbb{R}^{d \times C}$$

Dataset	d	n_{train}	n_{test}	Source
`simulated`	10	800	200	n/a
`yacht`	6	244	62	UCI
`energy`	8	614	154	UCI
`concrete`	8	824	206	UCI
`iWildCam`	157	20,000	5,000	WILDS
`emotion`	45	8,000	2000	DAIR-AI

Table 2: Benchmark dataset descriptions.

for classification. Each input feature $\phi_j(x)$ for $j = 1, \cdots, d$ is standardized to zero mean and unit variance (as are the targets y_i in regression). We now describe the datasets considered. The size and dimensionality of the resulting datasets are summarized in Tab. 2.

(a) `simulated`: This regression task entails prediction of a synthetic, real-valued response based on d-dimensional real vectors. The dataset is generated by sampling the inputs $x_1, ..., x_n$ and true parameter vector w^* from the d-dimensional standard normal distribution $\mathcal{N}(0, I_d)$ for $n = 1000$ and $d = 10$, and the noise $\epsilon_1, ..., \epsilon_n \in \mathcal{N}(0, 1)$. Then, $y_i = w^\top x_i + \epsilon_i$ for $i = 1, ..., n$. The feature map ϕ is taken to be the identity.

(b) `yacht`: This regression task entails prediction of the residuary resistance of a sailing yacht based on its physical attributes (Tsanas and Xifara, 2012). Each input $x \in \mathcal{X}$ is a sailing yacht and the feature map $\phi(x) \in \mathbb{R}^d$ lists $d = 6$ geometric attributes such as the length-beam ratio.

(c) `energy`: This regression task entails prediction of the cooling load of a building based on its physical attributes (Baressi Segota et al., 2020). Each input $x \in \mathcal{X}$ is a building and the feature map $\phi(x) \in \mathbb{R}^d$ lists $d = 8$ structural attributes such as the surface area, height, etc.

(d) `concrete`: This regression task entails prediction of the compressive strength of a concrete type based on its physical and chemical attributes (Yeh, 2006). Each input $x \in \mathcal{X}$ is a particular composition of concrete and the feature map $\phi(x) \in \mathbb{R}^d$ lists $d = 8$ physical/chemical attributes such as amount of cement vs water.

(e) `iWildCam`: This classification task entails prediction of an animal present in an image captured by various wilderness camera traps, with drastic variation in illumination, camera angle, background, vegetation, color, and relative animal frequencies (Beery et al., 2020). Each input $x \in \mathcal{X}$ is an image the feature map $\phi(x) \in \mathbb{R}^d$ for $d = 189$ is the output of the sequence of the following operations.

- A ResNet50 neural network (He et al., 2016) that is pretrained on ImageNet (Deng et al., 2009) is applied to the image x_i, resulting in vector x_i'.
- The $x_1', \ldots, x_n'$ are normalized to have unit norm.
- Principle Components Analysis (PCA) is applied, resulting in $d = 157$ components that explain 99% of the variance, resulting in vectors $x_i'' \in \mathbb{R}^{157}$.
- The $x_1'', \ldots, x_n''$ are standardized once again, giving $\phi(x_1), \ldots, \phi(x_n)$.

(f) `emotion`: This classification task entails prediction of the emotional content of a sentence taken from English Twitter archives (Saravia et al., 2018). Each input $x \in \mathcal{X}$ is an image the feature map $\phi(x) \in \mathbb{R}^d$ for $d = 189$ is the output of the sequence of the following operations.

- A BERT neural network (Devlin et al., 2019) (fine-tuned on $8,000$ held-out examples) is applied to the text x_i, resulting in vector x_i'.
- The $x_1', \ldots, x_n'$ are standardized to have unit norm.
- Principle Components Analysis (PCA) is applied, resulting in $d = 45$ components that explain 99% of the variance, resulting in vectors $x_i'' \in \mathbb{R}^{45}$.
- The $x_1'', \ldots, x_n''$ are standardized once again, giving $\phi(x_1), \ldots, \phi(x_n)$.

F.2 Objective

In the experiments, we consider minimizing regularized ordered risk minimization problems of the form

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & \mathcal{R}_\sigma(w) + \frac{\mu}{2}\|w\|_2^2, \\ \text{where} \quad & \mathcal{R}_\sigma(w) = \sum_{i=1}^n \sigma_i \ell_{(i)}(w), \end{aligned}$$

where the coefficients σ are defined using the spectrum of the spectral risk measure in question. We consider the mean, superquantile, extremile, and exponential spectral risk measure (ESRM), as defined in Sec. 2. The regularization parameter μ is chosen as $1/n$ in the experiments presented in the main text, whereas other choices of μ are shown in Appx. G. By adding a regularization $\|\cdot\|_2^2$ to the objective, the LSVRG algorithm is modified by considering a direction of the form $v_{\text{reg}}^{(t)} = v_{\text{non\_reg}}^{(t)} + \mu w^{(t)}$, where $v_{\text{non\_reg}}^{(t)} = \bar{g}^{(t)}$ is the direction presented in line 10 of Algorithm 7. All algorithms are initialized with $w^{(0)} = 0$.

F.3 Baseline Methods

The baseline methods described below rely on a *stochastic subgradient estimate*, or a random quantity $g^{(t)}$ that estimates $\nabla \mathcal{R}_\sigma(w^{(t)})$ if $\mathcal{R}_\sigma$ is differentiable at $w^{(t)}$ and a subgradient of $\partial \mathcal{R}_\sigma(w^{(t)})$ otherwise. As described in Sec. 3, we use

$$g^{(t)} := \sum_{j=1}^m \hat{\sigma}_j \nabla \ell_{i_{(j)}}(w^{(t)}) \tag{30}$$

for a minibatch $\{i_1, ..., i_m\}$ of size m with weights $\hat{\sigma}_j = \int_{\frac{j-1}{m}}^{\frac{j}{m}} s(t)\,\mathrm{d}t$, and $\ell_{i_{(j)}}$ being the ordered losses $\ell_{i_{(1)}} \leq \ldots \ell_{i_{(m)}}$ in the minibatch. We refer to the direction as $g^{(t)} = v_m^{(t)}$ in Algorithm 1.

SGD We refer to the stochastic subgradient method as SGD. The update can be written as

$$w^{(t+1)} := w^{(t)} - \eta(g^{(t)} + \mu w^{(t)}),$$

where $v_m^{(t)}$ is a stochastic estimate of the minibatch extremile subgradient (Equation (30)).

SRDA The stochastic regularized dual averaging (SRDA) (Xiao, 2009) update can be written as

$$w^{(t+1)} := \arg\min_{w \in \mathbb{R}^d} w^\top \bar{g}^{(t)} + \frac{\mu}{2}\|w\|_2^2 + \frac{1}{2\eta t}\|w\|_2^2,$$

where $\bar{g}^{(t)} = \sum_{i=0}^t g^{(i)}$ is the average of all stochastic subgradients (again computed by Equation (30)). Note that for $\Omega = \|\cdot\|_2^2/2$ and $w^{(0)} = 0$, Note that for $w^{(0)} = 0$,

$$\begin{aligned} w^{(t+1)} &= 0 - \frac{1}{\mu + 1/t\eta}\bar{g}^{(t)} \\ &= w^{(0)} - \sum_{s=0}^t \frac{1}{\mu t + 1/\eta} g^{(s)}. \end{aligned}$$

Thus, the SRDA solution at time $t+1$ can be seen as applying SGD with a *constant* learning rate of $\eta = 1/(\mu t/n + \beta)$ (as t refers to the value of only the last iteration). It is also seen that when $\mu = 0$ (no statistical regularization), SRDA reduces exactly to SGD.

F.4 Hyperparameter Selection

The fixed optimization hyperparameters include the minibatch size $m = 64$ (SGD, SRDA) and the epoch length $N = n$ (LSVRG). The statistical regularization parameter $\mu = 1/n$ is shown in the main text, whereas training curves for $\mu = 0.1/n$ and $\mu = 10/n$ are shown in Appx. G. Specifically, $c \in \{1, 2, 3, 4, 5\}$ be a seed that determines the randomness

for sampling the minibatch $\{i_1, ..., i_m\}$ at each iteration of SGD and SRDA, i_t at each iteration of LSVRG. Let T be the total number of iterations for the algorithm, and denote the trajectory of iterates seeded by c using learning rate η as $w_{c,\eta}^{(1)}, ..., w_{c,\eta}^{(T)}$. Then, define the quantity $L(\eta) = \frac{1}{5}\sum_{c=1}^{5} \mathcal{R}_\sigma(w_{c,\eta}^{(T)})$. The learning rate η is chosen in the set $\{3\times10^{-4}, 1\times10^{-3}, 3\times10^{-3}, 1\times10^{-2}, 3\times10^{-2}, 1\times10^{-1}, 3\times10^{-1}, 1\times10^{0}, 3\times10^{0}\}$ to minimize $L(\eta)$ for each algorithm. If any of the trajectories diverge, we consider $L(\eta) = +\infty$. Note that $\mathcal{R}_\sigma$ is computed using the *training set*, as we are selecting hyperparameters for optimization.

F.5 Compute Environment

All experiments were run on a workstation with Intel i9 processor (clock speed: 2.80GHz) with 32 virtual cores and 126G of memory. We did not use GPUs for any experiments. Code used for this project was written in Python 3.

F.6 Experimental Details on Clustering

Recall that we consider clustering n points $x_1, \dots, x_n$ into k clusters with centers $C = (c_1, \dots, c_k)$ by minimizing a weighted average of the distances of each point to its closest center, i.e., problems of the form

$$\min_{C\in\mathbb{R}^{d\times k}} \sum_{i=1}^{n} \sigma_i \ell_{(i)}(C) \quad \text{for } \ell_i(C) = \min_{\substack{z_i\in\{0,1\}^k \\ z_i^\top \mathbf{1}=\mathbf{1}}} \sum_{j=1}^{k} z_{ij}\|x_i - c_j\|_2^2.$$

We consider the weights σ_i to be the discretization of a spectrum s such that $\sigma_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} s(t)dt$ with s being one of the following examples:

1. uniform spectrum, $s(t) = \mathbf{1}_{[0,1]}(t)$ which corresponds to a classical kmeans objective of the form $\min_{C\in\mathbb{R}^{d\times k}} \frac{1}{n}\sum_{i=1}^{n} \ell_i(C)$,
2. a truncated spectrum, $s_q(t) = \mathbf{1}_{[0,q]}(t)/q$ for $q \in (0,1)$ that seeks to only consider minimizing losses with small enough values compared to the whole distribution,
3. an extremile spectrum $s_r(t) = r(1-t)^r$ for $r \geq 1$ that can be interpreted as minimizing the expected minimum of r random variables distributed as the losses (Daouia et al., 2019).

We consider a stochastic subradient descent with constant stepsize with mini-batch estimates given by the empirical L-statitics estimate on the mini-batches as described in Sec. 3.

F.6.1 Synthetic Data

As (Maurer et al., 2021) we consider as training data three cloud of Gaussians composed of 100 two dimensional points each with variance 0.1 along both axis and centers $(-3, 0), (0, 1)$ and $(3, 0)$ respectively. We add 100 outliers sampled from a Gaussian with variance 5 along both axis and center $(-1, -5)$. The test set consists in points sampled from the three aforementioned inlier Gaussians, 100 points per Gaussian. To test our method, we compute the number of correct assignments of the test points in their associated cluster after relabeling the clusters to match the true labeling. Namely, the groups found by a method may be correct but instead of labeling the first cloud of points by 1 the method may have assigned the label 1 to the second group and 2 to the first group for example, so we first find the permutation of the labels that leads to the highest accuracy.

We used mini-batches of size 64, a learning rate of 1 found by grid-search on log-10 scale, a uniform spectrum, a truncated spectrum with parameter $q = 0.75$ or an extremile spectrum with $r = 5$ and we initialize the centers at 0. In Fig. 7 we present the estimated centers found for each spectrum as well as the training and test losses and the training and test accuracies, where for the training accuracy we only consider the assignment of the inlier points.

F.6.2 Clustering Digits Images

We consider forming a subset of the MNIST dataset (LeCun et al., 1998) of 28×28 black and white images of handwritten digits by selecting 1000 images of the digit 1, 1000 images of the digit 3 each and 125 images of each other digit in $\{0, \dots, 9\} \setminus \{1, 3\}$ for a total of 2000 inlier examples and 1000 outlier examples. The images are standardized pixel by pixel. Our goal is to cluster the samples from 1 and 3 correctly even in the presence of outliers. We test our estimated centers on all images of the digits 1 and 3 from the test set of the MNIST database, that is, as in the synthetic experiment we test whether our estimated centers lead to the correct assignments of the test images in their respective group.

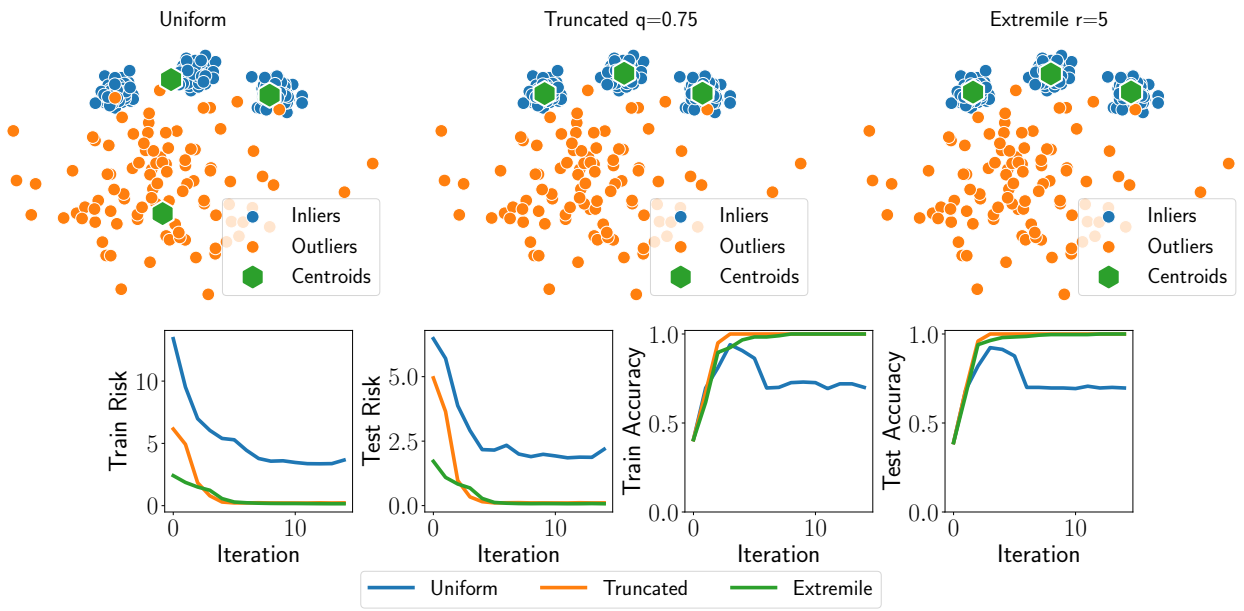

Figure 7: Clustering synthetic data points in the presence of outliers.

We consider mini-batches of size 256, a learning rate of 0.1 found by grid-search on a log-10 scale, a uniform spectrum, a truncated spectrum with parameter $q = 0.66$ or an extremile spectrum with $r = 2$ and we initialize the centers at 0. In Fig. 8 we present the estimated centers found for each spectrum as well as the training and test losses and the training and test accuracies, where for the training accuracy we only consider the assignment of the inlier points.

F.6.3 Clustering Images of Clothes

As Maurer et al. (2021) we also consider clustering images of clothes from the dataset FashionMNIST (Xiao et al., 2017) that consist in 28×28 black and white images of 10 clases of clothing such as: t-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot. The images are standardized pixel by pixel. We form a training set composed of 1000 images of trousers, 1000 images of sneakers, and 250 images of each of the other classes for a total of 2000 inliers and 2000 outliers. Our goal is to cluster teh trousers and the sneakers in the presence of the outliers. To test our estimators we use all images of trousers and sneakers from the test set of the FashionMNIST dataset.

We consider mini-batches of size 64, a learning rate of 1. found by grid-search on a log-10 scale, a uniform spectrum, a truncated spectrum with parameter $q = 0.5$ or an extremile spectrum with $r = 5$ and we initialize the centers at 0. In Fig. 9 we present the estimated centers found for each spectrum as well as the training and test losses and the training and test accuracies, where for the training accuracy we only consider the assignment of the inlier points.

Note that compared to Maurer et al. (2021) we obtain 100% accuracy of these methods on the test set. An approach by stochastic subgradient may be less sensitive to the initialization (performed with K-means++ by Maurer et al. (2021)).

G ADDITIONAL EXPERIMENTS

Optimization Effect of Varying Regularization Parameter We demonstrate the robustness of the algorithm comparison with respect to the *statistical regularization parameter* μ. Hyperparameters are selected in accordance with Appx. F.4. Fig. 10, Fig. 11, and Fig. 12 show the suboptimality trajectories for $\mu = 1/n$, $10/n$, and $0.1/n$, respectively. The same rankings of algorithms result from each of the three figures, that LSVRG generally outperforms SGD and SRDA.

Optimization Effect of Varying Risk Parameter We demonstrate the robustness of the algorithm comparison with respect to the statistical regularization parameter μ. Hyperparameters are selected in accordance with Appx. F.4. Fig. 13, Fig. 14, and Fig. 15 show the suboptimality trajectories for (q, r, ρ) set to $(0.25, 1.5, 0.5)$, $(0.5, 2, 1)$, and $(0.75, 2.5, 2)$, respectively. The same rankings of algorithms result from each of the three figures, that LSVRG generally outperforms SGD and SRDA. It should be noted that for the 0.75-superquantile, LSVRG suffers from slow convergence and is outperformed by SGD and

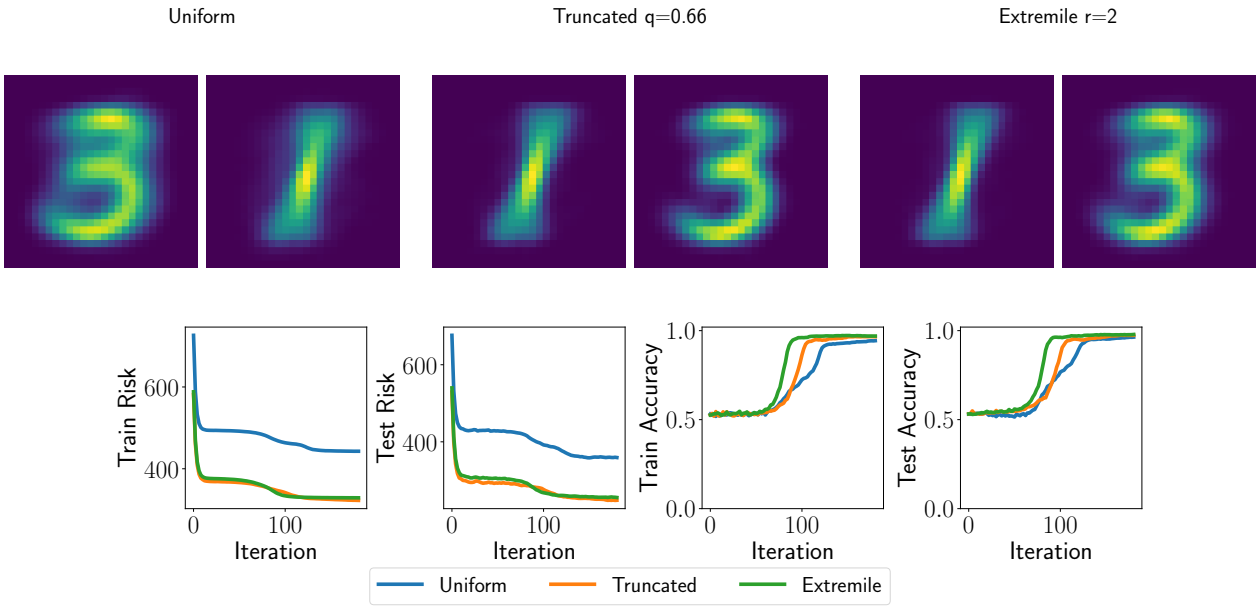

Figure 8: Clustering images of digits in the presence of outliers.

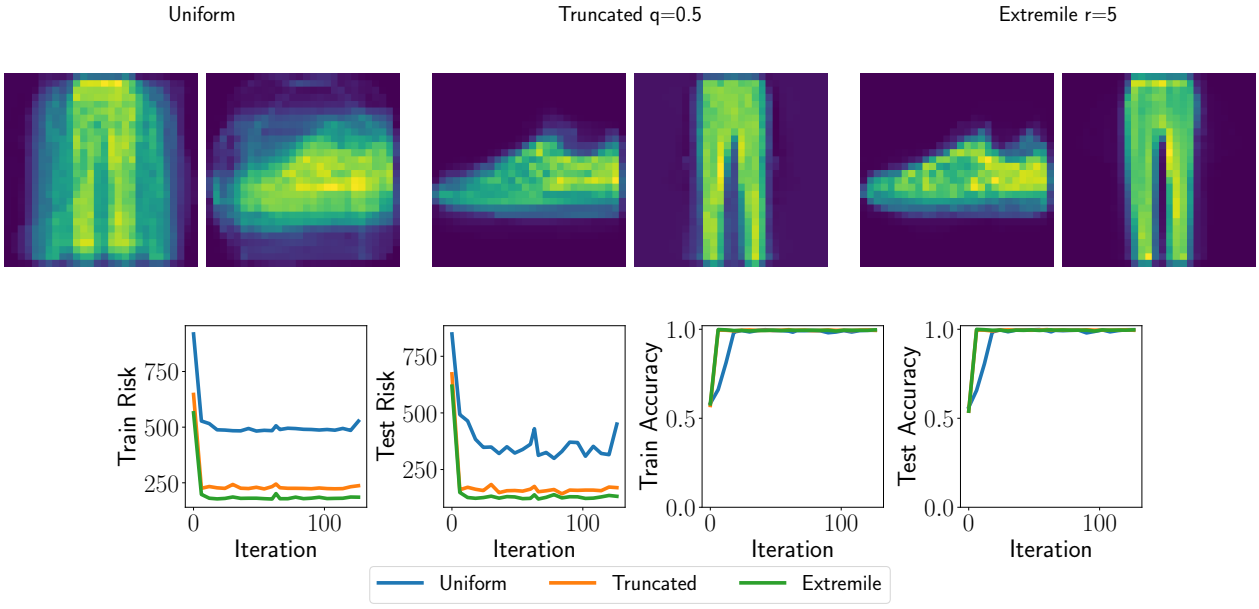

Figure 9: Clustering images of clothes in the presence of outliers.

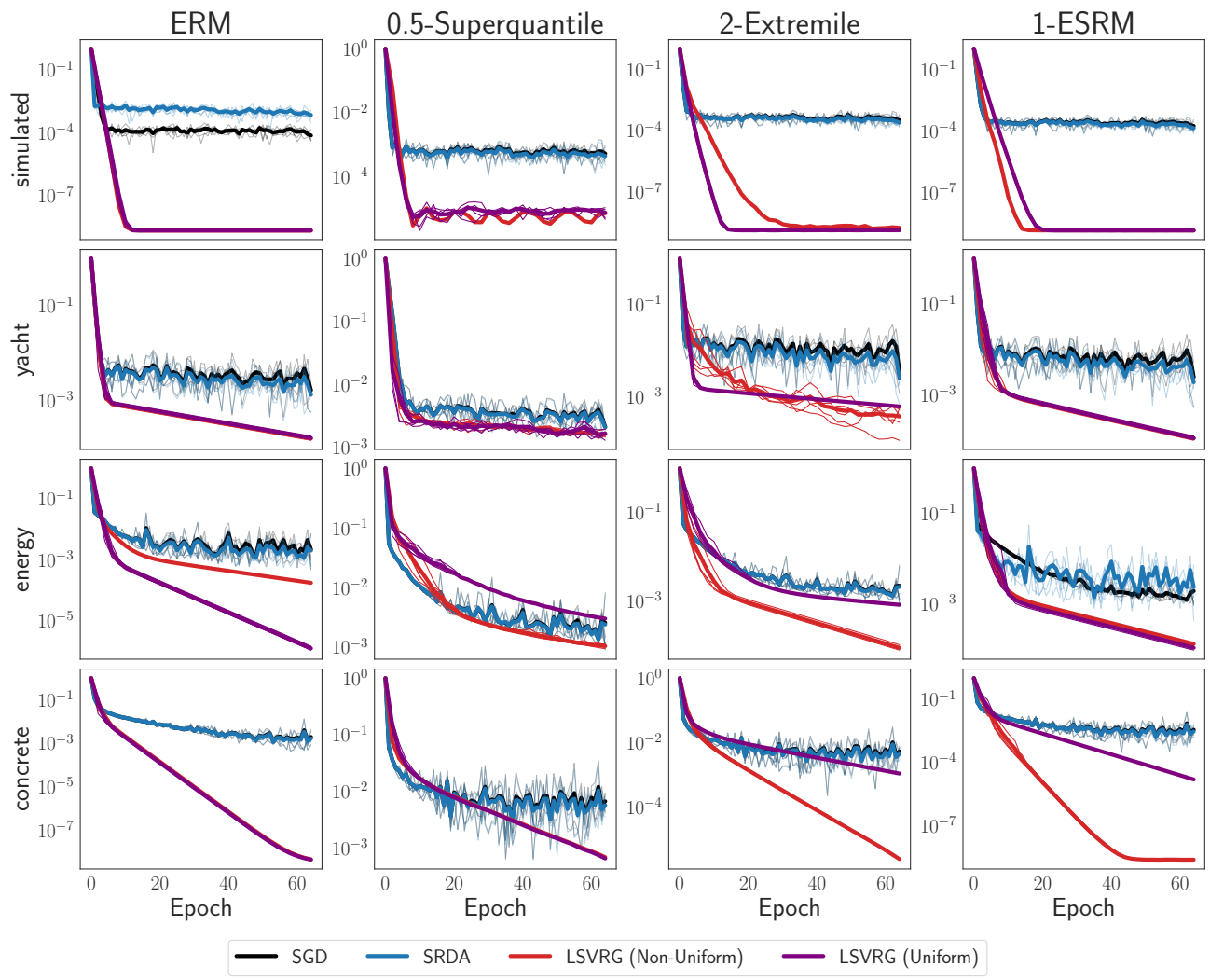

Figure 10: The suboptimality gap (base 10) for various optimization algorithms on spectral risk objectives for $\mu = 1/n$. The x-axis shows the number of effective passes through the data.

SRDA, suggesting that the superquantile is a particularly difficult learning objective.

Statistical Effect of Varying Risk Parameter We inspect how the test losses of the L-risk minimizers behave compared to the corresponding ERM solutions. Letting $\hat{w}_{\text{ERM}}$ be the approximate solution of ERM, whereas $\hat{w}_{\text{LRM}}$ is the approximate solution of an L-Risk minimization problem other than ERM, Fig. 16, Fig. 17, and Fig. 18 plot the following against p:

$$\ell_{(\lceil np \rceil)}\left(\hat{w}_{\text{ERM}}\right) - \ell_{(\lceil np \rceil)}\left(\hat{w}_{\text{LRM}}\right), \tag{31}$$

that is, the difference in the p-th quantile of the test loss of $\hat{w}_{\text{ERM}}$ and the p-th quantile of the test loss of $\hat{w}_{\text{LRM}}$. The plots are in order of "easy", "medium", and "hard" values of the risk parameters, corresponding to (q, r, ρ) being $(0.25, 1.5, 0.5)$, $(0.5, 2, 1)$, and $(0.75, 2.5, 2)$, respectively. The medium settings are shown primarily in the main text. The median test loss ($p = 0.5$) is similar between the L-risk minimizers and standard ERM across risk parameters. However, for $p > 0.5$, the ERM solution can make predictions with much higher loss, indicating that the tail is not controlled. The superquantile at parameters $q = 0.5$ generally fails to control test risk, even substantially underperforms in comparison to ERM in `energy`. On the other hand, the extremile and ESRM convincingly dominate ERM in the region $(0.9, 1)$ of the empirical quantile function for each of the risk parameters, with the extremile having a more pronounced effect.

Comparison between Smoothed and Non-smooth LSVRG We compare the implementation of LSVRG with smoothing presented in Alg. 5 to the non-smooth epoch-based implementation of LSVRG presented in Alg. 7. We consider the datasets

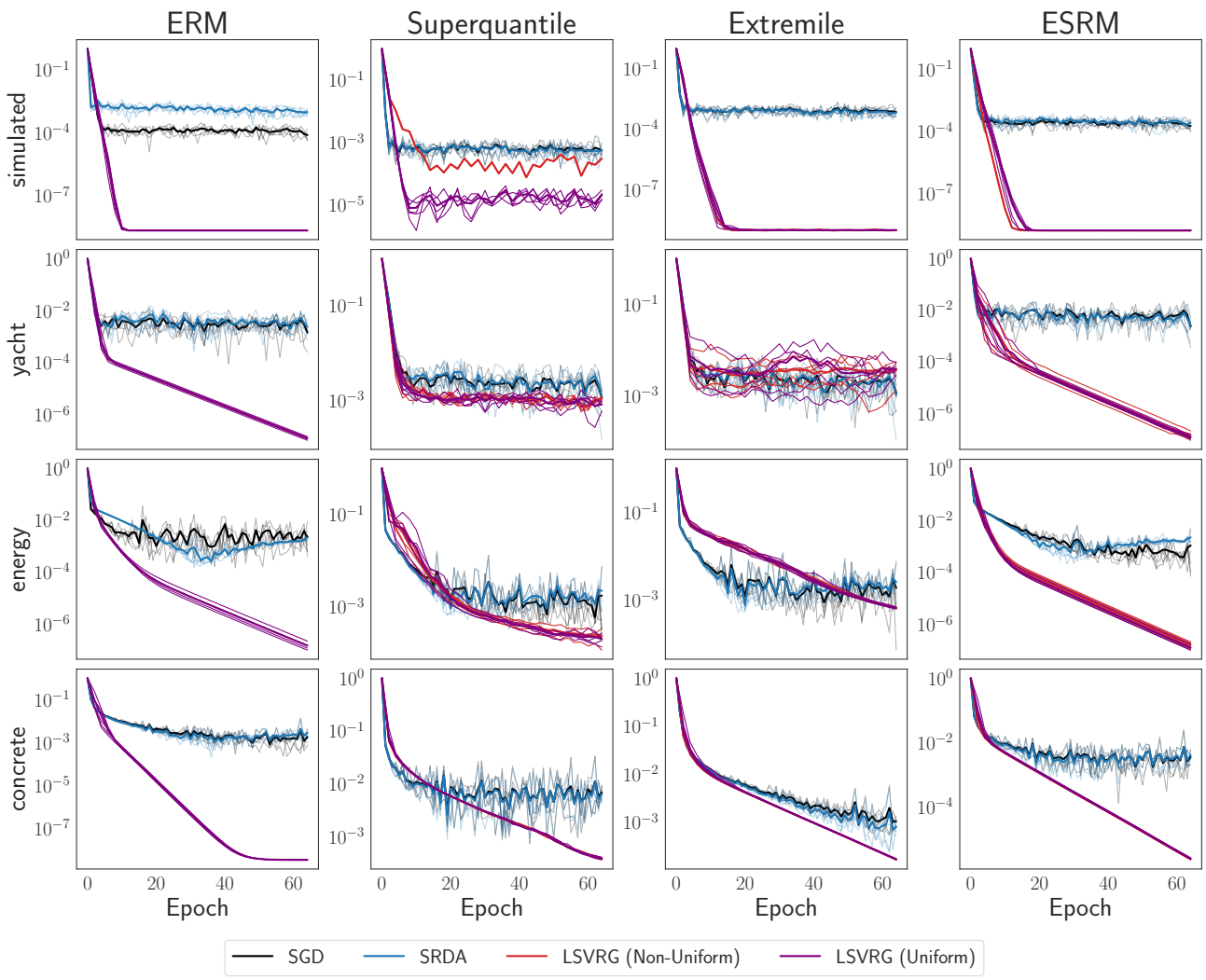

Figure 11: The suboptimality gap (base 10) for various optimization algorithms on spectral risk objectives for $\mu = 10/n$. The x-axis shows the number of effective passes through the data.

`simulated`, `yacht`, `energy` and `concrete` presented in Appx. F.1 and spectral risk measure objectives (2) defined by the empirical superquantile ($q = 0.5$), extremile ($r = 2$), and ESRM ($\rho = 1$) of the losses, plus an ℓ_2^2 regularization term of magnitude $1/n$.

We implemented the smoothed LSVRG algorithm (Alg. 5) with $N = n$, $q^* = 0$ and a smoothing given by either a centered negative entropy regularizer Ω_1 or a centered square Euclidean norm Ω_2, with Ω_1 and Ω_2 from Eq. (21). We consider using a smoothing coefficient of $\nu_1 = 10^{-3}$ for Ω_1 and $\nu_2 = n10^{-3}$ for Ω_2 (using the fact that the approximation done by Ω_2 has an approximation error of $\chi^2(s||u)/n$ as detailed in Appx. D). On the vertical axis we consider is the suboptimality gap $\frac{\mathcal{R}_\sigma(w^{(t)}) - \mathcal{R}_\sigma(w^*)}{\mathcal{R}_\sigma(w^{(0)}) - \mathcal{R}_\sigma(w^*)}$ for w^* computed by L-BFGS.

In Fig. 19, we observe that the non-smooth and smooth implementations of LSVRG generally match. For the ERM objective, this observation was expected since the permutahedron associated with the vector $u_n = \mathbf{1}/n$ reduces to $\{u_n\}$ since all entries of u_n are equal. Hence the maximization defining the smooth approximations $h_{\nu\Omega}$ given in Appx. D have a maximizer independent of the values of the losses and naturally given by u_n such that the smooth approximation of h reduces exactly to h for any choice of ν and Ω. For the other spectral risk measures, we observe some discrepancies between the non-smooth and the smooth implementations with the smooth implementation giving generally smoother curves as it is the case for the superquantile on the `simulated` dataset or the ESRM on the `concrete` dataset. However, such differences are not observed for, e.g., the superquantile on the `yacht`, `energy`, `concrete` datasets or the extremile and the ESRM objectives on the `simulated` and `yacht` datasets. Overall these experiments suggest that the non-smooth nature of

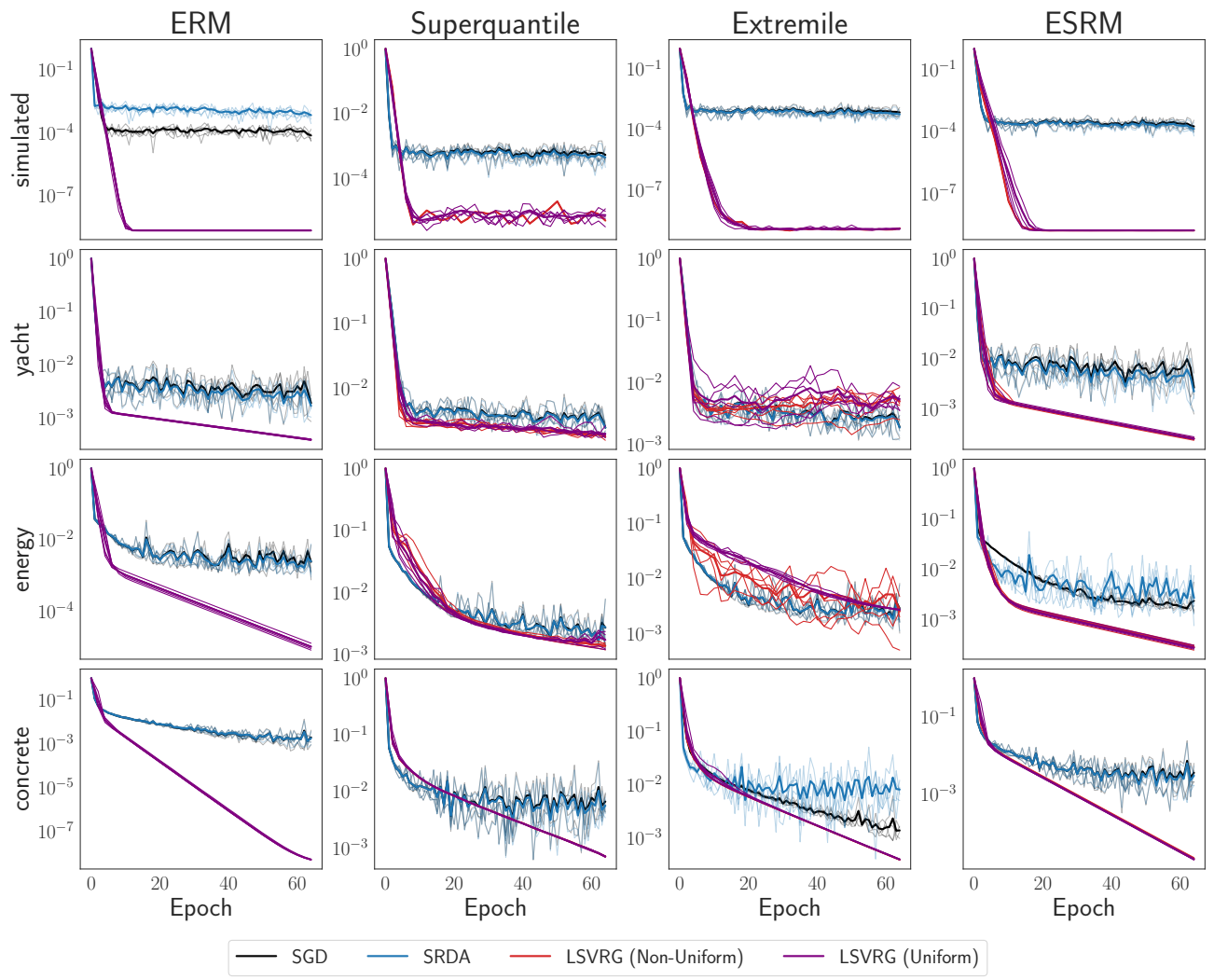

Figure 12: The suboptimality gap (base 10) for various optimization algorithms on spectral risk objectives for $\mu = 0.1/n$. The x-axis shows the number of effective passes through the data.

the problem has moderate impact on the performance of LSVRG. This behavior may be explained by the fact that the non-smoothness of the losses only intervene if the minimizer of the objective produces a vector of losses with ties which may not happen in practice. In addition note, that the negative entropy or the squared Euclidean smoothing generally give the same results (after appropriately scaling the smoothing coefficient of Ω_2 by n as suggested by the approximation errors given in Cor. 18 (Appx. D).

In Fig. 19, we also consider Alg. 5 with $N = 2n$, $q^* = 1/n$ and the same smoothing method as presented above. We scaled the horizontal axis by multiplying all algorithms by the total number of calls to the gradient oracles of the losses such that LSVRG in Alg. 7 is scaled by a factor 2 while Alg. 5 is scaled by a factor $\rho \geq 2$. We observe that the non-smooth implementation of LSVRG in Alg. 7 compares generally on par or better than the implementation of Alg. 5 after taking into account the total number of passes over the data, except for the ESRM risk on `concrete` and the extremile on `yacht`.

Run Time Experiments Fig. 20 contains plots of optimizer runtimes in each of the datasets considered. The values are calculated using the `time` module in Python 3 with logging disabled on the compute environment described in Appx. F.5. The two variants of LSVRG trade off run time for precision, as their suboptimality achieves ~ 1 order of magnitude improvement on `yacht`, up to ~ 4 orders of magnitude improvement on `concrete` over the SGD and SRDA baseline. SGD and SRDA also run ~ 2 orders of magnitude faster across datasets, but fail to converge due to both bias and variance.

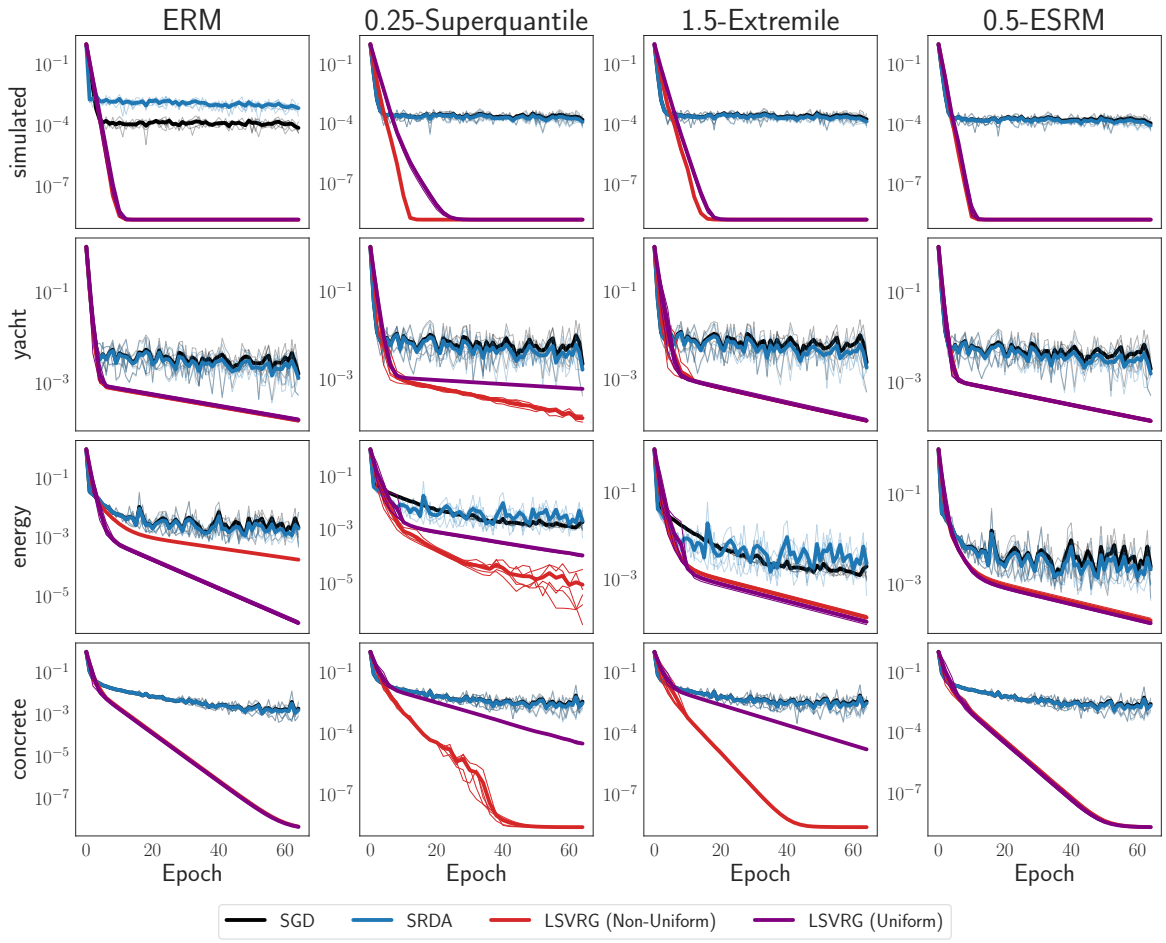

Figure 13: The suboptimality gap (base 10) for various optimization algorithms on ERM, q-superquantile, r-extremile, and ρ-ESRM objectives for(q, r, ρ) set to $(0.25, 1.5, 0.5)$. The x-axis shows the number of effective passes through the data.

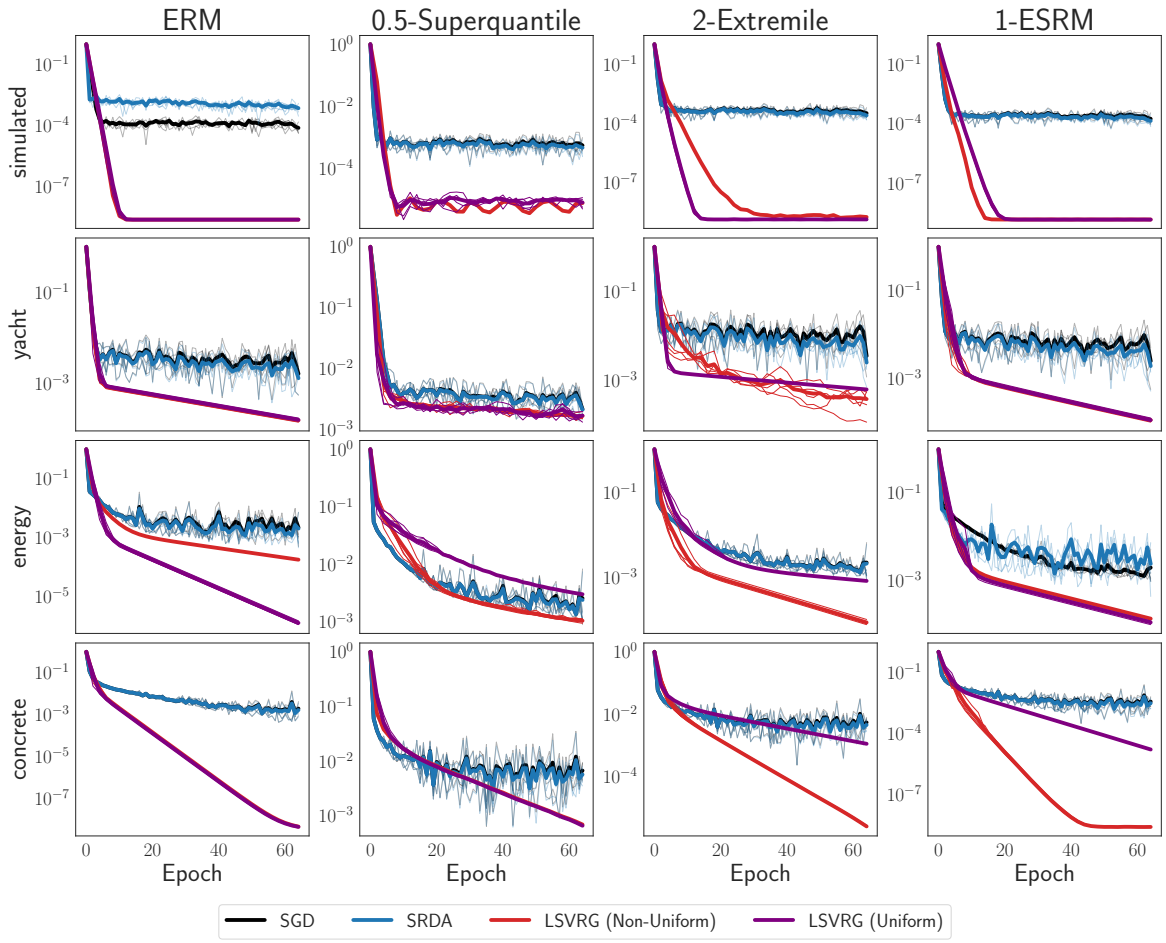

Figure 14: The suboptimality gap (base 10) for various optimization algorithms on ERM, q-superquantile, r-extremile, and ρ-ESRM objectives for(q, r, ρ) set to $(0.5, 2, 1)$. The x-axis shows the number of effective passes through the data.

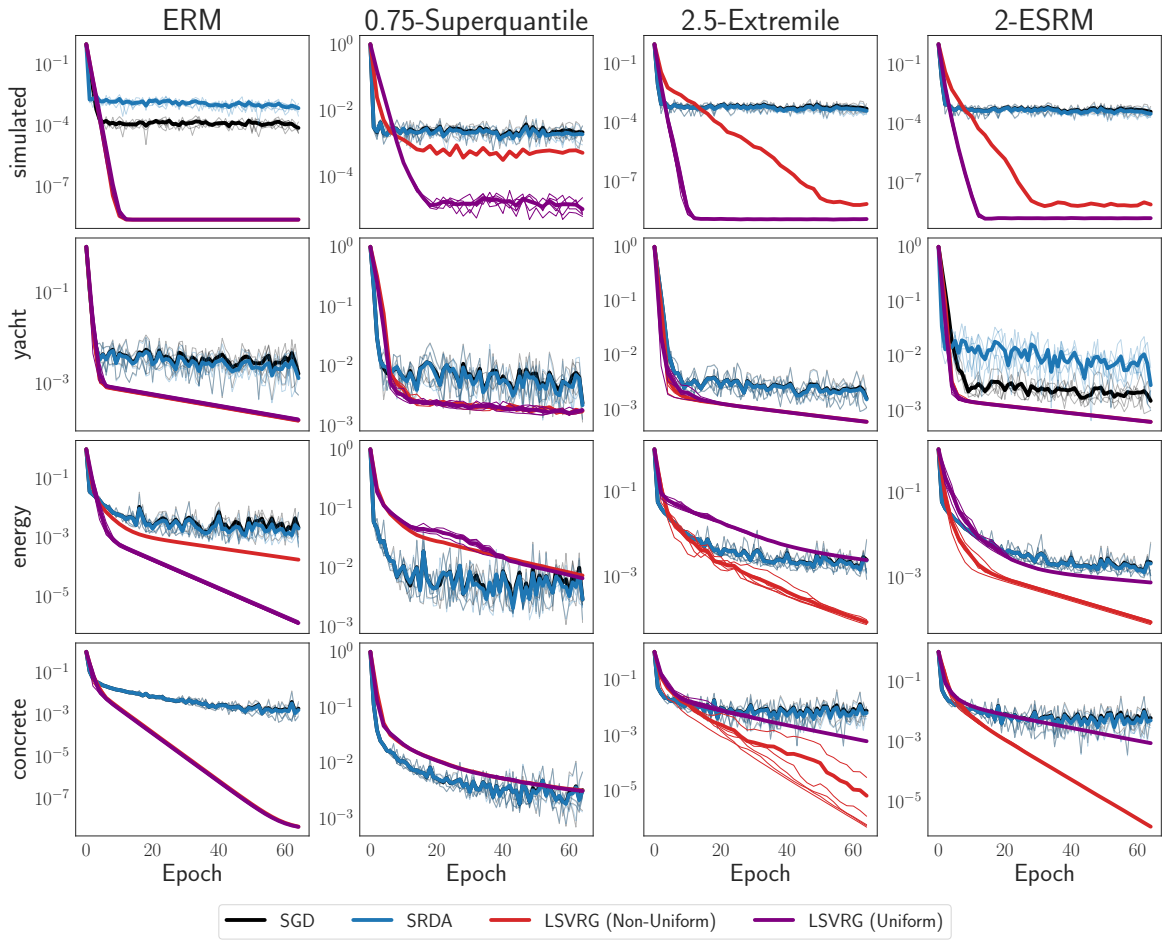

Figure 15: The suboptimality gap (base 10) for various optimization algorithms on ERM, q-superquantile, r-extremile, and ρ-ESRM objectives for(q, r, ρ) set to $(0.75, 2.5, 2)$. The x-axis shows the number of effective passes through the data.

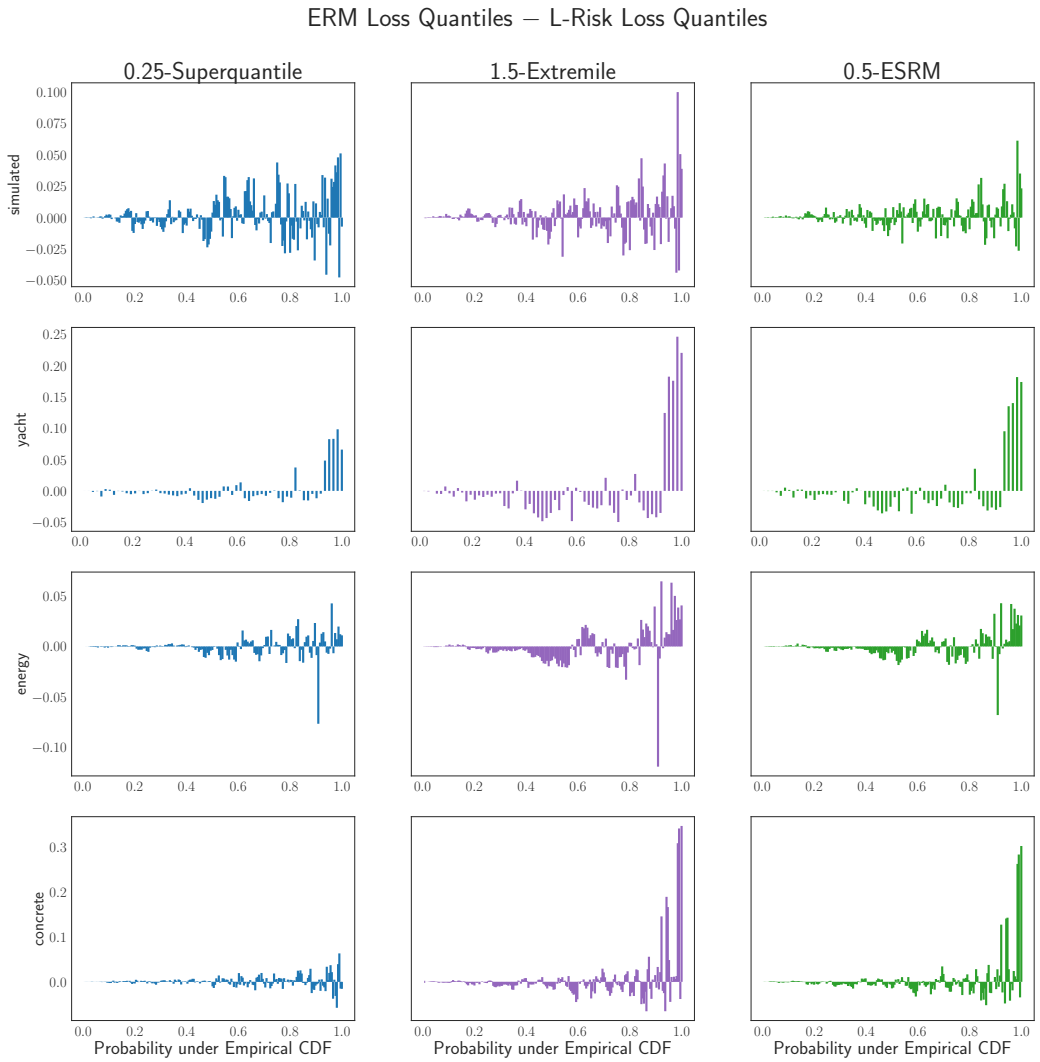

Figure 16: The difference between the empirical quantile function given by $\ell_{(1)}(\hat{w}_{\text{ERM}}), \ldots, \ell_{(n)}(\hat{w}_{\text{ERM}})$ and the empirical quantile function of an L-risk minimizer $\ell_{(1)}(\hat{w}_{\text{LRM}}), \ldots, \ell_{(n)}(\hat{w}_{\text{LRM}})$, where the L-risk is the q-superquantile (left column), r-extremile (middle column), or ρ-exponential spectral risk measure (right column). Each row represents a dataset out of `simulated`, `yacht`, `energy`, and `concrete`. Here, $(q, r, \rho) = (0.25, 1.5, 0.5)$, constituting L-risks that are "close" to ERM.

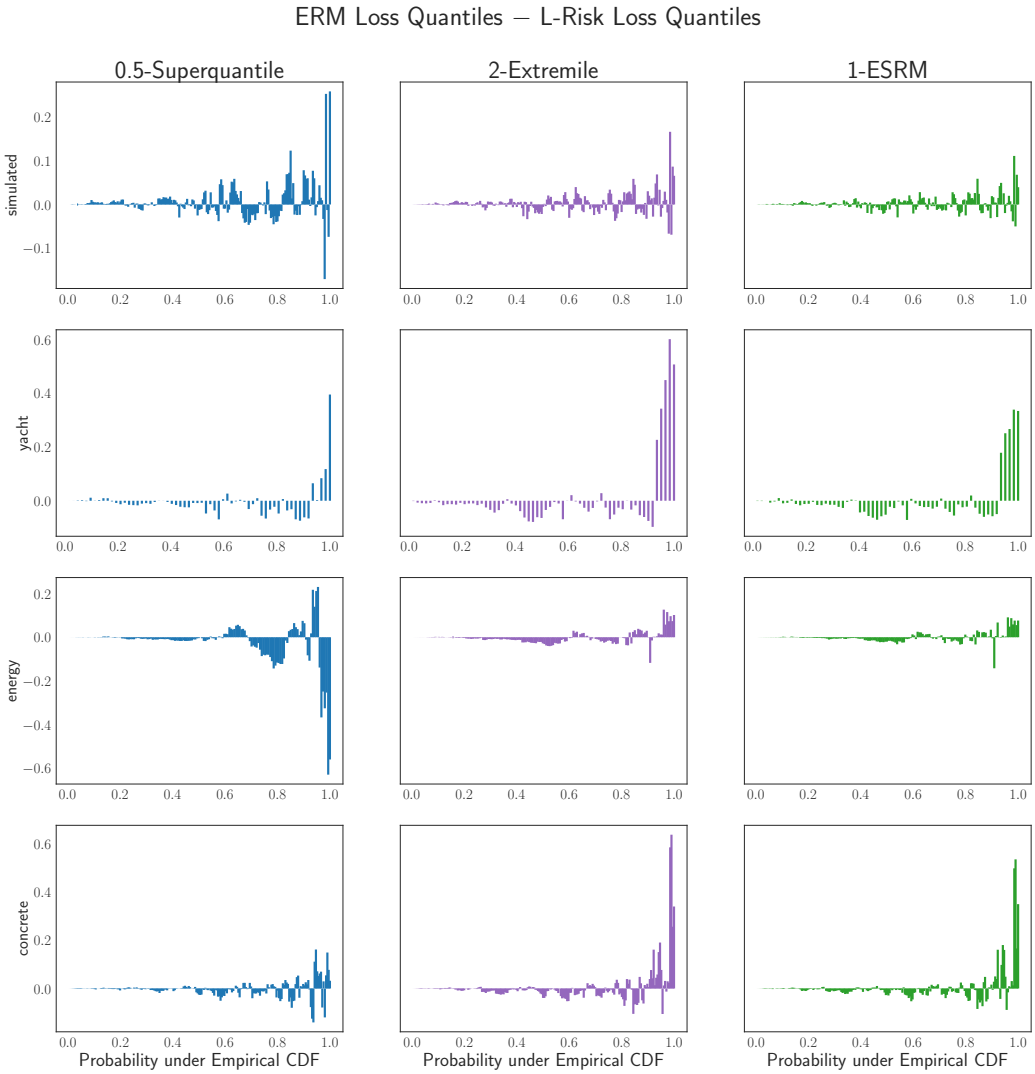

Figure 17: The difference between the empirical quantile function given by $\ell_{(1)}(\hat{w}_{\mathrm{ERM}}), \ldots, \ell_{(n)}(\hat{w}_{\mathrm{ERM}})$ and the empirical quantile function of an L-risk minimizer $\ell_{(1)}(\hat{w}_{\mathrm{LRM}}), \ldots, \ell_{(n)}(\hat{w}_{\mathrm{LRM}})$, where the L-risk is the q-superquantile (left column), r-extremile (middle column), or ρ-exponential spectral risk measure (right column). Each row represents a dataset out of `simulated`, `yacht`, `energy`, and `concrete`. Here, $(q, r, \rho) = (0.5, 2, 1)$, constituting L-risks that are "moderately far" from ERM.

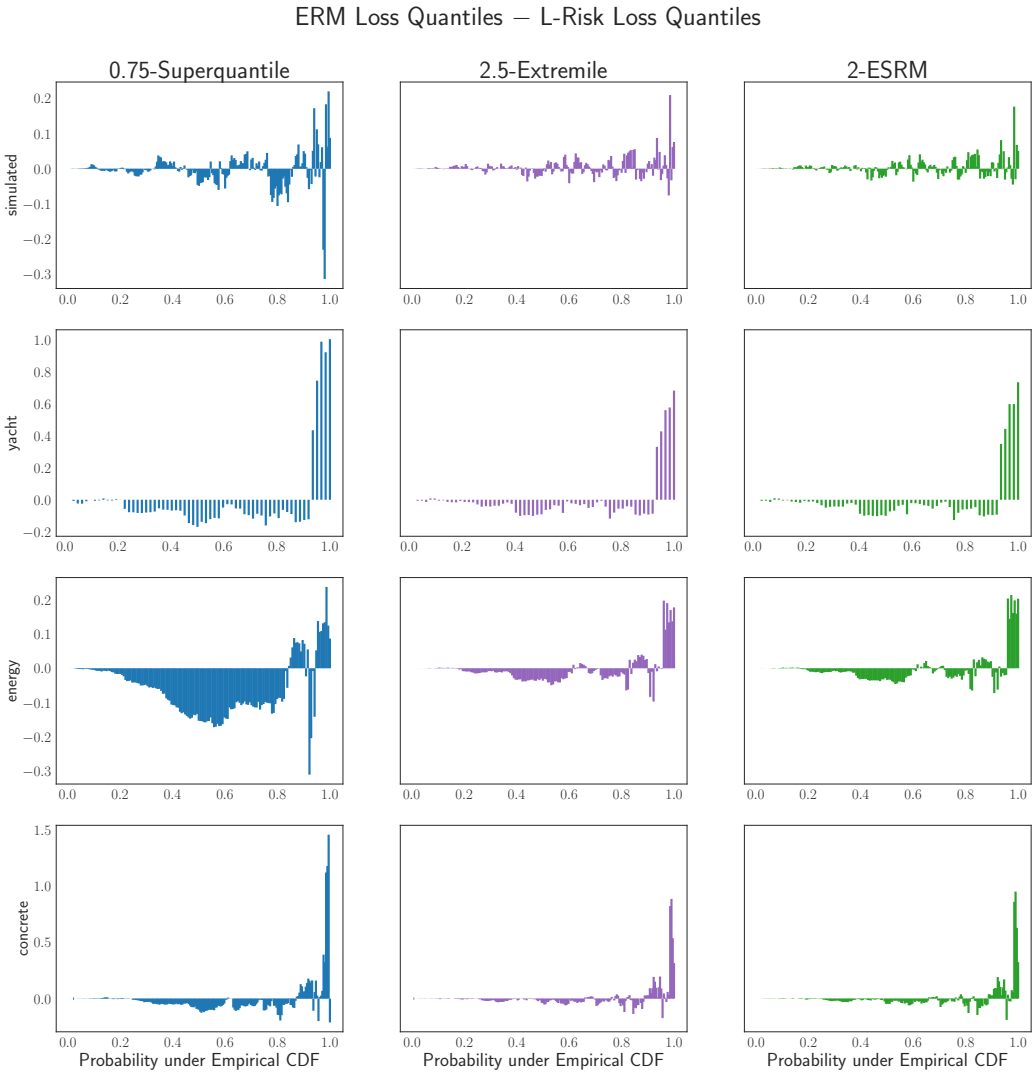

Figure 18: The difference between the empirical quantile function given by $\ell_{(1)}(\hat{w}_{\text{ERM}}), \ldots, \ell_{(n)}(\hat{w}_{\text{ERM}})$ and the empirical quantile function of an L-risk minimizer $\ell_{(1)}(\hat{w}_{\text{LRM}}), \ldots, \ell_{(n)}(\hat{w}_{\text{LRM}})$, where the L-risk is the q-superquantile (left column), r-extremile (middle column), or ρ-exponential spectral risk measure (right column). Each row represents a dataset out of `simulated`, `yacht`, `energy`, and `concrete`. Here, $(q, r, \rho) = (0.75, 2.5, 2)$, constituting L-risks that are "significantly far" from ERM.

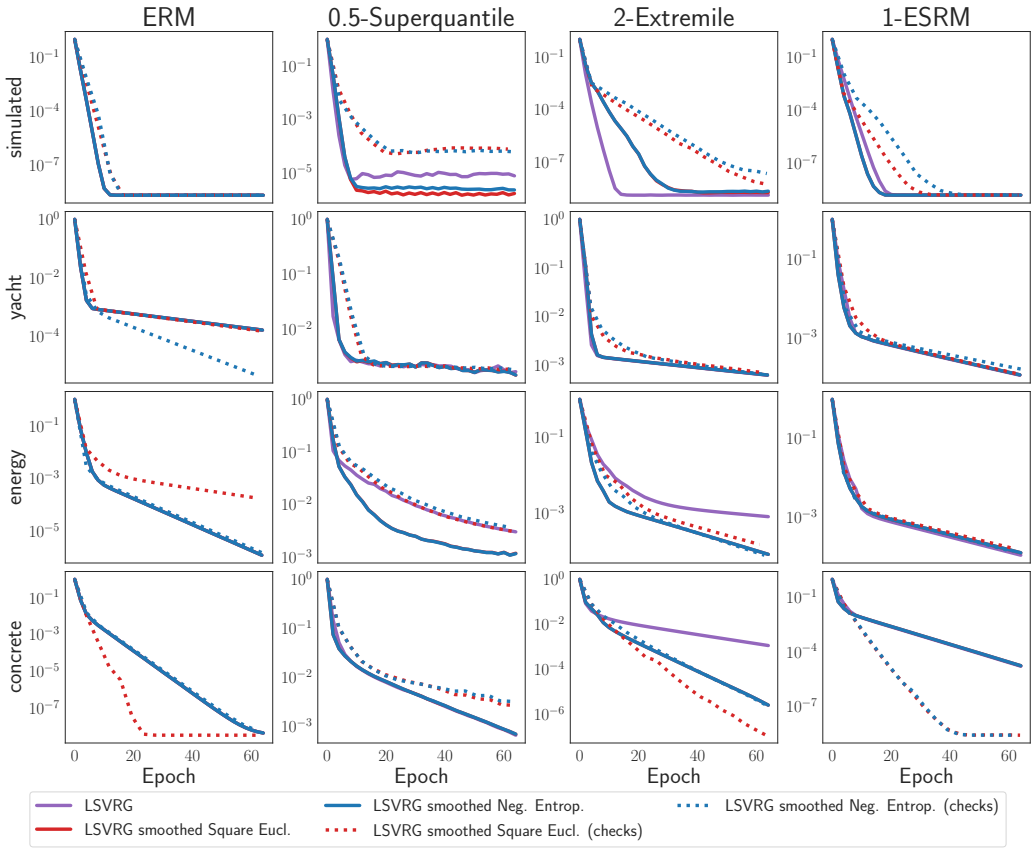

Figure 19: Comparison of the non-smooth implementation of LSVRG in Alg. 7 and the smoothed implementation of LSVRG in Alg. 5 with $N = n$, $q^* = 0$, and with a centered non-negative entropy smoothing function Ω_1 and $\nu_1 = 10^{-3}$ or a centered Euclidean smoothing function Ω_2 with $\nu_2 = n10^{-3}$ (see eq. (21) for the exact definitions of Ω_1, Ω_2).

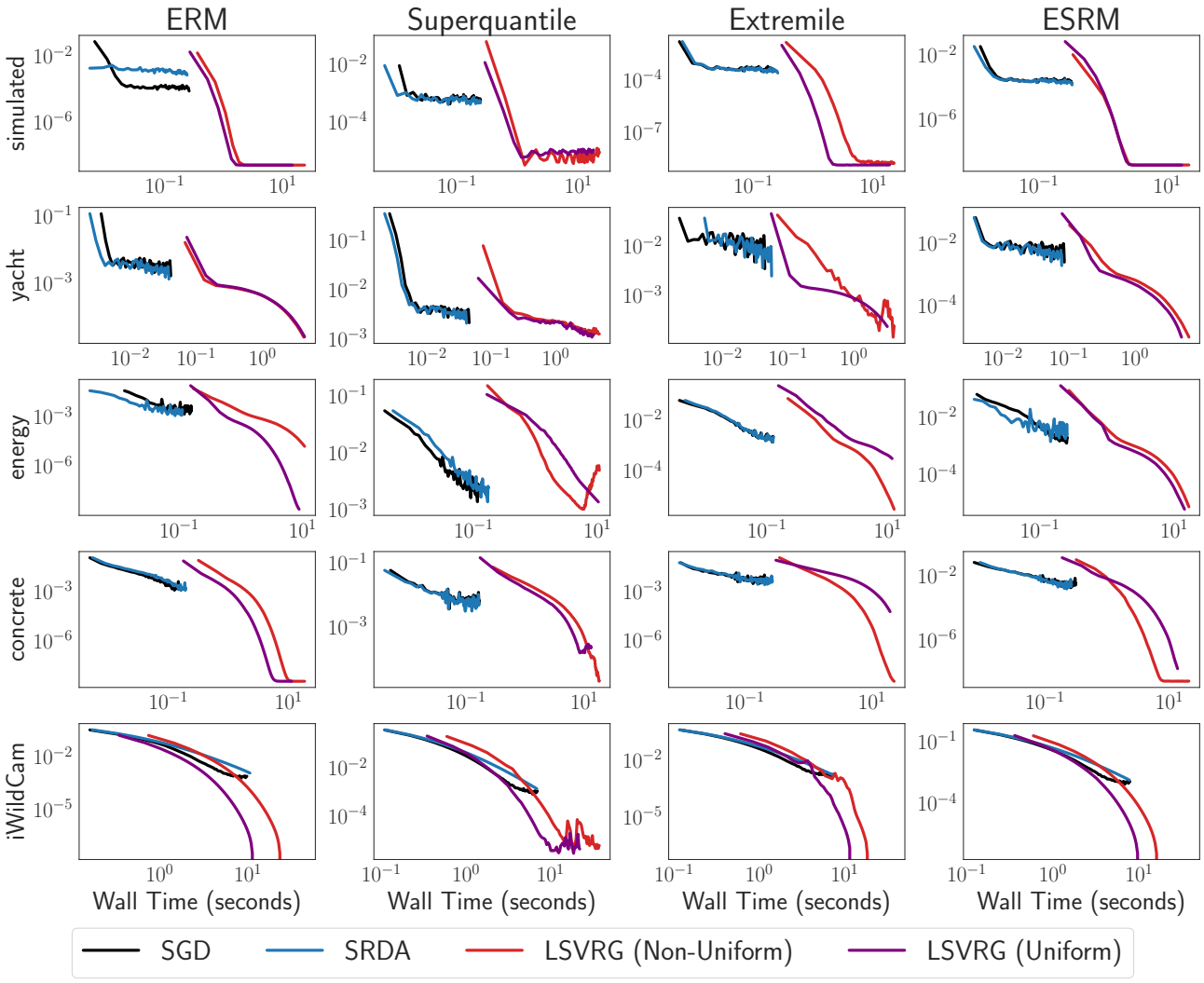

Figure 20: Algorithm run time of SGD, SRDA, and LSVRG optimizers on fives datasets (rows) and four objectives (columns). The y-axis plots the suboptimality in log scale, whereas the x-axis contains wall time in seconds in log scale.