Stochastic Optimization for Spectral Risk Measures

Ronak Mehta, Vincent Roulet, Krishna Pillutla, Lang Liu, Zaid Harchaoui

Background

Large prediction errors made by ML models can cause catastrophic/unfair outcomes. "Worst-case" performance is not captured by average loss when n is large!

IBM's Watson recommended 'unsafe and incorrect' treatments for cancer patients, investigation reveals

2 Killed in Driverless Tesla Car Crash, Officials Say

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

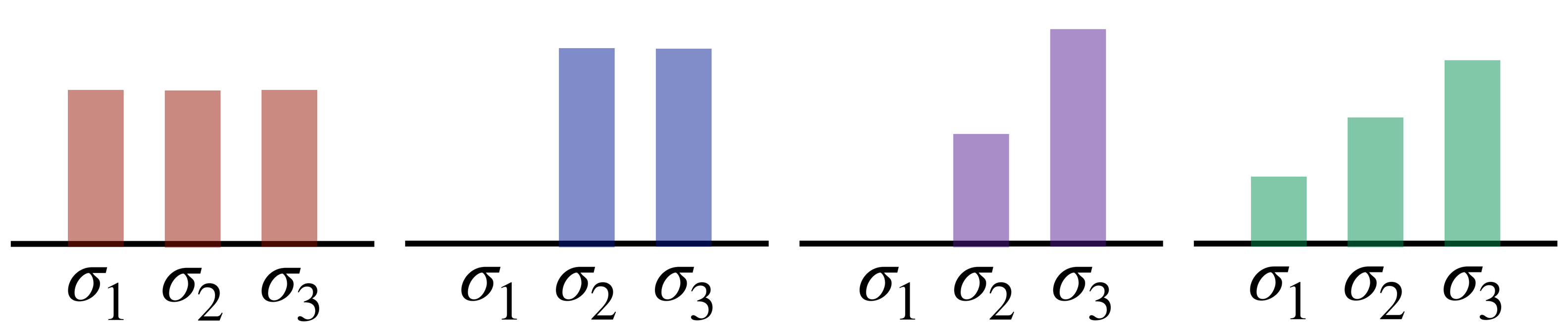
Setting

Spectral Risk Measures: A Robust Objective

Non-decreasing weights $\sigma_1 \leq \ldots \leq \sigma_n$

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

Model parameters. i-th smallest losses on training set.



Key Challenge 1: Optimizing SRMs **stochastically** (using only $O(1)$ gradient evaluations from oracles $\ell_1, \ldots, \ell_n$), as objective depends on sorted order of all training losses.

Key Challenge 2: Analyzing algorithms for non-smooth objective R_σ.

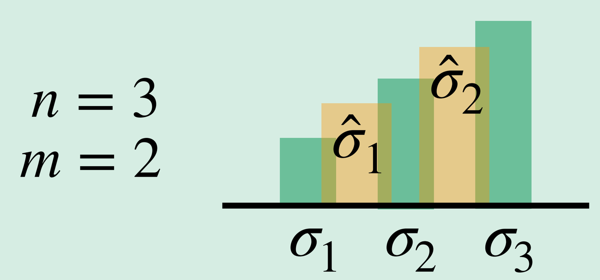
Algorithms

Proposition If the losses are convex and differentiable, then R_σ is convex, and if permutation π satisfies $\ell_{\pi(1)}(w) \leq \ldots \leq \ell_{\pi(n)}(w)$ (i.e. is an "argsort"), then $\sum_{i=1}^{n} \sigma_i \nabla \ell_{\pi(i)}(w)$ is a subgradient of R_σ at w.

Mini-batch SGD

1. Sample minibatch $i_1, \ldots, i_m$ uniformly .
2. Let π sort $\ell_{i_1}(w_t), \ldots, \ell_{i_m}(w_t)$.
3. Compute $v_t = \sum_{j=1}^{m} \hat{\sigma}_j \nabla \ell_{i_{\pi(j)}}(w_t)$.
4. Update $w_{t+1} = w_t - \eta_t v_t$.

Update direction v_t is biased for population subgradient, as $\hat{\sigma}_1, \ldots, \hat{\sigma}_m$ is a "coarsening" of the full batch SRM.



L-SVRG

1. Every $O(n)$ iterates, store checkpoint $\bar{w}$, let π sort $\ell_1(\bar{w}), \ldots, \ell_n(\bar{w})$, compute $\nabla R_\sigma(\bar{w})$.
2. Sample i uniformly.
3. Compute $v_t = n\sigma_i \nabla \ell_{\pi(i)}(w_t)$ and $c_t = \nabla R_\sigma(\bar{w}) - n\sigma_i \nabla \ell_{\pi(i)}(\bar{w})$
4. Update $w_{t+1} = w_t - \eta(v_t + c_t)$.

Update direction v_t is still biased, but asymptotically unbiased. Control variate c_t reduces variance, learning to convergence.

Theory

Consider the problem

$$\min_{w \in \mathbb{R}^d} R_\sigma(w) + (\mu/2)\|w\|_2^2,$$

where each ℓ_i is G-Lipschitz and L-smooth.

Theorem 1

Minibatch SGD suboptimality is

$$\lesssim c_\sigma \sqrt{\frac{n-m}{nm}} + \frac{G^2 \log(t)}{\mu t}$$

Bias, $c_\sigma \to 0$ when closer to ERM.

Theorem 2

L-SVRG suboptimality is

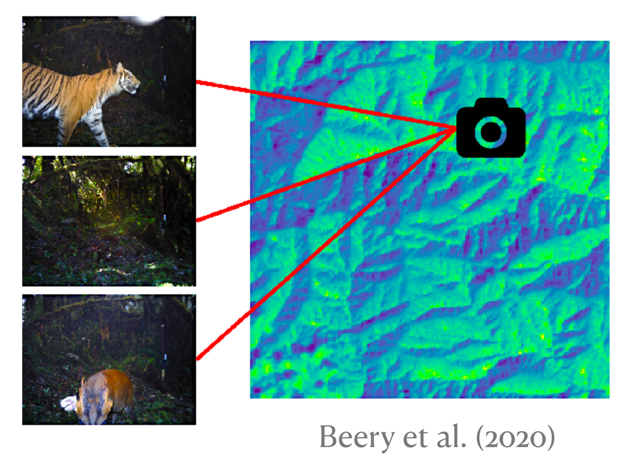
$$\lesssim c_\sigma G^2/\mu + \left(2^{0.25}\right)^{-\frac{t}{n+8\kappa}}$$

Smoothing error, $c_\sigma \to 0$ when closer to ERM.

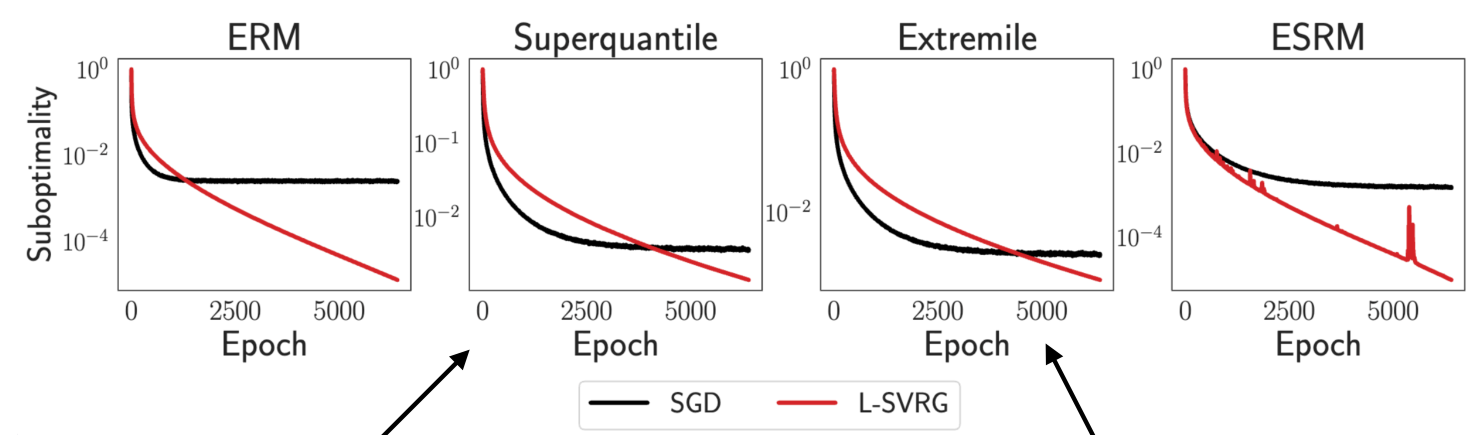
Linear rate $\kappa \sim n\sigma_n L/\mu$

Experiments

Fine-tuning image classification models on WILDS iWildCam



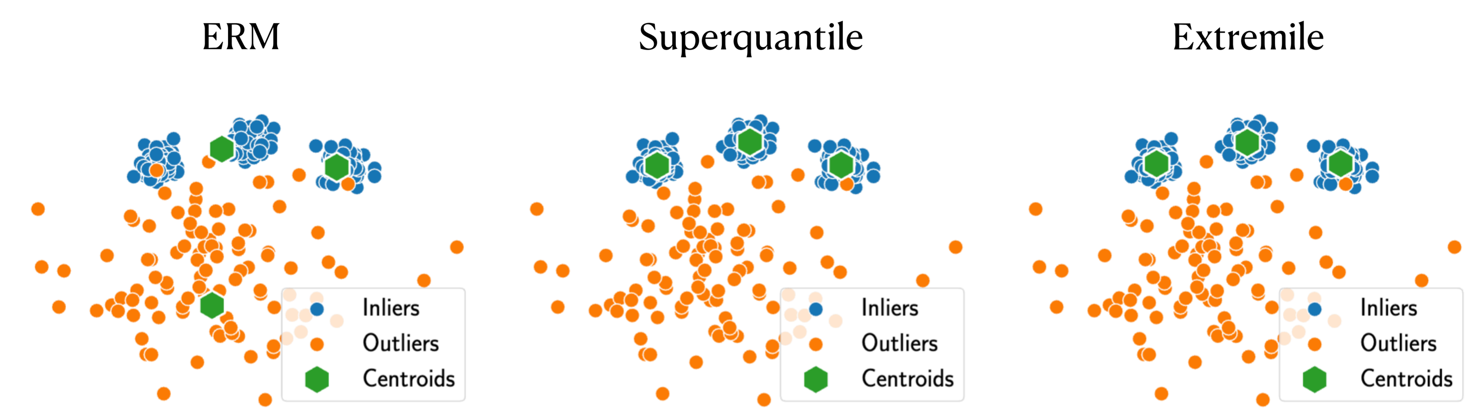
Beery et al. (2020)



SGD is hindered by bias and does not converge.

L-SVRG exhibits empirical linear convergence

Clustering in the presence of outliers



SRM minimizers are resistant to synthetic perturbations in unsupervised setting.

Full Paper & Code

ronakdm

www ronakdm.github.io

SCAN ME