Distributionally Robust Optimization with Bias and Variance Reduction

Ronak Mehta, Vincent Roulet, Krishna Pillutla, Zaid Harchaoui

Distributional Robustness

Standard Empirical Risk Minimization

model parameters (primal variables)

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

training data weights $\mathbf{1}/n = (1/n, \ldots, 1/n)$

vector of losses on data point i

Distributionally Robust (DR) Objectives

shifted weights (dual variables)

shift penalty

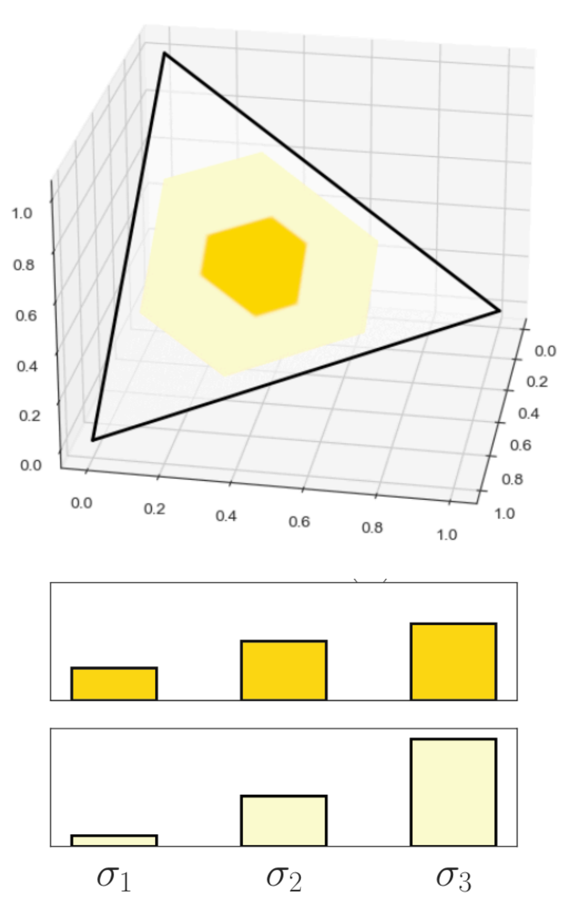
$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathcal{P}} \left\{ \sum_{i=1}^{n} q_i \ell_i(w) - \nu D(q \| \mathbf{1}/n) \right\}$$

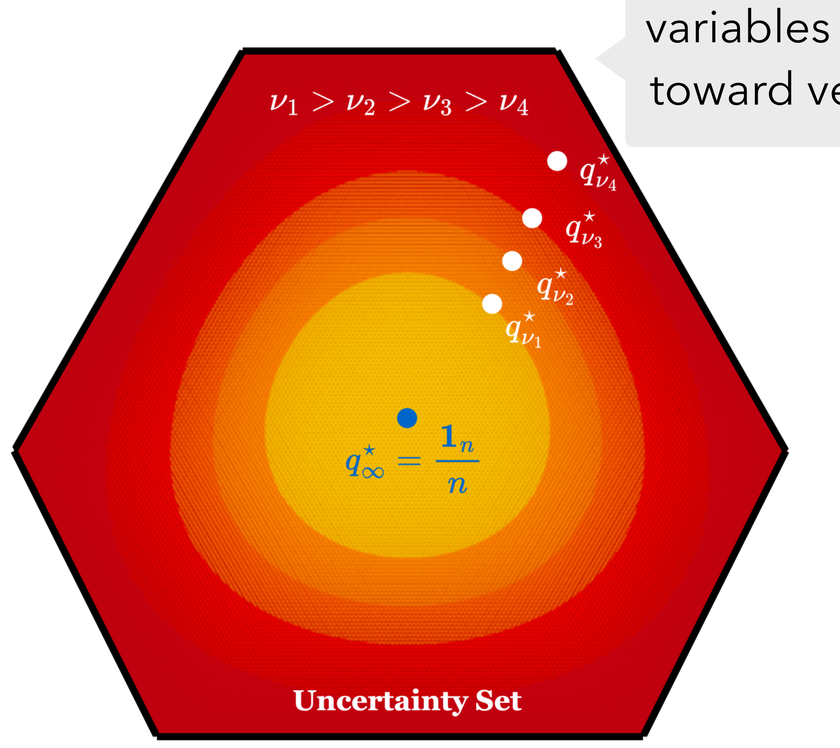
$$\mathcal{P} = \text{conv}\left(\text{permutations of } \sigma\right)$$

ambiguity set

most "skewed" weights possible

optimal dual variables tend toward vertex





Goal: construct a *stochastic, linearly convergent* optimization algorithm for DR objectives.

Algorithm: Prospect

Objective:

$$\min_{w \in \mathbb{R}^d} \max_{q \in \mathcal{Q}} \left\{ q^\top \ell(w) - \nu D(q \| \mathbf{1}/n) \right\} + \frac{\mu}{2} \|w\|_2^2$$

Input — learning rate η

Stored in Memory: — current primal iterate w

Estimates of loss/gradient at each data point $l_1, \ldots, l_n \quad g_1, \ldots, g_n$

Two estimates of dual-optimal variables at primal iterate $q_1, \ldots, q_n \quad \hat{q}_1, \ldots, \hat{q}_n$

Main Loop:

1. $i \sim \text{Unif}\{1, \ldots, n\}$. — sample data point
2. $v_1 = q_i \nabla \ell_i(w) + \mu w$ — compute gradient estimate
3. $v_2 = g_i - \sum_{j=1}^{n} \hat{q}_j g_j$ — compute variance reducer
4. $w \leftarrow w - \eta(v_1 - v_2)$ — **main update**
5. $q \leftarrow \text{argmax}_{q \in \mathcal{Q}} \sum_{i=1}^{n} q_i l_i - \nu D(q \| \mathbf{1}/n)$ — update all tables
6. $\hat{q}_i \leftarrow q_i, \quad l_i \leftarrow \ell_i(w), \quad g_i \leftarrow \nabla \ell_i(w)$

Key Idea: Instead of solving dual problem using true losses (which cost n oracle calls to compute), use lazily updated table of losses to approximate dual solution. Update direction has *asymptotically vanishing bias and variance*, as the tables estimates become exact in the limit.

Convergence Analysis

Assume the losses are convex, L-smooth and G-Lipschitz.

Define the condition numbers $\kappa_\ell = L/\mu + 1$ and $\kappa_\sigma = n\sigma_{\max}$, and constant $\kappa_\nu = G^2/(\nu\mu)$.

1. Prospect with $\eta \sim \text{poly}(n, \kappa_\ell, \kappa_\sigma, \kappa_\nu)^{-1}$ converges linearly.
2. If in addition, $\kappa_\nu \leq 1/6$, then for $\eta \sim (\kappa_\sigma(L + \mu))^{-1}$ and $\tau \sim n + \kappa_\sigma \kappa_\ell$, we have for iterates $w_0, w_1, \ldots$, that

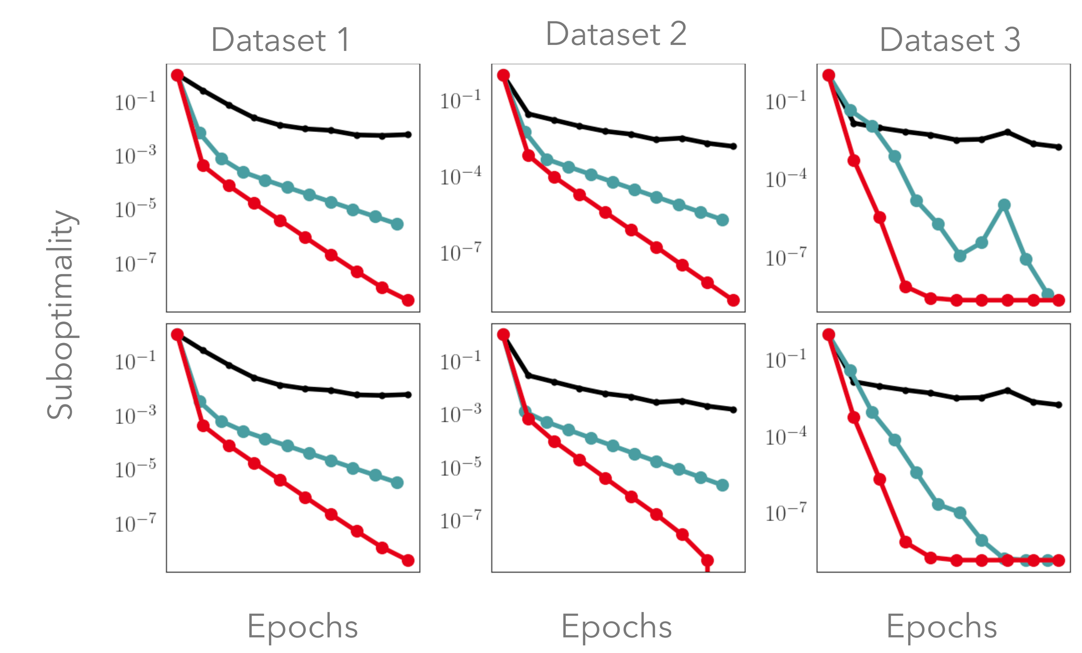
$$\mathbb{E}\|w_t - w^\star\|_2^2 \leq 6n^2 \exp(-t/\tau) \|w_0 - w^\star\|_2^2.$$

Experiments

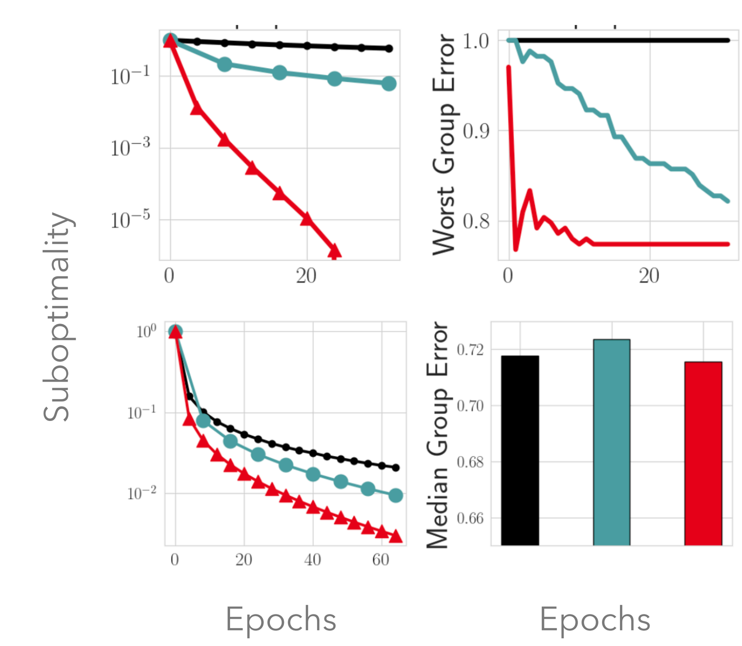
SGD — LSVRG — Prospect (Ours)

Training Distributionally Robust Linear Models

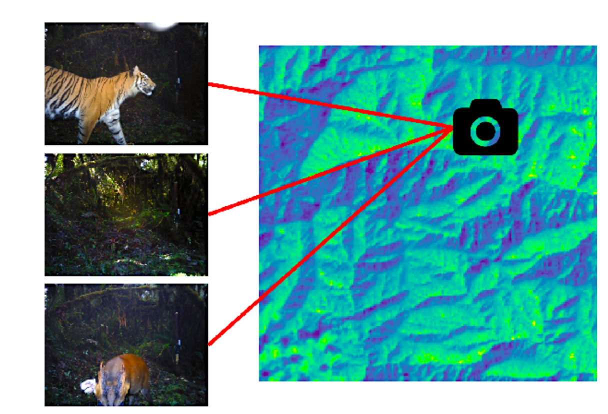
Prospect has best/close to best optimization performance in terms of gradient evaluations, across datasets.



Subpopulation Shift in Image/Text Classification



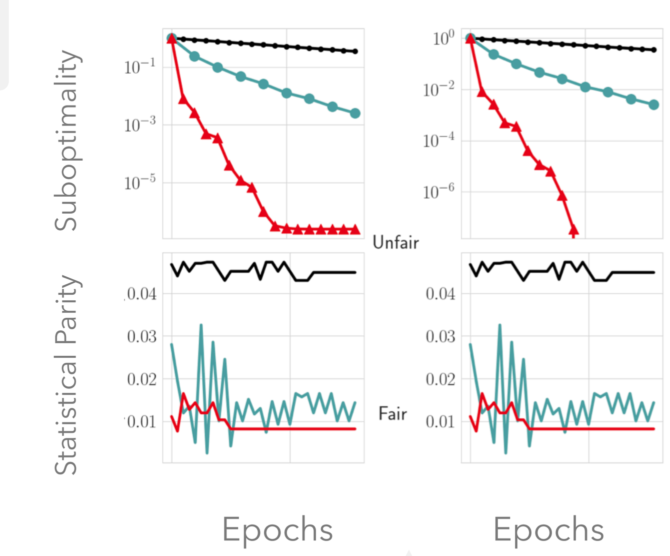
Worst group test error mitigated by Prospect solution on Amazon Reviews.



Beery et al. (2020)

Median group test error mitigated by Prospect solution on iWildCam.

Using SRMs to Promote Group Fairness



DR objective correlates with statistical parity fairness score.

Full Paper + Code