Drago: Primal-Dual Coupled Variance Reduction for Faster DRO

Ronak Mehta, Jelena Diakonikolas, Zaid Harchaoui

UNIVERSITY of WASHINGTON · IFML · NSF · Institute for Foundations of Data Science IFDS · Washington · Wisconsin · Santa Cruz · Chicago

Distributionally Robust Optimization

DRO is a popular framework for modeling distribution shift in the training phase.

model parameters (primal variables)

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell_i(w) + \frac{\mu}{2}\|w\|_2^2$$

losses on data point i

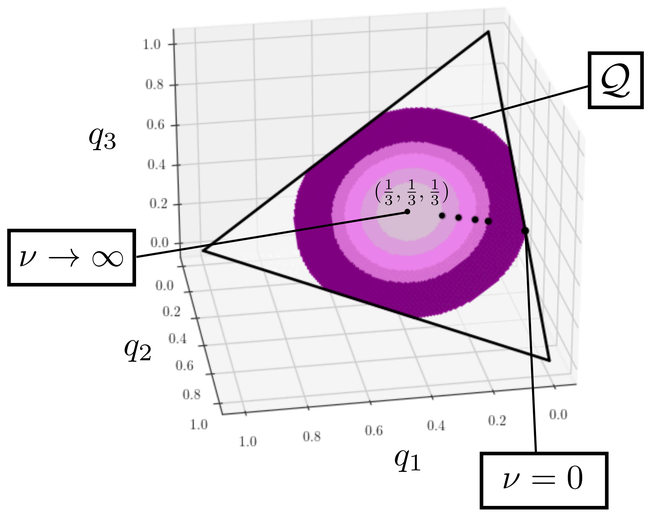
dual variables · dual regularization

$$\min_{w\in\mathbb{R}^d}\max_{q\in\mathcal{Q}} \sum_{i=1}^n q_i\ell_i(w) - \nu D(q\|\mathbf{1}/n) + \frac{\mu}{2}\|w\|_2^2$$

uncertainty set

Examples

$$\mathcal{Q} = \begin{cases} \{\mathbf{1}/n\} \\ \{q : \left\|q - \frac{\mathbf{1}}{n}\right\|_2 \le \rho\} \\ \Delta^{n-1} \end{cases}$$



Current algorithms for DRO either 1) do not converge at all, 2) converge sublinearly, or 3) converge linearly only with stringent conditions on the regularization constants of the problem.

Contribution: linearly-convergent algorithm with improved dependence on the sample size under all parameter regimes.

Context

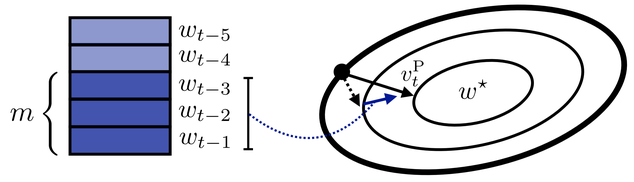
- Objective is a **nonbilinearly** coupled saddle-point problem with **nondecomposable** feasible set (contains constraint $\sum_{i=1}^n q_i = 1$).
- Dependence on n (sample size) is especially important in the "big data" regime; we use a carefully selected mini-batch size to improve overall runtime to $O(n^{3/2})$ instead of $O(n^2)$.
- The analysis handles **all parameter regimes** and **all common uncertainty** sets in a unified way.

Algorithm: Drago

High-Level: A Bregman-type proximal gradient method with stochastic gradient estimates, cyclically updated table of past iterates, and primal regularization for reducing variance in the dual.

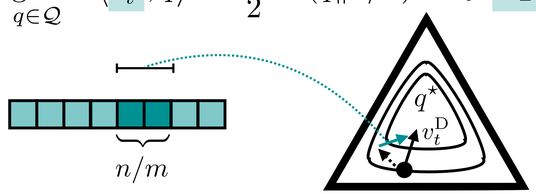
Primal: Sample-Based Variance Reduction

$$w_t = \arg\min_{w\in\mathbb{R}^d} \langle v_t^{\mathrm{P}}, w\rangle + \frac{a_t\mu}{2}\|w\|_2^2 + \frac{b_t\mu}{2}\sum_{\tau=1}^m \|w - w_{t-\tau}\|_2^2$$



Dual: Coordinate-Based Variance Reduction

$$q_t = \arg\max_{q\in\mathcal{Q}} \langle v_t^{\mathrm{D}}, q\rangle - \frac{a_t\nu}{2}D(q\|\mathbf{1}/n) - b_t\nu\Delta_D(q, q_{t-1})$$



Convergence Analysis

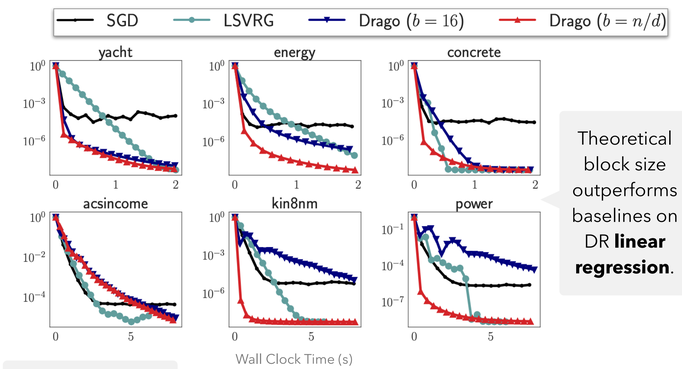
Assumptions and Notation

$$|\ell_i(w) - \ell_i(w')| \le G\|w - w'\|_2 \quad \text{(Lipschitz losses)}$$
$$\|\nabla\ell_i(w) - \nabla\ell_i(w')\|_2 \le L\|w - w'\|_2 \quad \text{(smooth losses)}$$
$$\kappa_{\mathcal{Q}} = n\max\{q_i : q\in\mathcal{Q}, i\in[n]\} \quad \text{(uncertainty)}$$

Theorem. Drago with block size n/d reaches suboptimality ε with global complexity of the order

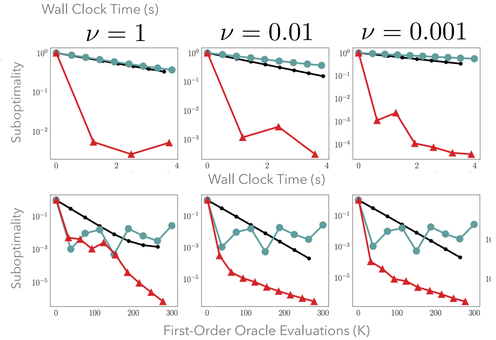
$$O\left(nd\left(\frac{\kappa_{\mathcal{Q}}L}{\mu} + \frac{\sqrt{n}G}{\sqrt{\mu\nu}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$$

Experiments



Theoretical block size outperforms baselines on DR **linear regression**.

Performance is resistant to changes in dual regularization on **text classification**.



Paper/Code