Coordinate Linear Variance Reduction for Generalized Linear Programming

Chaobing Song[1*], **Cheuk Yin (Eric) Lin[1*]**, Stephen J. Wright[1], Jelena Diakonikolas[1]

[1]*Department of Computer Sciences, University of Wisconsin-Madison*

**Equal Contribution*

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Overview

We study a class of **generalized linear programs (GLP)** in a large-scale setting, which includes a simple, possibly non-smooth convex regularizer and simple convex set constraints:

$$\min_{\mathbf{x}} \left\{ \mathbf{c}^T\mathbf{x} + r(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \right\} \qquad \text{(GLP).}$$

1) By reformulating (GLP) as an equivalent convex-concave min-max problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \mathbf{c}^T\mathbf{x} + r(\mathbf{x}) + \mathbf{y}^T\mathbf{A}\mathbf{x} - \mathbf{y}^T\mathbf{b} \right\} \qquad \text{(PD-GLP),}$$

we design an efficient, scalable first-order algorithm named ***Coordinate Linear Variance Reduction (CLVR)***. CLVR yields improved complexity results for (GLP) that depend on the max row norm of the linear constraint matrix $\mathbf{A}$ rather than the spectral norm. We further introduce two strategies to improve the convergence rates: 1) **Lazy updates** when the regularization term and constraints are coordinate-separable, and 2) an **adaptive restart scheme** when $r(\mathbf{x}) = 0$.

2) By introducing sparsely connected auxiliary variables, we show that **Distributionally Robust Optimization (DRO)** problems with ambiguity sets based on both f-divergence and Wasserstein metrics **can be reformulated as (GLPs)**.

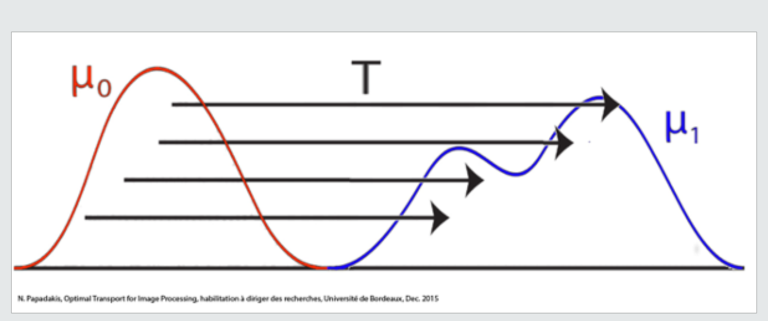
Generalized Linear Programs (GLP)

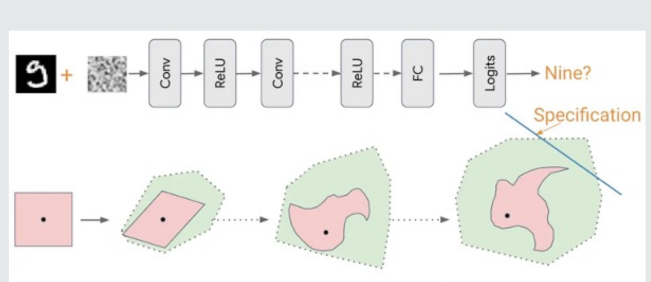
$$\min_{\mathbf{x}} \left\{ \mathbf{c}^T\mathbf{x} + r(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \right\} \qquad \text{(GLP)}$$

- $\mathcal{X}$ is a closed convex set in $\mathbb{R}^d$ admitting efficient projections
- r is a σ-strongly convex regularizer admitting efficiently computable proximal operators, where $\sigma \geq 0$. When r is only convex, we say $\sigma = 0$
- (GLP) reduces to a linear program (LP) when $r(\mathbf{x}) = 0$ and $\mathcal{X}$ polyhedral

Applications of GLP

- Linear programming
- Reinforcement learning [De Farias and Van Roy, 2003]
- Optimal transport [Villani, 2009]
- Neural network verification [Liu et al., 2020]
- **Distributionally robust optimization [This work]**





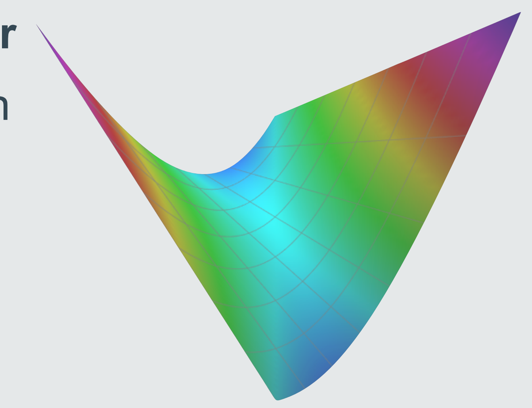
Our Algorithm (CLVR)

We aim to solve

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \mathcal{L}(\mathbf{x}, \mathbf{y}) := \mathbf{c}^T\mathbf{x} + r(\mathbf{x}) + \mathbf{y}^T\mathbf{A}\mathbf{x} - \mathbf{y}^T\mathbf{b} \right\} \qquad \text{(PD-GLP),}$$

where some of existing algorithms for solving (PD-GLP) include PDHG [CP11], SPDHG [CERS18], VRPDA² [SWD21], PURE-CD [ACF20]. We introduce three novel approaches to improve upon the existing results:

1. Based on VRPDA² and **by exploiting the linear structure** of $\mathcal{L}(\mathbf{x}, \mathbf{y})$ w.r.t. $\mathbf{y}$, CLVR removes an expensive initialization step requiring a single access to full-data
2. Also using **the linear structure**, CLVR uses an extrapolation term in the output point, which further cancels a variance term
3. When $\mathbf{A}$ **is sparse and when** $\mathcal{X}$ **and** r **are coordinate separable**, we can handle updates in a *lazy manner* only when coordinates are sampled.



$$O\left(\frac{nd\|A\|}{\epsilon}\right) \text{ in SPDHG } \xrightarrow{2} O\left(\frac{ndR}{\epsilon}\right) \text{ in CLVR } \xrightarrow{3} O\left(\frac{\mathrm{nnz}(A)R}{\epsilon}\right)$$

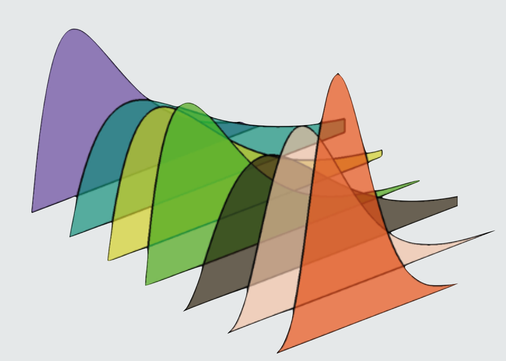
where $R = \max_{j\in[n]} \|A^j\|$ is the max row norm of $\mathbf{A}$ and $R \leq \|\mathbf{A}\| \leq \sqrt{n}R$

Novel Connection from DRO to GLP

DRO Formulation

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{p}\in\mathcal{P}} \sum_{i=1}^{n} p_i g(f(\mathbf{x}, \mathbf{a}_i), b_i)$$

where $\mathcal{P}$ is the uncertainty set around the uniform distribution $\mathbf{1}/n$ and $g(z, b)$ is the loss function.



- DRO can be seen as a robust generalization to empirical risk minimization problems. In DRO, we **minimize the worst case risk based on some ambiguity sets over the probability distribution of training data**

Reformulation to GLP

- We consider a simplified setting with linear predictors and binary classes

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{p}\in\mathcal{P}} \sum_{i=1}^{n} p_i g(b_i \mathbf{a}_i^T \mathbf{x})$$

and g can be a non-smooth loss function (e.g., hinge loss)

- **We show that such DRO problems based on f-divergence and Wasserstein metric can be reformulated into equivalent GLPs**

Adaptive Restart via Sharpness w.r.t LPMetric

- Standard form LP has a sharpness property w.r.t. the normalized duality gap [AHLL21], and it can be used **to obtain linear convergence rates** in first-order methods
- Instead of the normalized duality gap, we use **the classical LPMetric** [H52] as our measure of optimality and **showed its sharpness:**

$$\mathrm{LPMetric}(\mathbf{x}, \mathbf{y}) = \sqrt{\|\max\{\mathbf{x}, \mathbf{0}\}\|_2^2 + \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \|\max\{-\mathbf{A}^T\mathbf{y} - \mathbf{c}, \mathbf{0}\}\|_2^2 + |\max\{\mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{y}, 0\}|^2}$$

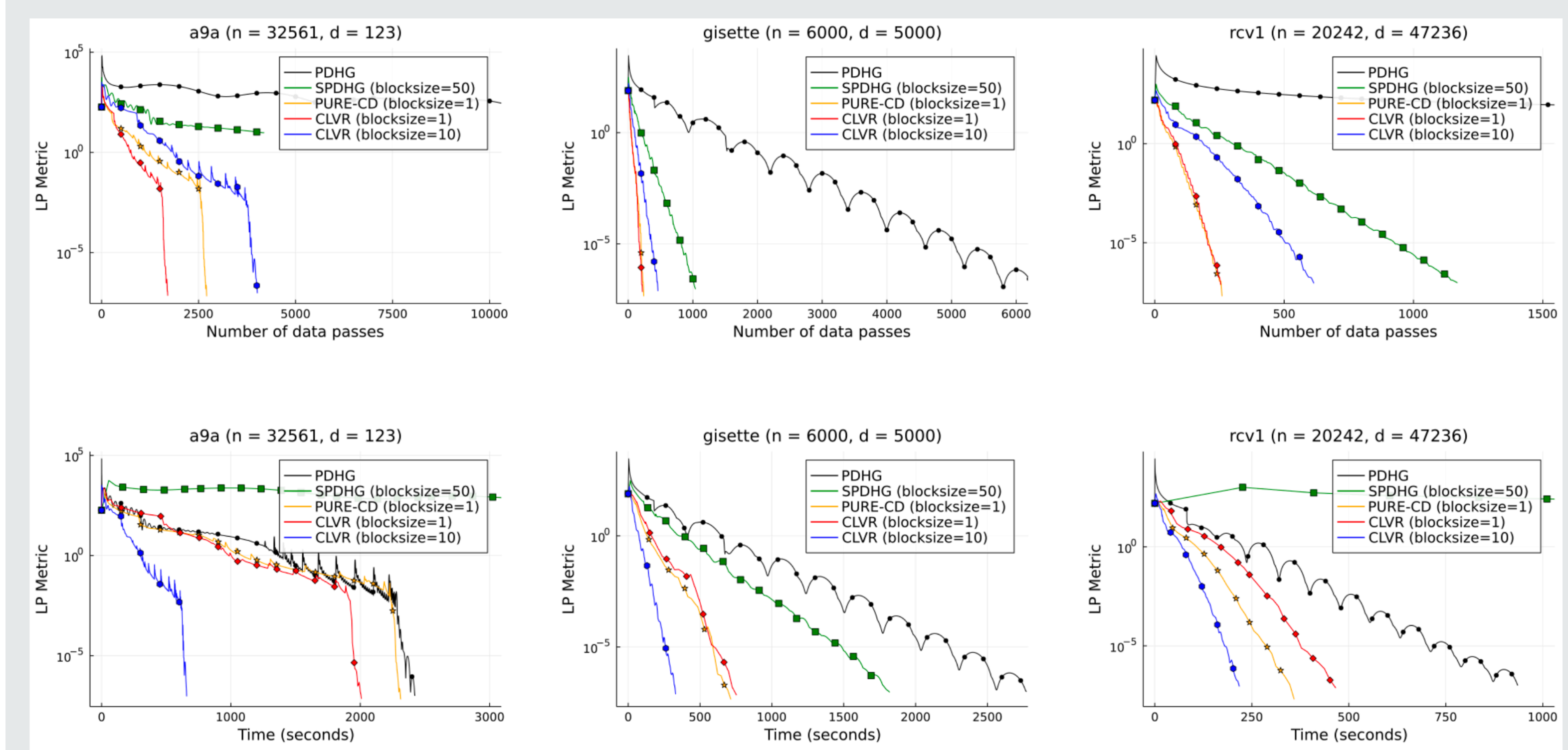
Numerical Experiments

We solve the linear program **reformulation for DRO with Wasserstein distance of ℓ_1 norm and hinge loss**, where we also apply our novel adaptive restart scheme using LPMetric.

Comparison Between Values of L when $R = 1$

Reformulated a9a $d = 130738, n = 97929$	Reformulated gisette $d = 44002, n = 28000$	Reformulated rcv1 $d = 269914, n = 155198$	Reformulated news20 $d = 5500750, n = 2770370$
117.3	65.9	196.4	1041.6

Comparison with Primal-dual Algorithms



Comparison with Production Linear Programming Solvers

Time (seconds)	Reformulated a9a $d = 130738, n = 97929$	Reformulated gisette $d = 44002, n = 28000$	Reformulated rcv1 $d = 269914, n = 155198$
JuMP+GLPK	899	$> 4 \times 10^4$	$> 4 \times 10^4$
JuMP+Gurobi(simplex)	893	2482	7008
JuMP+Gurobi(barrier)	**26**	1039.7	1039.5
CLVR	962	**697**	**582**

Acknowledgements

CS was supported in part by the NSF grant 2023239. JD and CYL acknowledge support from the NSF award 2007757. JD was also supported by the Office of Naval Research under contract number N00014-22-1-2348 and the Wisconsin Alumni Research Foundation. SW was supported by NSF grants 2023239 and 2224213, the DOE under subcontract 8F-30039 from Argonne National Laboratory, and the AFOSR under subcontract UTA20-001224 from UT-Austin. Part of this work was done while JD, CS, and SW were visiting the Simons Institute for the Theory of Computing.

Main References

[CERS18] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[SWD21] C. Song, S. J. Wright, and J. Diakonikolas. Variance reduction via primal-dual accelerated dual averaging for nonsmooth convex finite-sums. In *Proc. ICML'21*, 2021.

[AHLL21] D. Applegate, O. Hinder, H. Lu, and M. Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. *arXiv preprint arXiv:2105.12715*, 2021.