A First-Order Algorithmic Framework for Wasserstein Distributionally Robust Logistic Regression

Jiajin Li, Sen Huang, Anthony Man-Cho So
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

Problem Setting

Problem Setup: Consider the Wasserstein distance-induced distributionally robust logistic regression (**DRLR**) problem as follows

$$\inf_{\beta\in\mathbb{R}^n} \sup_{\mathbb{Q}\in B_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell_\beta(x,y)] \quad (1)$$

- $\ell_\beta(x,y) = \log(1+\exp(-y\beta^T x))$ with the feature label pair $(x,y)\in\Theta := \mathbb{R}^n \times \{+1,-1\}$
- $\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{(\hat{x}_i,\hat{y}_i)}$: Empirical distribution
- Wasserstein distance-induced ambiguity set: $B_\epsilon(\hat{\mathbb{P}}_N) = \{\mathbb{Q}\in\mathcal{P}(\Theta) : W(\mathbb{Q},\hat{\mathbb{P}}_N)\le\epsilon\}$

$$W(\mathbb{Q},\hat{\mathbb{P}}_N) = \inf_{\Pi\in\mathcal{P}(\Theta\times\Theta)} \left\{\int_{\Theta\times\Theta} d(\xi,\xi')\Pi(d\xi,d\xi') : \Pi(d\xi,\Theta)=\mathbb{Q}(d\xi),\ \Pi(\Theta,d\xi')=\hat{\mathbb{P}}_N(d\xi')\right\}$$

- $d(\xi,\xi') = \|x-x'\| + \frac{\kappa}{2}|y-y'|$ where κ represents the label reliability
- (1) admits a tractable conic reformulation (A). In particular, if $\kappa=+\infty$ and $\|.\|$ is the ℓ_∞-norm, (A) reduces to the well-known **ℓ_1-regularized logistic regression**

$$\inf_\beta \left\{\frac{1}{N}\sum_{i=1}^N \ell_\beta(\hat{x}_i,\hat{y}_i) + \epsilon\|\beta\|_1\right\}.$$

- DRO methodology provides a **principled** approach to regularization.

Limitations: Apply **interior-point** algorithms based off-the-shelf solvers (e.g., YALMIP) to tackle (A), which severely limits the applicablity of the DRO approach in large-scale learning problems.

Can we propose a lightweight first-order algorithm to solve (A) ($\kappa<\infty$) efficiently? Yes!

Key Steps in LP-ADMM

- Quadratic programming with box constraints (i.e., w.r.t (D))

$$\beta^{k+1} = \arg\min_{\beta\in\mathbb{R}^n}\left\{\|Z\beta-\mu^k-\frac{w^k}{\rho_k}\|_2^2 + \mathbb{I}_{\{\|\beta\|_\infty\le\lambda\}}\right\}.$$

- By exploiting the strongly convex property of the local model, we propose a **first-order** linearized update

$$\mu^{k+1} = \arg\min_{\mu\in\mathbb{R}^N}\left\{\frac{1}{N}\sum_{i=1}^N (h'(\mu_i^k)\mu_i + \max\{\mu_i-\lambda\kappa,0\}) + \frac{\rho_k}{2}\|\mu - Z\beta^{k+1}+\frac{w^k}{\rho_k}\|_2^2\right\}.$$

- **Geometrically increased** penalty parameter $\rho_{k+1}=\gamma\rho_k, \gamma>1$.

Our First-Order Algorithmic Framework

$$\begin{aligned}\min_{\beta,\mu}\ & \frac{1}{N}\sum_{i=1}^N (h(\mu_i)+\max\{\mu_i-\lambda\kappa,0\})\\ \text{s.t.}\ & Z\beta-\mu=0,\\ & \|\beta\|_\infty\le\lambda.\end{aligned} \quad (C)$$

$\downarrow$ β Exact Update

$$\begin{aligned}\min_\beta\ & \|Z\beta-\mu^k-\frac{w^k}{\rho}\|_2^2\\ \text{s.t.}\ & \|\beta\|_\infty\le\lambda.\end{aligned} \quad (D)$$

β-subproblem (LP-ADMM)

$\leftarrow$ Fix λ

$$\begin{aligned}\min_{\beta,\lambda}\ & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^N (h(\hat{y}_i\beta^T\hat{x}_i)+\\ & \max\{\hat{y}_i\beta^T\hat{x}_i-\lambda\kappa,0\})\\ \text{s.t.}\ & \|\beta\|_\infty\le\lambda.\end{aligned} \quad (B)$$

$\updownarrow$ Exact Reformulation

$$\begin{aligned}\min_{\beta,s,\lambda}\ & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^N s_i\\ \text{s.t.}\ & \ell_\beta(\hat{x}_i,\hat{y}_i)\le s_i,\ i\in[N],\\ & \ell_\beta(\hat{x}_i,-\hat{y}_i)-\lambda\kappa\le s_i\ i\in[N],\\ & \|\beta\|_\infty\le\lambda.\end{aligned} \quad (A)$$

ℓ_1 DRLR Problem

Note that $h(\mu)=\log(1+\exp(-\mu))$, $z_i=\hat{y}_i\hat{x}_i$ and w^k is the Lagrange multiplier.

Outline:

- **Step 1:** Invoking the fix λ strategy, we perform an **one-dimensional search** to update λ.
- **Step 2:** For the resulting problem (B), we apply the **operator splitting** technique to obtain (C).
- **Step 3:** Exploiting the specific local structure, we propose a novel LP-ADMM to solve (C), which further involves the subproblem of β-exact update (D).

Main Theorem

Suppose that $\{\beta^k,\mu^k,w^k\}_{k\ge0}$ is generated by the LP-ADMM algorithm. We have

- $\{\beta^k,\mu^k,w^k\}_{k\ge0}$ converges to a KKT point.
- The function value converges with rate $\mathcal{O}(\frac{1}{K})$.

Experiments & Results

Compare the CPU Running Time with YALMIP Solver:

Table 1: Synthetic Data & UCI Adult Data

Dataset	Samples	Features	YALMIP (s)	Ours (s)	Ratio
Synthetic	5000	100	287.67 ± 2.67	0.64 ± 0.03	**451**
Synthetic	10000	10	283.25 ± 18.98	0.50 ± 0.02	**563**
Synthetic	10000	100	1165.40 ± 26.52	1.37 ± 0.12	**852**
a1a	1605	123	25.63	2.93	**9**
a2a	2265	123	39.20	3.53	**11**
a3a	3185	123	57.79	4.26	**14**
a4a	4781	123	105.32	4.56	**23**
a5a	6414	123	155.42	4.39	**35**
a6a	11220	123	413.65	4.68	**88**
a7a	16100	123	738.12	5.41	**137**
a8a	22696	123	1396.45	5.81	**240**
a9a	32561	123	2993.30	7.08	**423**

Efficiency of LP-ADMM for β-subproblem: