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

Jiajin Li, Sen Huang, Anthony Man-Cho So
Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong
Shatin, N. T., Hong Kong
{jjli,hsen,manchoso}@se.cuhk.edu.hk

Abstract

Wasserstein distance-based distributionally robust optimization (DRO) has received much attention lately due to its ability to provide a robustness interpretation of various learning models. Moreover, many of the DRO problems that arise in the learning context admits exact convex reformulations and hence can be tackled by off-the-shelf solvers. Nevertheless, the use of such solvers severely limits the applicability of DRO in large-scale learning problems, as they often rely on general purpose interior-point algorithms. On the other hand, there are very few works that attempt to develop fast iterative methods to solve these DRO problems, which typically possess complicated structures. In this paper, we take a first step towards resolving the above difficulty by developing a first-order algorithmic framework for tackling a class of Wasserstein distance-based distributionally robust logistic regression (DRLR) problem. Specifically, we propose a novel linearized proximal ADMM to solve the DRLR problem, whose objective is convex but consists of a smooth term plus two non-separable non-smooth terms. We prove that our method enjoys a sublinear convergence rate. Furthermore, we conduct three different experiments to show its superb performance on both synthetic and real-world datasets. In particular, our method can achieve the same accuracy up to 800+ times faster than the standard off-the-shelf solver.

1 Introduction

One of the basic principles for dealing with the overfitting phenomenon in statistical learning is regularization [23]. Recently, there has been a flurry of works that aim to interpret regularization from a distributionally robust optimization (DRO) perspective; see, e.g., [19, 1, 7, 20, 18] and the references therein. The results in these works not only provide a probabilistic justification of existing regularization techniques but also offer a powerful alternative approach to tackle risk minimization problems. Indeed, it has been shown that the DRO formulations of various statistical learning problems admit polynomial-time solvable and exact convex reformulations [19, 7, 2, 21, 20], which can be tackled by off-the-shelf solvers (e.g., YALMIP). Nevertheless, the use of such solvers severely limits the applicability of the DRO approach in large-scale learning problems, as they often rely on general-purpose interior-point algorithms. On the other hand, there are very few works that address the design of fast iterative methods for solving the convex reformulations of DRO problems. This is in part due to the complicated structures that are often possessed by such reformulations. In fact, it is only recently that researchers have proposed stochastic gradient descent (SGD) algorithms for DRO with f-divergence-based ambiguity sets [17]. However, f-divergence measures can only compare distributions with the same support, while the Wasserstein distances do not have such a restriction. On another front, the works [15, 11] propose cutting-surface methods to deal with Wasserstein distance-based DRO problems. However, they tend to suffer a large computational burden.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

In this paper, we take a first step towards bridging the above-mentioned gap by proposing a new first-order algorithmic framework for solving the class of Wasserstein distance-based distributionally robust logistic regression (DRLR) problems considered in [19]. The starting point of our investigation is the following reformulation result; see Theorem 1 and Remark 2 in [19]:

$$\begin{aligned} \inf_{\beta} \sup_{\mathbb{Q}\in B_\epsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[\ell_\beta(x,y)] \triangleq \inf_{\beta,\,\lambda}\ & \lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N} \left(\ell_\beta(\hat{x}_i,\hat{y}_i) + \max\{\hat{y}_i\beta^T\hat{x}_i - \lambda\kappa, 0\}\right) \\ \text{s.t.}\ & \|\beta\|_* \le \lambda. \end{aligned} \tag{1.1}$$

Here, $x \in \mathbb{R}^n$ denotes a feature vector and $y \in \{-1,+1\}$ its associated label to be predicted; $\ell_\beta(x,y) = \log(1+\exp(-y\beta^T x))$ is the log-loss associated with the feature-label pair (x,y) and regression parameter $\beta \in \mathbb{R}^n$; $\{(\hat{x}_i,\hat{y}_i)\}_{i=1}^N$ are N training samples drawn from an unknown underlying distribution $\mathbb{P}^*$ on the feature-label space $\Theta = \mathbb{R}^n \times \{-1,+1\}$; $\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{(\hat{x}_i,\hat{y}_i)}$ denotes the empirical distribution associated with the training samples $\{(\hat{x}_i,\hat{y}_i)\}_{i=1}^N$; $B_\epsilon(\hat{\mathbb{P}}_N) = \{\mathbb{Q}\in\mathcal{P}(\Theta) : W(\mathbb{Q},\hat{\mathbb{P}}_N) \le \epsilon\}$ is the ball in the space $\mathcal{P}(\Theta)$ of probability distributions on Θ that is centered at the empirical distribution $\hat{\mathbb{P}}_N$ and has radius ϵ with respect to the Wasserstein distance

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

where $\xi = (x,y) \in \Theta$, $d(\xi,\xi') = \|x-x'\| + \frac{\kappa}{2}|y-y'|$ is the transport cost between two data points $\xi,\xi' \in \Theta$ induced by a generic norm $\|\cdot\|$ on $\mathbb{R}^n$ with $\|\cdot\|_*$ being its dual norm, and $\kappa > 0$ is a parameter that represents the reliability of the label measurements (the larger the κ, the more reliable are the measurements; when $\kappa = \infty$, the measurements are error-free). The formulation on the left-hand side of (1.1) is motivated by the desire to construct an ambiguity set around the empirical distribution $\hat{\mathbb{P}}_N$ that contains the true distribution $\mathbb{P}^*$, so that the resulting classifier has good out-of-sample performance. We refer the reader to [19] for a more detailed discussion.

A natural question that arises from (1.1) is how to solve the convex optimization problem on the right-hand side (RHS) efficiently. When $\kappa = \infty$, the RHS of (1.1) reduces to a classic regularized logistic regression problem [19, Remark 1]. As such, a host of practically efficient first-order methods (such as proximal gradient-type methods or stochastic (variance-reduced) gradient methods) with provable convergence guarantees (see, e.g., [22, 24, 28, 2]) can be applied. However, the algorithmic aspects of the practically more relevant case where $\kappa < \infty$ have not been well explored. Our proposed framework for tackling this case consists of two steps. First, by considering the optimality conditions of the RHS of (1.1), we can derive an upper bound λ^U on the optimal λ^*. This suggests that we can first initialize λ to a value in $[0,\lambda^U]$ and solve the resulting problem that involves only the variable β (the β-subproblem), then apply golden-section search to update λ, and then repeat the whole process until we find the optimal solution to (1.1). Second, which is the core step of our framework, is to design a fast iterative method for solving the β-subproblem. By treating λ as a constant, the RHS of (1.1) is equivalent to

$$\inf_{\|\beta\|_*\le\lambda} \frac{1}{N}\sum_{i=1}^{N} \left(h(\hat{y}_i\beta^T\hat{x}_i) + \max\{\hat{y}_i\beta^T\hat{x}_i - \lambda\kappa, 0\}\right) \tag{1.2}$$

with $h(u) = \log(1+\exp(-u))$. Although (1.2) has a relatively simple norm-ball constraint, its objective is non-smooth and non-separable. As such, most existing first-order methods (e.g., projected/proximal subgradient methods) are ill-suited for tackling it. To proceed, we apply the operator splitting technique to reformulate (1.2) as

$$\begin{aligned} \inf_{\beta,\,\mu}\quad & \frac{1}{N}\sum_{i=1}^{N} \left(h(\mu_i) + \max\{\mu_i - \lambda\kappa, 0\}\right) \\ \text{s.t.}\quad & Z\beta - \mu = 0,\ \|\beta\|_* \le \lambda, \end{aligned} \tag{1.3}$$

where Z is the $N \times n$ matrix whose i-th row is $\hat{y}_i\hat{x}_i^T$, and propose a new linearized proximal alternating direction method of multipliers (LP-ADMM) to fully exploit the structure of (1.3). In particular, our method differs substantially from the commonly used ADMM-variants in the literature [27, 14, 6, 12, 8, 9, 25] in the updates of the variables. For the β-update, we solve a norm-constrained quadratic optimization problem. Since such a problem can be rather ill-conditioned, we provide three different types of solvers to handle this task, namely, the accelerated projected gradient descent, coordinate minimization [10], and active set conjugate gradient methods [5]. For the μ-update, observing that the coupling matrix for μ in the linear equality constraint is the identity, the augmented Lagrangian function is already locally strongly convex in μ. Hence, instead of using a quadratic approximation of $h(\cdot)$ as in the vanilla proximal ADMM, we use a first-order approximation without step size selection; i.e.,

$$\mu^{k+1} = \arg\min_{\mu} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Big(h'(\mu_i^k)\mu_i + \max\{\mu_i - \lambda\kappa, 0\} \Big) - (w^k)^T (Z\beta^{k+1} - \mu) + \frac{\rho}{2} \|\mu - Z\beta^{k+1}\|_2^2 \right\},$$

where $w \in \mathbb{R}^N$ is the dual variable associated with the linear equality constraint in (1.3) and $\rho > 0$ is the penalty parameter in the augmented Lagrangian function. On the theoretical side, we prove that our proposed LP-ADMM enjoys an $\mathcal{O}(\frac{1}{K})$ convergence rate under standard assumptions. On the numerical side, we demonstrate via extensive experiments that our proposed method can be sped up substantially by adopting a geometrically increasing step size strategy. In particular, our method can achieve a hundred-fold speedup over the standard solver (which is the only other method that has been used so far to solve (1.1)) on both synthetic and real-world datasets without the need to tune an optimal penalty parameter in every iteration. To the best of our knowledge, our work is the first to propose a first-order algorithmic framework for solving the Wasserstein distance-based DRLR problem (1.1) for any $\kappa > 0$. Moreover, the proposed framework is sufficiently general that it can potentially be applied to other DRO problems, which could be of independent interest.

2 Preliminaries

Let us introduce some basic definitions and concepts. To allow for greater generality, consider the following problem:

$$\begin{aligned} \underset{x,y}{\text{minimize}} \quad & F(x,y) = f(y) + P(y) + g(x) \\ \text{subject to} \quad & Ax - y = 0. \end{aligned} \tag{2.1}$$

Here, $f : \mathbb{R}^N \to \mathbb{R}$ is a closed convex function that is continuously differentiable on $\operatorname{int}(\operatorname{dom}(f))$ with linear operator $A \in \mathbb{R}^{N \times n}$; $P : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is a closed proper convex function; $g(x)$ is the indicator function of a norm ball. It should be clear that problem (2.1) includes the β-subproblem (1.3) as a special case. Indeed, the latter can be written as

$$\begin{aligned} \underset{\mu,\beta}{\text{minimize}} \quad & F(\mu,\beta) = f(\mu) + P(\mu) + g(\beta) \\ \text{subject to} \quad & Z\beta - \mu = 0, \end{aligned}$$

where $f(\mu) = \frac{1}{N}\sum_{i=1}^{N} \left\{ \log(1+\exp(-\mu_i)) + \frac{1}{2}(\mu_i - \lambda\kappa) \right\}$, $P(\mu) = \frac{1}{2N}\sum_{i=1}^{N} |\mu_i - \lambda\kappa|$, and $g(\beta) = \mathbb{I}_{\{\|\beta\|_* \le \lambda\}}$. Now, the augmented Lagrangian function associated with (2.1) is given by

$$\mathcal{L}_\rho(x,y;w) = f(y) + P(y) + g(x) - w^T(Ax - y) + \frac{\rho}{2}\|Ax - y\|_2^2, \tag{2.2}$$

where w is the multiplier. We use $(\mathcal{X}^*, \mathcal{Y}^*)$ to denote the solution set of (2.1). A point (x^*, y^*) is optimal for (2.1) if there exists a w^* such that the following KKT conditions are satisfied:

$$\begin{cases} A^T w^* \in \partial g(x^*), \\ -w^* \in \nabla f(y^*) + \partial P(y^*), \\ Ax^* - y^* = 0. \end{cases} \tag{2.3}$$

Assumption 2.1. *There exists a point (x^*, y^*, w^*) satisfying the KKT conditions in* (2.3).

Assumption 2.2. *The gradient of the function f is Lipschitz continuous; i.e., there exists a constant $L_f > 0$ such that*

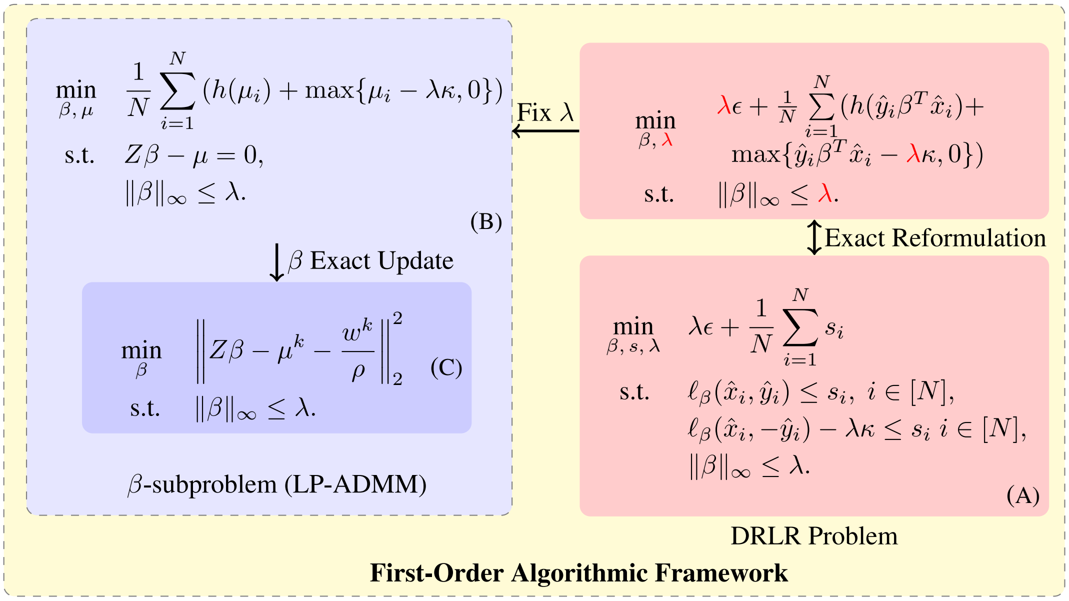
$$\|\nabla f(x) - \nabla f(y)\| \le L_f \|x - y\|, \forall x, y.$$

Definition 2.3 (Bregman Divergence)**.** *Let $f : \Omega \to \mathbb{R}$ be a function that is a) strictly convex, b) continuously differentiable, and c) defined on a closed convex set Ω. The Bregman divergence with respect to f is defined as*

$$B_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

3 First-Order Algorithmic Framework

In this section, we present our first-order algorithmic framework for solving the DRLR problem. For concreteness' sake, we take $\|\cdot\|$ in the transport cost to be the ℓ_1-norm in this paper. However, it should be mentioned that our framework is general enough to handle other norms as well.



$\min_{\beta,\mu} \; \frac{1}{N}\sum_{i=1}^{N}(h(\mu_i)+\max\{\mu_i-\lambda\kappa,0\})$
s.t. $Z\beta-\mu=0,$
$\|\beta\|_\infty \le \lambda.$
(B)
Fix λ
$\min_{\beta,\lambda} \; \lambda\epsilon+\frac{1}{N}\sum_{i=1}^{N}(h(\hat{y}_i\beta^T\hat{x}_i)+\max\{\hat{y}_i\beta^T\hat{x}_i-\lambda\kappa,0\})$
s.t. $\|\beta\|_\infty \le \lambda.$
Exact Reformulation
β Exact Update
$\min_{\beta} \; \left\|Z\beta-\mu^k-\frac{w^k}{\rho}\right\|_2^2$
s.t. $\|\beta\|_\infty \le \lambda.$
(C)
β-subproblem (LP-ADMM)
$\min_{\beta,s,\lambda} \; \lambda\epsilon+\frac{1}{N}\sum_{i=1}^{N}s_i$
s.t. $\ell_\beta(\hat{x}_i,\hat{y}_i)\le s_i,\ i\in[N],$
$\ell_\beta(\hat{x}_i,-\hat{y}_i)-\lambda\kappa\le s_i\ i\in[N],$
$\|\beta\|_\infty \le \lambda.$
(A)
DRLR Problem
First-Order Algorithmic Framework

Figure 1: First-order algorithmic framework for Wasserstein DRLR with ℓ_1-induced transport cost

We summarize the key components of our first-order algorithmic framework in Figure 1. As shown in [19], the original DRLR problem (i.e., LHS of (1.1)) can be reformulated as the convex program (A) using strong duality. A standard approach to tackling problem (A) is to use an off-the-shelf solver (e.g., YALMIP). To develop an efficient algorithmic framework, we focus on the RHS of (1.1) and proceed in two steps. Motivated by the structure of the RHS of (1.1), a natural first step is to fix λ to a certain value to obtain the problem (1.2), which involves only the variable β and will be referred to as the β-subproblem in the sequel. The second, which is also the core step of our framework, is to design a fast iterative algorithm to tackle the β-subproblem (1.2). The main difficulty of problem (1.2) comes from the two non-smooth non-separable terms. To overcome this difficulty, we introduce the auxiliary variable $\mu_i = \hat{y}_i\beta^T\hat{x}_i$ to split the non-separable non-smooth term $\max\{\hat{y}_i\beta^T\hat{x}_i - \lambda\kappa, 0\}$, thus leading to problem (B). Then, we propose a novel linearized proximal ADMM (LP-ADMM) algorithm to solve it efficiently. As will be shown in Section 4, the proposed LP-ADMM will converge at the rate $\mathcal{O}(1/K)$ when applied to the β-subproblem (B). In each iteration of our LP-ADMM algorithm, we perform an exact minimization for the β-update, which entails solving the box-constrained quadratic optimization problem (C) (here, w^k denotes the corresponding Lagrange multiplier). Towards that end, we provide three alternative solvers for problem (C), which target three different settings. Specifically, we use accelerated projected gradient descent in the well-conditioned case; coordinate minimization [10] in the high-dimensional case $N \ll d$; active set conjugate gradient method [5] in the ill-conditioned case. The details are given in Appendix B.

To implement the above framework, let us first show that there is a finite upper bound λ^U on the optimal λ^* to problem (A). Observe that the objective function in the RHS of (1.1) takes the form

$$\Omega(\lambda,\beta) = \lambda\epsilon + \frac{1}{N}\sum_{i=1}^{N}\left(h(\hat{y}_i\beta^T\hat{x}_i) + \max\{\hat{y}_i\beta^T\hat{x}_i - \lambda\kappa, 0\}\right) + \mathbb{I}_{\{\|\beta\|_\infty\le\lambda\}}.$$

Now, let $q(\lambda) = \inf_\beta \Omega(\lambda,\beta)$. As the function $\Omega(\cdot,\cdot)$ is jointly convex, we can conclude that $q(\cdot)$ is a convex (and hence unimodal) function on $\mathbb{R}$. Furthermore, the DRLR problem (A) satisfies the Mangasarian-Fromovitz constraint qualification (MFCQ), which implies that its KKT conditions are necessary and sufficient for optimality. As the following proposition shows, we can use the KKT system of problem (A) to derive the desired upper bound on λ^*:

Proposition 3.1. *Suppose that* $(\beta^*, \lambda^*, s^*)$ *is an optimal solution to problem* (A) *in Figure 1. Then, we have* $\lambda^* \leq \lambda^U = \frac{0.2785}{\epsilon}$.

Proof. Using a_{ij}, where $i = 1, \ldots, N$ and $j = 1, \ldots, 4$ to denote the multipliers associated with the constraints in problem (A), we can write down the KKT conditions of problem (A) as follows:

$$
\begin{array}{rl}
\min\limits_{\beta,\, s,\, \lambda} & \lambda\epsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\
\text{s.t.} & \ell_\beta(\hat{x}_i, \hat{y}_i) \leq s_i, \\
& \quad i \in [N], \\
& \ell_\beta(\hat{x}_i, -\hat{y}_i) - \lambda\kappa \leq s_i, \\
& \quad i \in [N], \\
& e_i^T\beta \leq \lambda,\ i \in [N], \\
& -e_i^T\beta \leq \lambda,\ i \in [N]
\end{array}
\Rightarrow
\begin{cases}
\displaystyle\sum_{i=1}^{N} a_{i1}\nabla_\beta \ell_\beta(\hat{x}_i, \hat{y}_i) + a_{i2}\nabla_\beta \ell_\beta(\hat{x}_i, -\hat{y}_i) + (a_{i3} - a_{i4})e_i = 0, \\
\displaystyle\sum_{i=1}^{N} \kappa a_{i2} + a_{i3} + a_{i4} = \epsilon, \\
a_{i1} + a_{i2} = \dfrac{1}{N},\ i \in [N], \\
a_{i1}(\ell_\beta(\hat{x}_i, \hat{y}_i) - s_i) = 0,\ i \in [N], \\
a_{i2}(\ell_\beta(\hat{x}_i, -\hat{y}_i) - \lambda\kappa - s_i) = 0,\ i \in [N], \\
a_{i3}(e_i^T\beta - \lambda) = 0,\ i \in [N], \\
a_{i4}(e_i^T\beta + \lambda) = 0,\ i \in [N], \\
a_{ij} \geq 0,\ i \in [N],\ j \in [4].
\end{cases}
$$

After some elementary manipulations (see Appendix A for details), we obtain

$$
\lambda \leq \frac{1}{N\epsilon}\sum_{i=1}^{N} \frac{\hat{y}_i\beta^T\hat{x}_i \exp(-\hat{y}_i\beta^T\hat{x}_i)}{1 + \exp(-\hat{y}_i\beta^T\hat{x}_i)} \leq \frac{0.2785}{\epsilon},
$$

as desired. □

Remark 3.2. *Although Proposition 3.1 applies to the case where the transport cost is induced by the* ℓ_1*-norm, the techniques used to prove it can also carry over to the* ℓ_2 *and* ℓ_∞ *cases. All we need to do is to modify the parts highlighted in blue above. Indeed, when the transport cost is induced by the* ℓ_2*-norm, the norm constraint in problem* (A) *becomes* $\|\beta\|_2 \leq \lambda$*, which is equivalent to* $\|\beta\|_2^2 \leq \lambda^2$*. On the other hand, when the transport cost is induced by the* ℓ_∞*-norm, the norm constraint becomes* $\|\beta\|_1 \leq \lambda$*, which can be expressed as* $B\beta \leq \lambda e_{2^n}$ *with* B *being the* $2^n \times n$ *matrix whose rows are all the possible arrangements of* $+1$*'s and* $-1'$*s.*

Proposition 3.1, together with the unimodality of $q(\cdot)$, suggests the following natural strategy for finding an optimal solution $(\beta^*, \lambda^*, s^*)$ to problem (A): initialize λ in (A) to a value in $[0, \lambda^U]$, solve the resulting β-subproblem (B), apply golden-section search to update λ, and repeat. The pseudo-code for the golden-section search on λ can be found in Appendix B. The β-subproblem (B) will be solved by our proposed LP-ADMM, which we present next.

4 LP-ADMM for the β-Subproblem and Its Convergence Analysis

To simplify notation, we consider the prototypical form (2.1) of the β-subproblem here. It can be shown that the β-subproblem (B) satisfies Assumptions 2.1 and 2.2 in Section 2. Now, we present our proposed LP-ADMM in Algorithm 1.

The x-update is standard in ADMM-type algorithms and leads to a box-constrained quadratic optimization problem. The crux of our algorithm lies in the local model used to perform the y-update. To understand the local model, observe that since the coupling matrix for y in the constraint $Ax - y = 0$ is the identity, the augmented Lagrangian function $\mathcal{L}_\rho(\cdot, \cdot; \cdot)$ in (2.2) is strongly convex in y. Thus, instead of using the quadratic approximation of $f(\cdot)$ as in the vanilla proximal ADMM, we can use the first-order approximation $y \mapsto \hat{f}(y; y^k) = f(y^k) + \nabla f(y^k)^T(y - y^k)$ at the current iterate y^k. This leads to the y-update

$$
y^{k+1} = \arg\min_{y} \left\{ \hat{f}(y; y^k) - \langle w^k, Ax^{k+1} - y \rangle + \frac{\rho}{2}\|y - Ax^{k+1}\|_2^2 + P(y) \right\},
$$

which, as can be easily verified, is equivalent to the update given in Algorithm 1. In fact, using such a first-order local model not only makes the resulting algorithm converge faster in practice but also eliminates the need to perform step size selection. The latter makes our algorithmic framework numerically more robust in general.

Algorithm 1: Linearized Proximal ADMM (LP-ADMM) for Solving (2.1)

```
Input: Choose initial point $(x^0, y^0, w^0) \in \mathbb{R}^n \times \mathbb{R}^N \times \mathbb{R}^N$ and number of iterations $K$;
       Initialized the penalty parameter $\rho_0$ and shrinking parameter $\gamma \geq 1$;
Output: $\{(x^k, y^k, w^k)\}_{k=1}^K$ and $\{F(x^k, y^k)\}_{k=1}^K$;
for each iteration do
    $x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ \frac{\rho_k}{2} \left\| Ax - y^k - \frac{w^k}{\rho_k} \right\|_2^2 + g(x) \right\};$
    $y^{k+1} = \arg\min_{y \in \mathbb{R}^N} \left\{ \frac{\rho_k}{2} \left\| y - \left( Ax^{k+1} - \frac{w^k + \nabla f(y^k)}{\rho_k} \right) \right\|_2^2 + P(y) \right\};$
    $w^{k+1} = w^k - \rho_k(Ax^{k+1} - y^{k+1});$
    $\rho_{k+1} = \gamma \rho_k$ (in particular, if $\gamma = 1$, then $\rho_k = \rho_{k+1} = \rho$);
end
```

Next, let us analyze the convergence behavior of the LP-ADMM. Based on the definition of the augmented Lagrangian function in (2.2), the optimality conditions of the subproblems in Algorithm 1 can be written as follows:

$$0 \in \rho A^T \left(Ax^{k+1} - y^k - \frac{w^k}{\rho} \right) + \partial g(x^{k+1}), \tag{4.1}$$

$$0 \in \nabla f(y^k) + \rho \left(y^{k+1} - Ax^{k+1} + \frac{w^k}{\rho} \right) + \partial P(y^{k+1}). \tag{4.2}$$

Using (4.1) and (4.2), we can establish the following basic properties concerning the iterates of our proposed LP-ADMM. The proofs can be found in Appendix A.

Proposition 4.1. *Suppose that we use a constant penalty parameter ρ that satisfies $\rho > (\sqrt{3}+1)L_f$. Let $\{(x^k, y^k, w^k)\}_{k \geq 0}$ be the sequence generated by the LP-ADMM and (x^*, y^*, w^*) be a point satisfying the KKT conditions* (2.3) *with $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$. Then, the following hold:*

(a) For all $k \geq 1$, $\|Ax^{k+1} - y^k\|_2^2 \geq \frac{1}{2}\|y^{k+1} - y^k\|_2^2 - \frac{L_f{}^2}{\rho^2}\|y^k - y^{k-1}\|_2^2$.

(b) For all $k \geq 0$ and (x, y) satisfying $Ax - y = 0$, we have $F(x^{k+1}, y^{k+1}) - F(x, y) \leq \frac{1}{2\rho}(\|w^k\|_2^2 - \|w^{k+1}\|_2^2) + \frac{\rho}{2}(\|y^k - y\|_2^2 - \|y^{k+1} - y\|_2^2) + c(\|y^k - y^{k-1}\|_2^2 - \|y^{k+1} - y^k\|_2^2) + (B_f(y, y^{k+1}) - B_f(y, y^k))$, where $c = \frac{\rho - 2L_f}{4}$.

(c) The sequence $\{\frac{1}{2\rho}\|w^k - w^*\|_2^2 + \frac{\rho}{2}\|y^k - y^*\|_2^2 - B_f(y^*, y^k)\}_{k \geq 0}$ is non-increasing and bounded below.

Armed with Proposition 4.1, we can prove the main convergence theorem for LP-ADMM.

Theorem 4.2. *Consider the setting of Proposition 4.1. Set $\bar{x}^K = \frac{1}{K}\sum_{k=1}^K x^k$ and $\bar{y}^K = \frac{1}{K}\sum_{k=1}^K y^k$. Then, the following hold:*

(a) The sequence $\{(x^k, y^k, w^k)\}_{k \geq 0}$ converges to a KKT point of problem (2.1).

(b) The sequence of function values converges at the rate $\mathcal{O}(\frac{1}{K})$:

$$F(\bar{x}^K, \bar{y}^K) - F(x^*, y^*) \leq \frac{(1/2\rho)\|w^0\|_2^2 + (\rho/2)\|y^* - y^0\|_2^2 + c\|y^0 - y^1\|_2^2}{K} = \mathcal{O}(\frac{1}{K}).$$

Remark 4.3. *The standard linearized ADMM in [14, 26, 25] involves the quadratic term $\frac{\eta}{2}\|y - y^k\|^2$, where η needs to satisfy $\eta > L_f$. Our LP-ADMM can be regarded as a linearized ADMM with $\eta = 0$. Using the first-order local model, the LP-ADMM achieves the fastest single-step update. Moreover, it is worth noting that the adaptive penalty strategy works well in practice, especially the geometrical increasing one (i.e., the blue line in Algorithm 1).*

5 Experiment Results

In this section, we present numerical results to demonstrate the effectiveness and efficiency of the different components in our proposed algorithmic framework. All experiments were conducted using MATLAB R2018a on a computer running Windows 10 with Intel® Core™ i5-8600 CPU (3.10 GHz) and 16 GB RAM. We conducted three different experiments to validate our theoretical results and show the high efficiency of our implementation of the proposed first-order algorithmic framework. To begin, we compare the CPU time of our framework with the YALMIP solver used by [19] on both synthetic and real datasets. Then, we present an empirical comparison of our LP-ADMM with other baseline first-order algorithms, including Projected SubGradient Method (SubGradient), Primal-Dual Hybrid Gradient (PDHG), Linearized-ADMM and Standard ADMM, on the β-subproblem. Lastly, we show the test data performance of the DRLR model on real datasets. We use the active set conjugate gradient method to solve the box-constrained quadratic optimization problem (C) in this section. Our code is available at `https://github.com/gerrili1996/DRLR_NIPS2019_exp`.

5.1 CPU Time Comparison with the YALMIP Solver

Our setup for the synthetic experiments is as follows. We first generate β from the standard n-dimensional Gaussian distribution $\mathcal{N}(0, I_n)$ and normalize it to obtain the ground truth $\beta_* = \beta/\|\beta\|$. Next, we generate the feature vectors $\{\hat{x}_i\}_{i=1}^N$ independently and identically (i.i.d) from $\mathcal{N}(0, I_n)$ and the noisy measurements $\{z_i\}_{i=1}^N$ i.i.d from the uniform distribution over $[0, 1]$. Lastly, we compute the ground truth labels $\{\hat{y}_i\}_{i=1}^N$ via $\hat{y}_i = 2 \times \text{int}(z_i < \frac{1}{1+\exp(-\beta_*^T \hat{x}_i)}) - 1$. We set the DRLR model parameters to be $\kappa = 1, \epsilon = 0.1$ and the default parameters of our Adaptive LP-ADMM to be $\rho_0 = 0.001, \gamma = 1.05$. All the experiment results reported here were averaged over 30 independent trials over random seeds. Table 1 summarizes the comparison of CPU times on different scales in the synthetic setting. Our experiment results indicate that the proposed LP-ADMM with adaptive penalty strategy can be over 800 times faster than YALMIP, a state-of-the-art optimization solver, and the performance gap grows considerably with problem size.

Table 1: CPU time comparison: LP-ADMM vs. YALMIP (used in [19]) in the synthetic setting

(N, d)	YALMIP (s)	Non-Adaptive (s)	Adaptive (s)	Ratio
(10,3)	2.40 ± 0.18	0.06 ± 0.02	0.07 ± 0.02	**37**
(100,3)	3.29 ± 0.05	0.14 ± 0.03	0.06 ± 0.01	**54**
(100,10)	3.34 ± 0.03	0.21 ± 0.03	0.08 ± 0.01	**44**
(500,10)	7.92 ± 0.17	0.58 ± 0.16	0.14 ± 0.01	**55**
(500,50)	8.53 ± 0.17	0.60 ± 0.03	0.24 ± 0.01	**36**
(1000,50)	16.44 ± 0.44	0.96 ± 0.07	0.25 ± 0.02	**67**
(1000,100)	19.16 ± 0.48	1.69 ± 0.11	0.38 ± 0.01	**51**
(3000,50)	65.87 ± 1.54	2.40 ± 0.15	0.32 ± 0.01	**206**
(3000,100)	113.94 ± 2.05	3.84 ± 0.20	0.47 ± 0.04	**243**
(5000,100)	287.67 ± 2.67	7.08 ± 0.66	0.64 ± 0.03	**451**
(10000,10)	283.25 ± 18.98	5.03 ± 0.76	0.50 ± 0.02	**563**
(10000,100)	1165.40 ± 26.52	19.75 ± 3.74	1.37 ± 0.12	**852**

We also tested our proposed method on the real datasets `a1a-a9a` downloaded from LIBSVM[1]. Note that the data matrices from these datasets are ill-conditioned and highly sparse, which should be contrasted with the well-conditioned and dense ones in the synthetic setting. Table 2 shows the comparison of CPU times on the real datasets. We observe that our methods work exceptionally well, especially in the large-scale case (i.e., `a9a`).

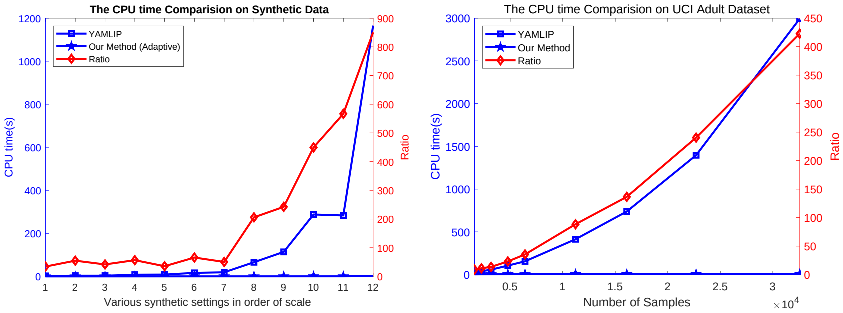
As standard ADMM-type algorithms are very sensitive to the choice of penalty parameters, it is hard for us to tune the optimal penalty parameter for each subproblem with different λ. Thus, we use a constant penalty parameter for the non-adaptive case. In fact, since we do not need to perform a careful penalty parameter selection in our method, we can achieve an even greater speed by using an adaptive penalty strategy. Moreover, it is worth noticing that our approaches achieve higher-accuracy solutions compared with the YALMIP solver in all the experiments.

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`

Table 2: CPU time comparison: LP-ADMM vs. YALMIP (used in [19]) on UCI adult datasets

Dataset	Data statistics		CPU time (s)		Ratio
	Samples	Features	YALMIP	Ours	
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**

Figure 2: CPU time comparison with YALMIP using interior point algorithm



5.2 Efficiency of LP-ADMM for β-Subproblem

To further demonstrate the efficacy of our proposed LP-ADMM on the β-subproblem, we present an empirical comparison of our algorithm with other first-order methods in the synthetic setting. The implementation details are given as follows: λ is regarded as a constant (i.e., $\lambda = 0.1$), the DRLR model parameters are the same as in Section 5.1, and the first-order methods used include

(a) Two-block Standard ADMM (SADMM) [3]: for both β- and μ-updates, we perform exact minimization, which are done using accelerated projected gradient descent and semi-smooth Newton method [13], respectively (pseudo-codes are given in Appendix B);

(b) Primal-Dual Hybrid Gradient (PDHG) [4];

(c) Linearized-ADMM (LADMM): all the ingredients are the same as LP-ADMM, except that the μ-update involves the classic quadratic term;

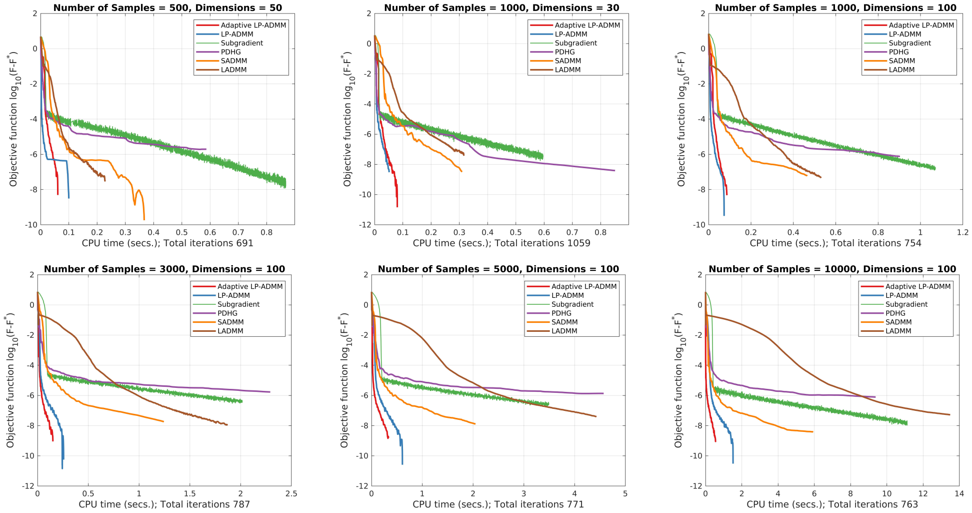
(d) Projected Subgradient Method (SubGradient).

The convergence curves for various synthetic cases are shown in Figure 3. The performance of our methods significantly dominates those of other methods, which agrees with our theoretical findings in Section 4. Compared with LADMM, we show practical advantages of the first-order local model for the μ-update. In addition, LP-ADMM and Adaptive LP-ADMM have similar performance in small instances, but the latter has better performance in large instances. In summary, we have demonstrated that the usefulness and efficiency of all the components in our first-order algorithmic framework. In particular, as the data matrices in the real datasets are ill-conditioned, all the baseline approaches cannot achieve a high accuracy but our proposed LP-ADMM can.

5.3 Test Data Performance of the DRLR Model

In this subsection, we compare the test data performance of the DRLR model with two classic models, namely, Logistic Regression (LR) and Regularized Logistic Regression (RLR). The latter refers to $\min \frac{1}{N} \sum_{i=1}^{N} \ell_\beta(\hat{x}_i, \hat{y}_i) + \epsilon \|\beta\|_\infty$. If the training data labels are error-free (which corresponds to $\kappa = \infty$), then the DRLR model reduces to RLR [19]. We use grid search with cross-validation to

Figure 3: Comparison of LP-ADMM with other first-order methods on β-subproblem: y-axis is the sub-optimality gap $\log(F^k - F^*)$; "total iterations" refers to that taken by Adaptive ADMM



select the parameters of the DRLR model (i.e., $\epsilon = 0.3, \kappa = 7$). In addition, we randomly select 60% of the data to train the models and the rest to test the performance. As before, all the results reported here were averaged over 30 independent trials. Table 3 shows the average classification accuracy on the test data. We observe that the DRLR model consistently outperforms the two classic models over all datasets. Thus, the distributionally robust optimization approach opens a new door for ameliorating the poor test performance in practice.

Table 3: Average classification accuracy on test datasets

Dataset	LR	RLR ($\kappa = \infty$)	DRLR ($\kappa = 7$)
a1a	83.13%	83.82%	**84.01%**
a2a	83.68%	83.93%	**84.24%**
MNIST(0 vs 3)	99.15%	99.45%	**99.55%**
MNIST(0 vs 4)	99.39%	99.54%	**99.75%**
MNIST(0 vs 6)	97.88%	98.86%	**98.92%**
MNIST(2 vs 3)	96.87%	97.31%	**97.40%**
MNIST(2 vs 5)	96.67%	97.45%	**97.77%**
MNIST(5 vs 8)	94.80%	94.91%	**95.18%**
MNIST(5 vs 9)	97.21%	98.11%	**98.47%**
MNIST(6 vs 9)	99.54%	99.59%	**99.80%**

6 Conclusion and Future Work

In this paper, we have proposed a first-order algorithmic framework to solve a class of Wasserstein distance-based distributionally robust logistic regression (DRLR) problem. The core step of our framework is the efficient solution of the β-subproblem. Towards that end, we have developed a novel ADMM-type algorithm (the LP-ADMM) and established its sublinear rate of convergence. We have also conducted extensive experiments to verify the practicality of our framework. It is worth noting that problem (1.2) actually enjoys the Luo-Tseng error bound property when the transport cost is induced either by the ℓ_1-norm or ℓ_∞-norm [16]. However, this does not immediately imply the linear convergence of our proposed LP-ADMM, as the method involves both primal and dual updates. Thus, it is interesting to see whether our proposed LP-ADMM or some other practically efficient first-order methods can provably achieve a linear rate of convergence when applied to the β-subproblem.

References

[1] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein Profile Inference and Applications to Machine Learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[2] Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal Transport Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes. *arXiv preprint arXiv:1810.02403*, 2018.

[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[4] Antonin Chambolle and Thomas Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[5] Wanyou Cheng, Qunfeng Liu, and Donghui Li. An Accurate Active Set Conjugate Gradient Algorithm with Project Search for Bound Constrained Optimization. *Optimization Letters*, 8(2):763–776, 2014.

[6] Wei Deng and Wotao Yin. On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[7] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein Distributional Robustness and Regularization in Statistical Learning. *arXiv preprint arXiv:1712.06050*, 2017.

[8] Xiang Gao and Shu-Zhong Zhang. First-Order Algorithms for Convex Optimization with Nonseparable Objective and Coupled Constraints. *Journal of the Operations Research Society of China*, 5(2):131–159, 2017.

[9] Mingyi Hong and Zhi-Quan Luo. On the Linear Convergence of the Alternating Direction Method of Multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

[10] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A Dual Coordinate Descent Method for Large–Scale Linear SVM. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pages 408–415, 2008.

[11] Changhyeok Lee and Sanjay Mehrotra. A Distributionally-Robust Approach for Finding Support Vector Machines. Manuscript, available at `http://www.optimization-online.org/DB_HTML/2015/06/4965.html`, 2015.

[12] Min Li, Defeng Sun, and Kim-Chuan Toh. A Majorized ADMM with Indefinite Proximal Terms for Linearly Constrained Convex Composite Optimization. *SIAM Journal on Optimization*, 26(2):922–950, 2016.

[13] Xudong Li, Defeng Sun, and Kim-Chuan Toh. A Highly Efficient Semismooth Newton Augmented Lagrangian Method for Solving Lasso Problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018.

[14] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation. In *Advances in Neural Information Processing Systems*, pages 612–620, 2011.

[15] Fengqiao Luo and Sanjay Mehrotra. Decomposition Algorithm for Distributionally Robust Optimization Using Wasserstein Metric with an Application to a Class of Regression Models. *European Journal of Operational Research*, 278(1):20–35, 2019.

[16] Zhi-Quan Luo and Paul Tseng. Error Bounds and Convergence Analysis of Feasible Descent Methods: A General Approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[17] Hongseok Namkoong and John C Duchi. Stochastic Gradient Methods for Distributionally Robust Optimization with f-Divergences. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2016.

[18] Hongseok Namkoong and John C Duchi. Variance-based Regularization with Convex Objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.

[19] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally Robust Logistic Regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[20] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via Mass Transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[21] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying Some Distributional Robustness with Principled Adversarial Training. *arXiv preprint arXiv:1710.10571*, 2017.

[22] Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors. *Optimization for Machine Learning*. Neural Information Processing Series. MIT Press, Cambridge, Massachusetts, 2012.

[23] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer Science+Business Media, LLC, second edition, 2000.

[24] Lin Xiao and Tong Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[25] Yangyang Xu. Accelerated First-Order Primal-Dual Proximal Methods for Linearly Constrained Composite Convex Programming. *SIAM Journal on Optimization*, 27(3):1459–1484, 2017.

[26] Junfeng Yang and Xiaoming Yuan. Linearized Augmented Lagrangian and Alternating Direction Methods for Nuclear Norm Minimization. *Mathematics of computation*, 82(281):301–329, 2013.

[27] Wotao Yin. Analysis and Generalizations of the Linearized Bregman Method. *SIAM Journal on Imaging Sciences*, 3(4):856–877, 2010.

[28] Zirui Zhou and Anthony Man-Cho So. A Unified Approach to Error Bounds for Structured Convex Optimization Problems. *Mathematical Programming: Series A and B*, 165(2):689–728, 2017.

Appendix

This supplementary document is the appendix section of the paper titled "First-Order Algorithmic Framework for Wasserstein Distributionally Robust Logistic Regression". It is organized as follows. In Section A, we present several technical lemmas and complete the proof of the main convergence result for the LP-ADMM. In Section B, we give the implementation details of our algorithmic framework, including pseudo-codes of the inner solvers and baseline algorithms. In Section C, we provide additional experiment results to support our findings in this paper.

A : Technical Proof

Note that the objective function in problem (A) takes the form

$$\Omega(\lambda,\beta)=\lambda\epsilon+\frac{1}{N}\sum_{i=1}^{N}\big(h(\hat{y}_i\beta^T\hat{x}_i)+\max\{\hat{y}_i\beta^T\hat{x}_i-\lambda\kappa,0\}\big)+\mathbb{I}_{\{\|\beta\|_\infty\leq\lambda\}}.$$

As the function $\Omega(\cdot,\cdot)$ is jointly convex, it is natural to conclude that the function $\lambda\mapsto q(\lambda)=\inf_\beta\Omega(\lambda,\beta)$ is also convex (and hence unimodal) on $\mathbb{R}$. The following lemma shows that this is indeed the case.

Lemma 6.1. *Suppose that $\Omega:\mathcal{X}\times\mathcal{Y}\longrightarrow\mathbb{R}\cup\{+\infty\}$ is a convex function that is proper and bounded below. Then, the function $x\mapsto q(x)=\inf_{y\in\mathcal{Y}}\Omega(x,y)$ is also convex.*

Recall that the β-subproblem of Wasserstein DRLR with ℓ_p-induced transport cost (where $p\in\{1,2,\infty\}$) is

$$\begin{array}{ll}\min\limits_{\beta,\,s,\,\lambda} & \lambda\epsilon+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}s_i\\ \text{s.t.} & \ell_\beta(\hat{x}_i,\hat{y}_i)\leq s_i,\ i\in[N],\\ & \ell_\beta(\hat{x}_i,-\hat{y}_i)-\lambda\kappa\leq s_i,\ i\in[N],\\ & \|\beta\|_q\leq\lambda,\end{array}\tag{6.1}$$

where q satisfies $\frac{1}{p}+\frac{1}{q}=1$.

Proposition 6.2. *Suppose that (β^*,λ^*,s^*) is an optimal solution to* (6.1). *Then, we have $\lambda^*\leq\lambda^U=\frac{0.2785}{\epsilon}$.*

Proof. We analyze the case where $p=1$ and other two cases are similar. From the KKT system in the proof of Proposition 3.1, we deduce

$$\begin{aligned}0&=\sum_{i=1}^{N}a_{i1}\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i)+a_{i2}\nabla_\beta\ell_\beta(\hat{x}_i,-\hat{y}_i)+(a_{i3}-a_{i4})e_i\\&=\sum_{i=1}^{N}\frac{1}{N}\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i)+a_{i2}(\nabla_\beta\ell_\beta(\hat{x}_i,-\hat{y}_i)-\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i))+\sum_{i=1}^{N}(a_{i3}-a_{i4})e_i\\&=\sum_{i=1}^{N}(\frac{1}{N}\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i)+a_{i2}\hat{y}_i\hat{x}_i)+\sum_{i=1}^{N}(a_{i3}-a_{i4})e_i.\end{aligned}$$

Multiplying β gives

$$\begin{aligned}
0 &= \sum_{i=1}^{N}(\frac{1}{N}\beta^T\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i) + a_{i2}\hat{y}_i\beta^T\hat{x}_i) + \sum_{i=1}^{N}(a_{i3}-a_{i4})e_i^T\beta \\
&= \sum_{i=1}^{N}(\frac{1}{N}\beta^T\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i) + a_{i2}\hat{y}_i\beta^T\hat{x}_i) + \lambda\sum_{i=1}^{N}(a_{i3}+a_{i4}) \\
&= \sum_{i=1}^{N}(\frac{1}{N}\beta^T\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i) + a_{i2}\hat{y}_i\beta^T\hat{x}_i) + \lambda(\epsilon - \kappa\sum_{i=1}^{N}a_{i2}) \\
&= \sum_{i=1}^{N}(\frac{1}{N}\beta^T\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i) + a_{i2}(\hat{y}_i\beta^T\hat{x}_i - \lambda\kappa)) + \lambda\epsilon \\
&= \sum_{i=1}^{N}(\frac{1}{N}\beta^T\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i) + a_{i2}(s_i - \ell_\beta(\hat{x}_i,\hat{y}_i))) + \lambda\epsilon \\
&\geq \frac{1}{N}\sum_{i=1}^{N}\beta^T\nabla_\beta\ell_\beta(\hat{x}_i,\hat{y}_i) + \lambda\epsilon.
\end{aligned}$$

Plugging in the explicit formula for $\ell_\beta(\cdot,\cdot)$ yields

$$\lambda \leq \frac{1}{N\epsilon}\sum_{i=1}^{N}\frac{\hat{y}_i\beta^T\hat{x}_i\exp(-\hat{y}_i\beta^T\hat{x}_i)}{1+\exp(-\hat{y}_i\beta^T\hat{x}_i)}. \tag{6.2}$$

Note that $\phi(t) = \frac{te^{-t}}{1+e^{-t}} = \frac{t}{e^t+1}$ and $\phi'(t) = \frac{e^t - te^t + 1}{(e^t+1)^2}$ is strictly decreasing. Thus, $\phi(t)$ has a unique maximizer and $\phi(t) \leq 0.2785$. Therefore, $\lambda \leq \lambda^U = \frac{0.2785}{\epsilon}$. □

Convergence Analysis of LP-ADMM

Lemma 6.3. *The sequence* $\{(x^k, y^k, z^k)\}_{k\geq 0}$ *generated by the LP-ADMM satisfies*

$$\|Ax^{k+1} - y^k\|_2^2 \geq \frac{1}{2}\|y^{k+1} - y^k\|_2^2 - \frac{{L_f}^2}{\rho^2}\|y^k - y^{k-1}\|_2^2$$

for all $k \geq 1$.

Proof. Note that the y-update rule (3) takes the form

$$y^{k+1} = \arg\min_y\left\{\frac{\rho}{2}\|y - (Ax^{k+1} + \frac{w^k + \nabla f(y^k)}{\rho})\|_2^2 + P(y)\right\}, \tag{6.3}$$

which can be further rewritten as

$$y^{k+1} = \operatorname{prox}_{\frac{1}{\rho}g(.)}(Ax^{k+1} - \frac{w^k + \nabla f(y^k)}{\rho}). \tag{6.4}$$

Based on the expansiveness property of proximal operator and the y-update rule (6.4), we have

$$\begin{aligned}
\|y^{k+1} - y^k\|_2^2 &= \|\operatorname{prox}_{\frac{1}{\rho}g(.)}(Ax^{k+1} - \frac{w^k + \nabla f(y^k)}{\rho}) - \operatorname{prox}_{\frac{1}{\rho}g(.)}(Ax^k - \frac{w^{k-1} + \nabla f(y^{k-1})}{\rho})\|_2^2 \\
&\leq \|Ax^{k+1} - Ax^k + \frac{w^{k-1} - w^k}{\rho} + \frac{\nabla f(y^k) - \nabla f(y^{k-1})}{\rho}\|_2^2 \\
&\overset{(\text{♥})}{=} \|Ax^{k+1} - y^k + \frac{1}{\rho}(\nabla f(y^k) - \nabla f(y^{k-1})\|_2^2 \\
&\leq 2(\|Ax^{k+1} - y^k\|_2^2 + \frac{{L_f}^2}{\rho^2}\|y^k - y^{k-1}\|_2^2),
\end{aligned}$$

where (♥) is derived from $w^{k+1} = w^k - \rho(Ax^{k+1} - y^{k+1})$. Rearranging the inequality, we can conclude the proof. □

Lemma 6.4. *Let* $\{(x^k, y^k, w^k)\}_{k\geq 0}$ *be the sequence generated by the LP-ADMM. Set* $c = \frac{\rho - 2L_f}{4}$*. Then, for any* x, y *satisfying* $Ax - y = 0$*, we have*

$$F(x^{k+1}, y^{k+1}) - F(x, y) \leq \frac{1}{2\rho}(\|w^k\|_2^2 - \|w^{k+1}\|_2^2)$$
$$+ \frac{\rho}{2}(\|y^k - y\|_2^2 - \|y^{k+1} - y\|_2^2) + c(\|y^k - y^{k-1}\|_2^2 - \|y^{k+1} - y^k\|_2^2) + (B_f(y, y^{k+1}) - B_f(y, y^k)).$$

Proof. Using the optimality conditions (4.1) and (4.2), we have

$$-\rho A^T(Ax^{k+1} - y^k - \frac{w^k}{\rho}) \in \partial g(x^{k+1}), \tag{6.5}$$
$$\nabla f(y^{k+1}) - \nabla f(y^k) - w^{k+1} \in \nabla f(y^{k+1}) + \partial P(y^{k+1}), \tag{6.6}$$

where (6.6) is derived from $w^{k+1} = w^k - \rho(Ax^{k+1} - y^{k+1})$. By the convexity of $g(\cdot)$,

$$\begin{aligned} g(x^{k+1}) - g(x) &\leq \langle \partial g(x^{k+1}), x^{k+1} - x\rangle \\ &= -\langle \rho A^T(Ax^{k+1} - y^k - \frac{w^k}{\rho}), x^{k+1} - x\rangle \\ &= -\rho\langle Ax^{k+1} - y^{k+1} + y^{k+1} - y^k - \frac{w^k}{\rho}, A(x^{k+1} - x)\rangle \\ &= -\rho\langle \frac{1}{\rho}(w^k - w^{k+1}) + y^{k+1} - y^k - \frac{w^k}{\rho}, A(x^{k+1} - x)\rangle \\ &= \langle w^{k+1} - \rho(y^{k+1} - y^k), Ax^{k+1} - y\rangle \quad \text{(i.e., Dual Update and Dual Feasibility)} \\ &= \langle w^{k+1}, Ax^{k+1} - y\rangle + \langle \rho(y^k - y^{k+1}), Ax^{k+1} - y\rangle \\ &= \langle w^{k+1}, Ax^{k+1} - y\rangle + \frac{\rho}{2}(\|y^k - y\|_2^2 - \|y^{k+1} - y\|_2^2 + \|Ax^{k+1} - y^{k+1}\|_2^2 \\ &\quad - \|Ax^{k+1} - y^k\|_2^2), \end{aligned}$$

where the last equality holds as

$$(x_1 - x_2)^T(x_3 + x_4) = \frac{1}{2}(\|x_4 - x_2\|_2^2 - \|x_4 - x_1\|_2^2 + \|x_3 + x_1\|_2^2 - \|x_3 - x_2\|_2^2).$$

Similarly, by the convexity of the function $f(\cdot) + P(\cdot)$, we have

$$\begin{aligned} f(y^{k+1}) + P(y^{k+1}) - f(y) - P(y) &\leq \langle \nabla f(y^{k+1}) + \partial P(y^{k+1}), y^{k+1} - y\rangle \\ &= \langle \nabla f(y^{k+1}) - \nabla f(y^k) - w^{k+1}, y^{k+1} - y\rangle \\ &= \langle \nabla f(y^{k+1}) - \nabla f(y^k), y^{k+1} - y\rangle - \langle w^{k+1}, y^{k+1} - y\rangle. \end{aligned}$$

Summing the above two inequalities, we have

$$\begin{aligned} F(x^{k+1}, y^{k+1}) - F(x, y) &\leq \langle w^{k+1}, Ax^{k+1} - y^{k+1}\rangle + \frac{\rho}{2}\|Ax^{k+1} - y^{k+1}\|_2^2 \\ &+ \frac{\rho}{2}(\|y^k - y\|_2^2 - \|y^{k+1} - y\|_2^2) \\ &+ \underbrace{\langle \nabla f(y^{k+1}) - \nabla f(y^k), y^{k+1} - y\rangle - \frac{\rho}{2}\|Ax^{k+1} - y^k\|_2^2}_{(*)}. \end{aligned}$$

Note that the term $\langle w^{k+1}, Ax^{k+1} - y^{k+1}\rangle + \frac{\rho}{2}\|Ax^{k+1} - y^{k+1}\|_2^2$ can be reformulated as $\frac{1}{2\rho}(\|w^k\|_2^2 - \|w^{k+1}\|_2^2)$ via plugging in the dual update rule. In details,

$$\begin{aligned} \langle w^{k+1}, Ax^{k+1} - y^{k+1}\rangle + \frac{\rho}{2}\|Ax^{k+1} - y^{k+1}\|_2^2 &= \langle Ax^{k+1} - y^{k+1}, w^{k+1} + \frac{\rho}{2}(Ax^{k+1} - y^{k+1})\rangle \\ &= \langle \frac{w^k - w^{k+1}}{\rho}, \frac{1}{2}(w^k + w^{k+1})\rangle \\ &= \frac{1}{2\rho}(\|w^k\|_2^2 - \|w^{k+1}\|_2^2). \end{aligned}$$

Besides, it is more tricky to give an upper bound on the last term in the above gap function. Based on the three-point property of Bregman divergence, we have

$$\begin{aligned}\langle \nabla f(y^{k+1}) - \nabla f(y^k), y^{k+1} - y\rangle &= B_f(y^{k+1}, y^k) + B_f(y, y^{k+1}) - B_f(y, y^k)\\ &\leq \frac{L_f}{2}\|y^{k+1} - y^k\|_2^2 + (B_f(y, y^{k+1}) - B_f(y, y^k)).\end{aligned} \tag{6.7}$$

Applying Lemma 6.3 to the term $-\frac{\rho}{2}\|Ax^{k+1} - y^k\|_2^2$, we have

$$-\frac{\rho}{2}\|Ax^{k+1} - y^k\|_2^2 \leq \frac{{L_f}^2}{2\rho}\|y^k - y^{k-1}\|_2^2 - \frac{\rho}{4}\|y^{k+1} - y^k\|_2^2. \tag{6.8}$$

Adding (6.7) and (6.8), we can obtain

$$(*) \leq \frac{{L_f}^2}{2\rho}\|y^k - y^{k-1}\|_2^2 - (\frac{\rho}{4} - \frac{L_f}{2})\|y^{k+1} - y^k\|_2^2 + (B_f(y, y^{k+1}) - B_f(y, y^k)). \tag{6.9}$$

If the penalty parameter satisfies the condition $\rho \geq (\sqrt{3} + 1)L_f$, then

$$(*) \leq c(\|y^k - y^{k-1}\|_2^2 - \|y^{k+1} - y^k\|_2^2) + (B_f(y, y^{k+1}) - B_f(y, y^k)). \tag{6.10}$$

□

Proposition 6.5. *Suppose that we use a constant penalty parameter ρ that satisfies $\rho > (\sqrt{3} + 1)L_f$. Let $\{(x^k, y^k, w^k)\}_{k\geq 0}$ be the sequence generated by the LP-ADMM and (x^*, y^*, w^*) be a point satisfying the KKT conditions* (2.3) *with $x^* \in \mathcal{X}, y^* \in \mathcal{Y}$. Then, the following hold:*

(a) For all $k \geq 1$, $\|Ax^{k+1} - y^k\|_2^2 \geq \frac{1}{2}\|y^{k+1} - y^k\|_2^2 - \frac{{L_f}^2}{\rho^2}\|y^k - y^{k-1}\|_2^2$.

(b) For all $k \geq 0$ and (x, y) satisfying $Ax - y = 0$, we have $F(x^{k+1}, y^{k+1}) - F(x, y) \leq \frac{1}{2\rho}(\|w^k\|_2^2 - \|w^{k+1}\|_2^2) + \frac{\rho}{2}(\|y^k - y\|_2^2 - \|y^{k+1} - y\|_2^2) + c(\|y^k - y^{k-1}\|_2^2 - \|y^{k+1} - y^k\|_2^2) + (B_f(y, y^{k+1}) - B_f(y, y^k))$, where $c = \frac{\rho - 2L_f}{4}$.

(c) The sequence $\{\frac{1}{2\rho}\|w^k - w^*\|_2^2 + \frac{\rho}{2}\|y^k - y^*\|_2^2 - B_f(y^*, y^k)\}_{k\geq 0}$ is non-increasing and bounded below.

Proof. Parts (a) and (b) have been already proven in Lemmas 6.2 and 6.3. By the convexity of $F(\cdot, \cdot)$,

$$F(x^*, y^*) - F(x^{k+1}, y^{k+1}) \leq -\langle w^*, Ax^{k+1} - y^{k+1}\rangle. \tag{6.11}$$

Subsequently, combining (6) and (6.7),

$$\begin{aligned}F(x^{k+1}, y^{k+1}) - F(x^*, y^*) \leq{}& \langle w^{k+1}, Ax^{k+1} - y^{k+1}\rangle + \frac{\rho}{2}\|Ax^{k+1} - y^{k+1}\|_2^2\\ &+ \frac{\rho}{2}(\|y^k - y^*\|_2^2 - \|y^{k+1} - y^*\|_2^2) - (B_f(y^*, y^k) - B_f(y^*, y^{k+1}))\\ &+ \frac{L_f}{2}\|y^{k+1} - y^k\|_2^2 - \frac{\rho}{2}\|Ax^{k+1} - y^k\|_2^2.\end{aligned} \tag{6.12}$$

Summing up the terms (6.11) and (6.12), we have

$$\begin{aligned}0 \leq{}& \frac{1}{\rho}\langle w^{k+1} - w^*, w^k - w^{k+1}\rangle + \frac{\rho}{2}\|Ax^{k+1} - y^{k+1}\|_2^2 + \frac{\rho}{2}(\|y^k - y^*\|_2^2 - \|y^{k+1} - y^*\|_2^2)\\ &- (B_f(y^*, y^k) - B_f(y^*, y^{k+1})) + \frac{L_f}{2}\|y^{k+1} - y^k\|_2^2 - \frac{\rho}{2}\|Ax^{k+1} - y^k\|_2^2.\end{aligned} \tag{6.13}$$

Note that $m_{k+1} = \frac{1}{2\rho}\|w^{k+1} - w^*\|_2^2 + \frac{\rho}{2}\|y^{k+1} - y^*\|_2^2 - B_f(y^*, y^{k+1})$. Based on $(x_1 - x_2)^T(x_1 - x_3) = \frac{1}{2}(\|x_1 - x_2\|_2^2 + \|x_1 - x_3\|_2^2 - \|x_2 - x_3\|_2^2)$, we have,

$$0 \leq m_k - m_{k+1} + \frac{L_f}{2}\|y^{k+1} - y^k\|_2^2 - \frac{\rho}{2}\|Ax^{k+1} - y^k\|_2^2. \tag{6.14}$$

On top of Lemma 6.3,

$$(-\frac{L_f^2}{2\rho}+\frac{\rho}{4}-\frac{L_f}{2})\|y^{k+1}-y^k\|_2^2 \le m_k - m_{k+1} + \frac{L_f^2}{2\rho}(\|y^k-y^{k-1}\|_2^2 - \|y^{k+1}-y^k\|_2^2). \quad (6.15)$$

We construct a new sequence $m'_{k+1} = m_{k+1} + \frac{L_f^2}{2\rho}\|y^{k+1}-y^k\|_2^2$. Since the term $-\frac{L_f^2}{2\rho}+\frac{\rho}{4}-\frac{L_f}{2}$ is positive if $\rho > (\sqrt{3}+1)L_f$, it is easy to conclude that the sequence m'_{k+1} is non-increasing and bounded below. Moreover, we have

$$\sum_{k=0}^{\infty} \|y^{k+1}-y^k\|_2^2 < +\infty.$$

Thus, we conclude that the sequence $\{m_k\}$ is non-increasing, bounded below and $\|y^{k+1}-y^k\| \to 0$. Due to the dual update rule, we automatically have $\|x^{k+1}-x^k\| \to 0$ and $\|w^{k+1}-w^k\| \to 0$. □

Theorem 6.6. *Consider the setting of Proposition 6.5. Set* $\bar{x}^K = \frac{1}{K}\sum_{k=1}^K x^k$ *and* $\bar{y}^K = \frac{1}{K}\sum_{k=1}^K y^k$. *Then, the following hold:*

(a) The sequence $\{(x^k, y^k, w^k)\}_{k\ge 0}$ *converges to a KKT point of problem* (2.1).

(b) The sequence of function values converges at the rate $\mathcal{O}(\frac{1}{K})$*:*

$$F(\bar{x}^K, \bar{y}^K) - F(x^*, y^*) \le \frac{(1/2\rho)\|w^0\|_2^2 + (\rho/2)\|y^* - y^0\|_2^2 + c\|y^0-y^1\|_2^2}{K} = \mathcal{O}(\frac{1}{K}).$$

Proof. (a): By Proposition 6.5, we know that $\{m_k\}$ is a convergent sequence. Thus, $\{(x^k, y^k, w^k)\}$ is a bounded sequence. Every bounded sequence in $\mathbb{R}^n$ contains a convergent subsequence. Note that $\{(x^{k_j}, y^{k_j}, w^{k_j})\}$ is the corresponding subsequence. Hence, it has a limit point, which we denote by $(x^\infty, y^\infty, w^\infty)$.

- **Step 1:** Prove that the accumulation point $(x^\infty, y^\infty, w^\infty)$ is a KKT point.

As the dual update $\frac{1}{\rho}(w^k - w^{k+1}) = Ax^{k+1} - y^{k+1}$ and $\|w^k - w^{k+1}\| \to 0$, we have $\|Ax^{k+1} - y^{k+1}\| \to 0$ and $Ax^\infty - y^\infty = 0$. Thus, any accumulation point is dual feasible. Moreover, by the convexity of $F(\cdot,\cdot)$ and the optimality conditions (4.1) and (4.2),

$$\begin{aligned} F(x^*, y^*) \ge{} & F(x^{k_j}, y^{k_j}) - \langle \rho A^T(Ax^{k_j+1} - y^{k_j} - \frac{w^{k_j}}{\rho}), x^* - x^{k_j}\rangle \\ & + \langle \nabla f(y^{k_j+1}) - \nabla f(y^{k_j}) - w^{k_j+1}, y^* - y^{k_j}\rangle. \end{aligned} \quad (6.16)$$

As $j \to \infty$, we have

$$F(x^*, y^*) \ge F(x^\infty, y^\infty) - \langle w^\infty, A(x^* - x^\infty) + y^\infty - y^*\rangle. \quad (6.17)$$

Since the point (x^∞, y^∞) is dual feasible, $F(x^*, y^*) \ge F(x^\infty, y^\infty)$. As (x^∞, y^∞) and (x^*, y^*) are both feasible solutions, (x^∞, y^∞) is an optimal solution.

Moreover, based on the convexity of $g(\cdot)$ and the x-update optimality condition (4.1), we have

$$g(x) \ge g(x^{k_j}) - \langle \rho A^T(Ax^{k_j+1} - y^{k_j} - \frac{w^{k_j}}{\rho}), x - x^{k_j}\rangle. \quad (6.18)$$

As $j \to \infty$, we have

$$g(x) \ge g(x^\infty) + \langle A^T w^\infty, x - x^\infty\rangle. \quad (6.19)$$

Thus, $A^T w^\infty \in \partial g(x^\infty)$ and similarly $-w^\infty \in \nabla f(y^\infty) + \partial P(y^\infty)$. Therefore, $(x^\infty, y^\infty, w^\infty)$ is a KKT point.

- **Step 2:** Prove that the whole sequence $\{(x^k, y^k, w^k)\}_{k\ge 0}$ converges to $(x^\infty, y^\infty, w^\infty)$.

By choosing $(x^*, y^*, w^*) = (x^\infty, y^\infty, w^\infty)$ in Proposition 6.5, we have $\frac{1}{2\rho}\|w^{k_j+1} - w^\infty\|_2^2 + \frac{\rho}{2}\|y^{k_j+1} - y^\infty\|_2^2 - B_f(y^\infty, y^{k_j+1}) \to 0$. Subsequently, by Proposition 6.5(a), it is easy to conclude that $\frac{1}{2\rho}\|w^{k+1} - w^\infty\|_2^2 + \frac{\rho}{2}\|y^{k+1} - y^*\|_2^2 - B_f(y^\infty, y^{k+1}) \to 0$. Therefore, $(x^k, y^k, w^k) \to (x^\infty, y^\infty, w^\infty)$. As the point $(x^\infty, y^\infty, w^\infty)$ can be any limit point of the sequence $\{(x^k, y^k, w^k)\}_{k\geq 0}$, we conclude that $(x^\infty, y^\infty, w^\infty)$ is a KKT point.

(b): On top of Proposition 6.5(b), we have

$$\begin{aligned}F(x^{k+1}, y^{k+1}) - F(x^*, y^*) \leq \frac{1}{2\rho}(\|w^k\|_2^2 - \|w^{k+1}\|_2^2) + \frac{\rho}{2}(\|y^k - y^*\|_2^2 - \|y^{k+1} - y^*\|_2^2) \\ + c(\|y^k - y^{k-1}\|_2^2 - \|y^{k+1} - y^k\|_2^2) + (B_f(y^*, y^{k+1}) - B_f(y^*, y^k)).\end{aligned}$$

Since $F(\cdot, \cdot)$ is convex, by the Jensen inequality, we have

$$\begin{aligned}&F(\bar{x}^K, \bar{y}^K) - F(x^*, y^*) \leq \frac{1}{K}\sum_{k=1}^{K}(F(x^k, y^k) - F(x^*, y^*)) \\ &\leq \frac{1}{K}\left\{\frac{1}{2\rho}(\|w^0\|_2^2 - \|w^K\|_2^2) + \frac{\rho}{2}(\|y^0 - y^*\|_2^2 - \|y^K - y^*\|_2^2) + c(\|y^1 - y^0\|_2^2 - \|y^K - y^{K-1}\|_2^2)\right\} \\ &\leq \frac{1}{K}\left\{\frac{1}{2\rho}\|w^0\|_2^2 + \frac{\rho}{2}\|y^0 - y^*\|_2^2 + c\|y^1 - y^0\|_2^2\right\}.\end{aligned}$$

Noting that $B_f(y^*, y^K) - B_f(y^*, y^0) < 0$ when K is sufficiently large, we complete the proof. □

B: Implementation Details

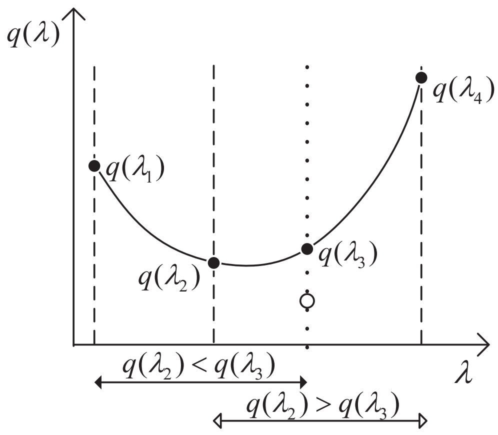


Figure 4: Schematics of the Golden Section Search Method

Algorithm 2: Golden Section Search Algorithm

```
Input: $\lambda_1 = 0$, $\lambda_4 = \lambda^U$, $r = 0.618$;
Output: Optimal solution $(\beta^*, \lambda^*)$;
while not converge do
    $\lambda_2 = r\lambda_1 + (1-r)\lambda_4$;
    $\lambda_3 = (1-r)\lambda_1 + r\lambda_4$;
    $q(\lambda_1) =$ LP-ADMM$(\lambda_1)$;
    $q(\lambda_2) =$ LP-ADMM$(\lambda_2)$;
    $q(\lambda_3) =$ LP-ADMM$(\lambda_3)$;
    $q(\lambda_4) =$ LP-ADMM$(\lambda_4)$;
    if $q(\lambda_2) < q(\lambda_3)$ then
        set $\lambda_4 = \lambda_3$
    else
        set $\lambda_1 = \lambda_2$
    end
end
```

Primal Dual Hybrid Gradient Method

$$\min_{\|x\|_\infty \le \lambda} \max_{\|y\|_\infty \le 1} \frac{1}{N} \sum_{i=1}^{N} \left\{ \log(1 + \exp(-a_i^T x)) + \frac{1}{2}(a_i^T x - b_i) \right\} + \frac{1}{2N} y^T (Ax - b) \qquad (6.20)$$

Standard ADMM

$$\begin{aligned} \min_{x,\,y} \quad & f(y) + g(z) + \mathbb{I}_{\{\|x\|_\infty \le \lambda\}} \\ \text{s.t.} \quad & Ax = y, \\ & z = y - b. \end{aligned} \qquad (6.21)$$

The corresponding augmented Lagrangian function is

$$\mathcal{L}_\rho(x, y; u, v) = f(y) + g(z) + u^T(Ax - y) + v^T(z - y + b) + \frac{\rho}{2}\|Ax - y\|_2^2 + \frac{\rho}{2}\|z - y + b\|_2^2.$$

Remark 6.7. *We can regard (y, z) as a block in problem* (6.21) *and apply the standard ADMM to solve it. Moreover, though $g(\cdot)$ is a non-smooth function, it statisfies the semi-smooth property and its generalized Hessian matrix can be easily computed due to the log-loss function. Namely,*

$$(y^{k+1}, z^{k+1}) = \arg\min_{y} \left\{ f(y) + \frac{1}{2}\|y - d_1\|^2 + \min_{z}\{g(z) + \frac{1}{2}\|z - (y - d_2)\|^2\} \right\}, \tag{6.22}$$

where $d_1 = Ax^{k+1} + \frac{\mu^k}{\rho}$ and $d_2 = b + \frac{v^k}{\rho}$.

```
Algorithm 3: Standard ADMM
Input: $\rho = 10$, $\sigma = 1$, $k = 1$, randomly generate $x^1$, $y^1$, $z^1$, $u^1$, $v^1$;
Output: Optimal solution $(x^*, y^*, z^*)$;
while not converge do
    $x^{k+1} = \underset{x}{\text{argmin}} \left\{ \frac{\rho}{2}\|Ax - y^k + \frac{u^k}{\rho}\|_2^2 + g(x) \right\}$;      ▷ Conjugate Gradient Method
    $y^{k+1} = \underset{y \in \mathbb{R}^N}{\text{argmin}} \left\{ f(y) + \frac{\rho}{2}\|y - Ax^{k+1} - \frac{u^k}{\rho}\|_2^2 \right\}$;      ▷ Semi-Smooth Newton
    $z^{k+1} = \underset{z \in \mathbb{R}^N}{\text{argmin}} \left\{ g(z) + \frac{\rho}{2}\|z - y^{k+1} + b + \frac{v^k}{\rho}\|_2^2 \right\}$;
    $u^{k+1} = u^k + \sigma\rho(Ax^{k+1} - y^{k+1})$;
    $v^{k+1} = v^k + \sigma\rho(z^{k+1} - y^{k+1} + b)$;
    $k = k + 1$;
end
```

```
Algorithm 4: Semi-smooth Newton Method
Input: $k = 1$, random generate $y^1$;
Output: Optimal solution $y^*$;
while not converge do
    $\nabla f(y^k) = -\frac{e^{-y^k}}{(1+e^{-y^k})} + \rho(y - Ax - \frac{u}{\rho}) + \rho(y - z - b - \frac{v}{\rho})$;
    if $y_i^k > (b + \frac{v}{\rho})_i$ or $y_i^k < (b + \frac{v}{\rho})_i + \frac{1}{\rho}$ then
        $h_i = 1$;
    else
        $h_i = 0$;
    end
    $\nabla^2 f(y^k) = \frac{e^{-y^k}}{(1+e^{-y^k})^2} + \rho + \rho h$;      ▷ computing the generalized Hessian
    $g^k = (\nabla^2 f(y^k))^{-1} \cdot \nabla f(y^k)$;
    $y^{k+1} = y^k + ss \cdot g^k$;      ▷ ss is returned by line search
    $k = k + 1$;
end
```

Note that the subproblem solvers are the accelerated projected gradient method and semi-smooth Newton method, respectively.

Solver for Quadratic Minimization with Box constraints

$$\begin{aligned} \min_{x} \quad & \|Ax - b\|_2^2 \\ \text{s.t.} \quad & \|x\|_\infty \leq \lambda. \end{aligned} \tag{6.23}$$

Algorithm 5: Accelerated Projected Gradient

```
Input: $ss = 1/\lambda_{max}(A)$, $k = 1$, random generate $x^1 = x_{old}$;
Output: Optimal solution $x^*$;
while not converge do
    $\beta_k = \frac{k}{k+3}$;
    $y^k = x^k + \beta(x^k - x_{old})$;
    $g^k = A^T(Ax^k - b)$;
    $w^k = y^k - ss \cdot g^k$;
    $w^k = \text{proj}_{\{\|w\|_\infty \leq \lambda\}}(w^k)$;
    $x_o = x^k$;
    $x^k = w^k$;
    $k = k + 1$;
end
```

Algorithm 6: Conjugate Gradient with Active Set

```
Input: Randomly initialize $x^1$, $g^1 = A^T(Ax^1 - b)$, $k = 1$;
Bound set $\bar{B} = \{i : |x_i| = \lambda \text{ and } -g_i \cdot x_i \geq 0\}$, Free set $\bar{F} = \{i : i \notin \bar{B}\}$;
$r_i^k = \begin{cases} -g_i^k, & i \in \bar{F}, \\ 0, & \text{otherwise} \end{cases}$
Output: Optimal solution $x^*$;
while not converge do
    Update $\bar{F}$;
    $r_i^k = \begin{cases} -g_i^k, & i \in \bar{F}, \\ 0, & \text{otherwise} \end{cases}$;
    if $k = 1$ or $\bar{F}$ changed then
        $p^k = r^k$
    else
        $\beta_k = \frac{\|r^k\|^2}{\|r^{k-1}\|^2}$;
        $p^k = r^k + \beta_k p^{k-1}$
    end
    $\alpha_k = \frac{\|r^k\|^2}{\|Ap^k\|^2}$;
    $\tilde{x}^k = x^{k-1} + \alpha_k p^k$;
    $x^k = \text{proj}_{\{\|x\|_\infty \leq \lambda\}}(\tilde{x}^k)$;
    if $\tilde{x}^k = x^k$ then
        $g^k = g^k + \alpha_k A p^k$
    else
        $g^k = A^T(Ax^k - b)$
    end
    $k = k + 1$;
end
```

Remark 6.8. *As problem* (6.23) *takes a similar form as the dual formulation of SVM, (i.e., see the form (4) in [10]), we use the same coordinate minimization algorithm as our alternative inner solver.*

Table 4: The CPU time (s) of our proposed first-order algorithmic framework on real datasets

Dataset	Samples	Features	CPU time (s)	Sparsity Level
mushrooms	8123	112	8.47	0.8125
phishing	11055	68	4.01	0.7543
w1a	2477	300	13.15	0.9618
w2a	3470	300	15.13	0.9612
w3a	4912	300	15.96	0.9612
w4a	7366	300	19.35	0.9611
w5a	9888	300	19.50	0.9612
w6a	17188	300	11.42	0.9611
w7a	24692	300	12.07	0.9611
w8a	49749	300	14.86	0.9612
MNIST(0 vs 3)	12054	752	118.68	0.7641
MNIST(0 vs 4)	11765	754	117.81	0.7784
MNIST(0 vs 6)	11841	720	118.79	0.7577
MNIST(2 vs 3)	12089	752	127.43	0.7792
MNIST(2 vs 5)	11379	771	134.80	0.7912
MNIST(5 vs 8)	11272	771	111.54	0.7883
MNIST(5 vs 9)	11370	780	123.89	0.8109
MNIST(6 vs 9)	11867	780	139.14	0.8077

C: Additional Experiment Results

Here, the sparsity level is the ratio of the zero elements to the total elements. The stopping criterion used in our proposed LP-ADMM is the dual infeasibility $\|Ax^{k+1} - y^{k+1}\|_2 \leq 10^{-6}$ in all experiments.