Journal of Machine Learning Research 25 (2024) 1-40 Submitted 3/23; Revised 4/24; Published 5/24

Adjusted Wasserstein Distributionally Robust Estimator in Statistical Learning

Yiling Xie YXIE350@GATECH.EDU
Xiaoming Huo HUO@GATECH.EDU
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA

Editor: Po-Ling Loh

Abstract

We propose an *adjusted* Wasserstein distributionally robust estimator—based on a nonlinear transformation of the Wasserstein distributionally robust (WDRO) estimator in statistical learning. The classic WDRO estimator is asymptotically biased, while our adjusted WDRO estimator is asymptotically *unbiased*, resulting in a smaller asymptotic mean squared error. Further, under certain conditions, our proposed adjustment technique provides a general principle to de-bias asymptotically biased estimators. Specifically, we will investigate how the adjusted WDRO estimator is developed in the generalized linear model, including logistic regression, linear regression, and Poisson regression. Numerical experiments demonstrate the favorable practical performance of the adjusted estimator over the classic one.

Keywords: distributionally robust optimization; asymptotic normality; Wasserstein distance; unbiased estimator; generalized linear model

1. Introduction

Wasserstein distributionally robust optimization (WDRO) has appeared as a promising tool to achieve "robust" decision-making (Mohajerin Esfahani and Kuhn, 2018; Blanchet and Murthy, 2019; Gao and Kleywegt, 2022). WDRO has attracted intense research interest in the past few years. It is well-known that WDRO admits tractable reformulations (Mohajerin Esfahani and Kuhn, 2018) and has a powerful out-of-sample performance guarantee (Gao, 2022). People also have been actively exploring its applications in financial portfolio selection (Blanchet et al., 2022a), statistical learning (Chen and Paschalidis, 2018; Shafieezadeh-Abadeh et al., 2019), neural networks (Sinha et al., 2018), automatic control (Yang, 2020), transportation (Carlsson et al., 2018), and energy systems (Wang et al., 2018), among many others.

WDRO can be applied in statistical learning (Chen and Paschalidis, 2018; Kuhn et al., 2019; Nguyen et al., 2022). In general, the statistical learning model can be written as the following optimization problem:

$$\min_{\beta \in B} \mathbb{E}_{P_*} \left[L(f(\mathbf{X}, \beta), Y) \right],$$

where $\mathbf{X} \in \Omega \subset \mathbb{R}^d$ denotes the feature variable, Ω is a convex set, Y denotes the response variable, P_* is the true data-generating distribution of $(\mathbf{X}, Y)$, $f(\cdot, \beta)$ is the hypothesis function parameterized by $\beta \in B \subset \mathbb{R}^d$, B is a compact convex set, and L is the loss function. Considering the true data-generating distribution P_* is usually unknown, the empirical risk minimization can be applied to estimate the ground-truth hypothesis function $f(\cdot, \beta_*)$ parameterized by $\beta_* \neq \mathbf{0}$. However, the empirical risk minimization estimators are sensitive to perturbations and suffer from overfitting (Smith and Winkler, 2006; Shalev-Shwartz and Ben-David, 2014). To obtain robust estimators with desirable generalization abilities, distributionally robust optimization is proposed, which minimizes the worst-case expected loss among an ambiguity set $\mathcal{U}$ of distributions. In this paper, we are interested in the Wasserstein ambiguity set, and then the resulting problem is the so-called Wasserstein distributionally robust optimization. The Wasserstein ambiguity is defined as the ball centered at the empirical distribution $\mathbb{P}_n$ and contains all distributions close to $\mathbb{P}_n$ in the sense of the Wasserstein distance. We denote the WDRO estimators—the solutions to the WDRO problem—by β_n^{DRO}. More details will be stated in Section 4.

The asymptotic distribution of the WDRO estimator β_n^{DRO} can be obtained under certain regularity conditions. However, the associated convergence results imply that the WDRO estimator β_n^{DRO} has an asymptotic bias. From the perspective of parameter estimation, the asymptotic bias indicates an inaccurate estimation of the ground-truth parameter β_*. Inspired by this phenomenon, we provide a general adjustment technique to de-bias the asymptotically biased estimators. The asymptotic behavior of the asymptotically biased estimator under different transformations is also discussed.

We obtain the adjusted WDRO estimator, denoted by β_n^{ADRO}, by applying the proposed adjustment technique to the WDRO problem. It will be shown that the adjusted WDRO estimator β_n^{ADRO} could be computed exactly simply using the given samples and the value of the classic WDRO estimator β_n^{DRO}, making it convenient to apply the proposed technique. Also, the existence and the asymptotic unbiasedness of the adjusted WDRO estimator β_n^{ADRO} could be promised under mild conditions, enabling broad applications of the proposed technique. In addition, since the proposed adjusted WDRO estimator β_n^{ADRO} is transformed from the classic WDRO estimator β_n^{DRO}, the out-of-sample guarantee of the WDRO estimator β_n^{DRO} could promise the generalization capacity of the proposed adjusted WDRO estimator β_n^{ADRO}.

Since the generalized linear model includes multiple widely-used regression models and is easy to interpret and implement, we will articulate how to apply the adjustment strategy in the setting of the generalized linear model, including linear regression, logistic regression, and Poisson regression. Then, we carry out the numerical experiments in the generalized linear model. Our numerical experiments illustrate that the proposed estimator β_n^{ADRO} has a superior performance even if the sample size is not very large.

1.1 Related Work

We review the existing work related to the proposed adjusted WDRO estimator. WDRO is broadly applied to solve parameter-estimation problems (Kuhn et al., 2019; Shafieezadeh-Abadeh et al., 2019; Aolaritei et al., 2022; Nguyen et al., 2022). Multiple algorithms have been developed (Li et al., 2019; Luo and Mehrotra, 2019; Blanchet et al., 2022c) and can be

applied to compute the estimators in the WDRO framework. While intense work focuses on adapting WDRO to different machine learning problems, deriving the tractable reformulations, and solving the WDRO problems efficiently, people have begun to investigate the statistical properties of WDRO estimators in recent few years, e.g., Blanchet et al. (2021, 2022b); Xie and Huo (2024), evaluating the behavior of WDRO through the lens of statistics. Notably, the asymptotic distribution of the WDRO estimator has been proven to be normal and has an asymptotic bias (Blanchet et al., 2022b). In this paper, we propose a nonlinear transformation to overcome this shortcoming. It will be shown that the estimator obtained from the nonlinear transformation has an asymptotically smaller mean squared error, indicating the proposed estimator is more accurate in the asymptotic sense. In the literature of WDRO, the generalization bounds, i.e., the upper confidence bounds on the out-of-sample loss, have been established to guarantee the out-of-sample performance of the WDRO estimator (Mohajerin Esfahani and Kuhn, 2018; Shafieezadeh-Abadeh et al., 2019; Gao, 2022). Since the proposed adjusted WDRO estimator is transformed from the classic WDRO estimator, we can also develop the generalization bounds for the associated adjusted WDRO estimator.

1.2 Organization of this Paper

The remainder of this paper is organized as follows. In Section 2, we introduce the adjustment technique that could de-bias the general asymptotically biased estimators under certain conditions. In Section 3, we discuss the asymptotic behavior of the WDRO problem. In Section 4, we give the formulation of the adjusted WDRO estimator in statistical learning. In Section 5, we show how to develop the adjusted WDRO estimators in the generalized linear model. Numerical experiments are conducted and analyzed in Section 6. The proofs are relegated to the appendix whenever possible.

2. Adjustment Technique

In this section, we first discuss the properties of transformations on the asymptotically biased estimators, based on which we provide a general strategy to de-bias the asymptotically biased estimators under certain conditions. The proposed adjustment technique will be further illustrated in detail in the WDRO setting in Section 4.

Suppose the estimator $\beta_n \in \mathbb{R}^d$ is obtained by the following parameter-estimation procedure:

$$\beta_n \in \arg\min_{\beta} l(\mathbb{P}_n, \beta),$$

where l is the loss and depends on the empirical distribution $\mathbb{P}_n$ and parameter β. Also, suppose that the estimator β_n has the following convergence in distribution:

$$\sqrt{n}(\beta_n - \beta_*) \Rightarrow \mathcal{N}(f(\beta_*), D), \tag{1}$$

where $\Rightarrow$ means "converge in distribution", $D \in \mathbb{R}^{d\times d}$ is the asymptotic covariance matrix, $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\beta_* \in \mathbb{R}^d$ is the ground-truth parameter. We focus on the scenario when $f \neq 0$.

For the estimator β_n with the limiting distribution in (1), our goal is to look for some (deterministic) transformation ϕ_n to obtain a more accurate estimation of β_* in the asymp-

totic sense. The following proposition states that the "best" transformations have a unique formulation.

Proposition 1 *Suppose β_n is an estimator of ground-truth parameter β_* and has the following convergence in distribution:*

$$\sqrt{n}(\beta_n - \beta_*) \Rightarrow \mathcal{N}(f(\beta_*), D),$$

where f is differentiable at some neighborhood $\mathcal{B}(\beta_*)$ of β_*. Assume the transformation ϕ_n is differentiable at $\mathcal{B}(\beta_*)$ and satisfies $\phi_n(\beta) \to \phi(\beta)$ and $\phi_n'(\beta) \to \phi'(\beta)$ for every β in $\mathcal{B}(\beta_*)$, where ϕ is differentiable, and ϕ_n' and ϕ' are the gradients of ϕ_n and ϕ. Under this assumption, the least asymptotic mean squared error of $\phi_n(\beta_n)$ is $\mathrm{tr}(D)$, which can be achieved if and only if the transformation ϕ_n has the following formulation

$$\phi_n(\beta) = \beta - \frac{1}{\sqrt{n}} g(\beta) + o\left(\frac{1}{\sqrt{n}}\right), \tag{2}$$

where g is some differentiable function at $\mathcal{B}(\beta_*)$ satisfying $g(\beta_*) = f(\beta_*)$, resulting in the following convergence in distribution:

$$\sqrt{n}(\phi_n(\beta_n) - \beta_*) \Rightarrow \mathcal{N}(0, D).$$

Proposition 1 demonstrates that for the asymptotically biased estimator β_n, the transformation ϕ_n should take the formulation (2) to achieve the least asymptotic mean squared error $\mathrm{tr}(D)$. Meanwhile, the resulting estimator $\phi_n(\beta_n)$ is asymptotically unbiased.

The transformation ϕ_n in the formulation (2) is desirable, and one can simply let $g = f$ to define the transformation ϕ_n in (2). However, the function f is usually unknown. For example, in the limiting distribution of the WDRO estimator, f depends on the unknown ground-truth data-generating distribution. In this regard, the function f should be approximated accordingly.

Suppose we have a sequence of (stochastic) functions f_n to approximate the function f. Our adjustment transformation is defined in terms of f_n and based on the formulation of ϕ_n shown in (2). Certain conditions should be imposed to f_n to promise that the estimator obtained by our adjustment transformation is asymptotically unbiased and could have the asymptotic mean squared error $\mathrm{tr}(D)$. More details are described in Assumption 2 and Theorem 3.

Before introducing Theorem 3, we state our assumptions of functions f_n.

Assumption 2 *Given function f, f_n and β_*, we assume that*

- *The function f_n is differentiable at some neighborhood $\mathcal{B}(\beta_*)$ of β_*.*
- *The sequence $\sup_{\beta \in \mathcal{B}(\beta_*)} \|f_n'(\beta)\|$ is bounded in probability.*
- *$f_n(\beta_*) \to_p f(\beta_*)$, where $\to_p$ means "converge in probability".*

Equipped with Assumption 2, we give our main result in the following theorem.

Theorem 3 (Adjustement Technique) *Suppose β_n is an estimator of ground-truth parameter β_* and has the following convergence in distribution:*

$$\sqrt{n}(\beta_n - \beta_*) \Rightarrow \mathcal{N}(f(\beta_*), D),$$

where f is differentiable at some neighborhood $\mathcal{B}(\beta_*)$ of β_*. If we have the function f_n satisfying Assumption 2 and the transformation $\mathcal{A}_n$ defined by

$$\mathcal{A}_n(\beta_n) = \beta_n - \frac{1}{\sqrt{n}} f_n(\beta_n),$$

then we have that

$$\sqrt{n}\left(\mathcal{A}_n(\beta_n) - \beta_*\right) \Rightarrow \mathcal{N}(0, D). \tag{3}$$

The convergence (3) in Theorem 3 demonstrates that the proposed adjusted estimator $\mathcal{A}_n(\beta_n)$ is asymptotically unbiased and the asymptotic covariance matrix remains unchanged, resulting in a smaller asymptotic means square error $\mathrm{tr}(D)$, which is the least asymptotic mean squared error stated in Proposition 1. In this regard, to de-bias the asymptotically biased estimators, one only needs to have a sequence of functions f_n satisfying Assumption 2.

2.1 Sequential Delta Method

Notice that the transformations ϕ_n discussed in Proposition 1 depend on n. In this way, when we discuss the asymptotic distribution of $\phi_n(\beta_n)$, the classic delta method is not applicable. To resolve this issue, we have developed a sequential delta method based on the extended continuous mapping theorem, seeing Theorem 1.11.1 in Van der Vaart and Wellner (1996). The sequential delta method may have an independent research interest, so we state it in the following theorem.

Theorem 4 (Sequential Delta Method) *Let ϕ_n and $\phi : \mathbb{D} \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ be functions defined on a subset of $\mathbb{R}^d$. Suppose ϕ_n and ϕ are differentiable at the neighborhood $\mathcal{B}(\vartheta) \subset \mathbb{D}$ of $\vartheta \in \mathbb{D}$, and $\phi_n(\theta) \to \phi(\theta)$ and $\phi_n'(\theta) \to \phi'(\theta)$ hold for every $\theta \in \mathcal{B}(\vartheta)$, where ϕ' and ϕ_n' are gradients of the functions ϕ and ϕ_n. Let T_n be random vectors taking their values in $\mathbb{D}$. If $r_n(T_n - \vartheta) \Rightarrow \mathcal{N}(\mu, \Sigma)$ for numbers $r_n \to \infty$, then we have that*

$$r_n(\phi_n(T_n) - \phi_n(\vartheta)) \Rightarrow \mathcal{N}(\phi'(\vartheta)\mu, \phi'(\vartheta)\Sigma\phi'(\vartheta)^\top).$$

3. WDRO Problem

This section discusses the problem formulation of WDRO and gives the asymptotic distribution of the WDRO estimator.

3.1 Problem Formulation

The WDRO problem can be written as

$$\beta_n^{DRO} \in \arg\min_{\beta \in B} \sup_{P \in \mathcal{U}_{\rho_n}(\mathbb{P}_n)} \mathbb{E}_P\left[L(f(\mathbf{X}, \beta), Y)\right], \tag{4}$$

where the feature variable $\mathbf{X}$ belongs to the convex set $\Omega \subset \mathbb{R}^d$, the response variable Y can be continuous or discrete, f is the hypothesis function parametrized by $\beta \in B \subset \mathbb{R}^d$, B is a compact convex set, $\mathcal{U}_{\rho_n}(\mathbb{P}_n)$ is the Wasserstein uncertainty set, and L is the loss function. The Wasserstein uncertainty set is defined by

$$\mathcal{U}_{\rho_n}(\mathbb{P}_n) = \{P : W_p(P, \mathbb{P}_n) \leq \rho_n\}, \tag{5}$$

where $\mathbb{P}_n$ is the empirical distribution of the samples $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), ..., (\mathbf{X}_n, Y_n)\}$ generated by true data-generating distribution P_*,

$$W_p(P, \mathbb{P}_n) = \left(\inf_{\gamma \in \Gamma(P, \mathbb{P}_n)} \left\{ \int_{Z^2} d^p(z, z') d\gamma(z, z') \right\} \right)^{1/p},$$

$\Gamma(P, \mathbb{P}_n)$ is the set of distributions with marginals P and $\mathbb{P}_n$, d is some metric in space $Z = \mathbf{X} \times Y$, and $W_p(P_1, P_2)$ is the so-called p-Wasserstein distance.

3.2 Asymptotic Distribution of the WDRO Estimator

In this subsection, we study the asymptotic distribution of the WDRO estimator in the supervised statistical learning.

Blanchet et al. (2022b) have derived the asymptotic distribution of the WDRO estimator in the unsupervised learning setting. In our study, we first let the cost function be infinite if the response variables are different and then adapt the asymptotic distribution of the WDRO estimator to the supervised statistical learning setting.

To adapt the results, we should specify the hyperparameters of the Wasserstein uncertainty set and clarify some regularity conditions, which should be satisfied for the loss function L and the underlying data-generating distribution P_* of $(\mathbf{X}, Y)$.

Assumption 5 *The hyperparameters of the Wasserstein uncertainty set $\mathcal{U}_{\rho_n}(\mathbb{P}_n)$ in (5) are prescribed as follows,*

- $\rho_n = \tau/\sqrt{n}$, $\tau > 0$,
- $p = 2$,
- $d\left((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\right) = \begin{cases} \|\mathbf{x}_1 - \mathbf{x}_2\|_2 & y_1 = y_2 \\ \infty & y_1 \neq y_2 \end{cases}$.

Remark 6 *We justify the choices of hyperparameter in Assumption 5 as follows,*

- *We choose the radius to be of the square-root order $\mathcal{O}(1/\sqrt{n})$ because the powerful out-of-sample performance guarantee can be proved (Gao and Kleywegt, 2022), and the confidence region can be constructed (Blanchet et al., 2022b) with the square-root order.*
- *We choose the 2-Wasserstein distance since the 2-Wasserstein distance applies to the quadratic loss, and the associated WDRO problem could be solved by iterative algorithms (Blanchet et al., 2022c).*

- *The distance function d is infinite when $y_1 \neq y_2$, admitting distributional ambiguities only with respect to the feature variable $\mathbf{X}$. In the classification problem, the distance function d can be applied to tasks where the samples are correctly labeled (Gao et al., 2017). In the regression problem, the distance function d can help recover several popular regularized estimators, including square-root LASSO estimator (Blanchet et al., 2019; Shafieezadeh-Abadeh et al., 2019).*

Assumption 7 *The loss function $L(f(\mathbf{x}, \beta), y)$ satisfies:*

a. *The loss function $L(f(\mathbf{x}, \beta), y)$ is twice continuously differentiable w.r.t. $\mathbf{x}$ and β.*

b. *For each variable $\mathbf{x} \in \Omega$ and y, the loss function $L(f(\mathbf{x}, \beta), y)$ is convex w.r.t. β.*

c. *For each parameter $\beta \in B$ and variable y, the function $\left\| \frac{\partial^2 L(f(\mathbf{x},\beta),y)}{\partial \mathbf{x}^2} \right\|_2$ is uniformly continuous w.r.t. $\mathbf{x}$ and uniformly bounded by a continuous function $M(\beta)$.*

Assumption 8 *The underlying data-generating distribution P_* of $(\mathbf{X}, Y)$ satisfies:*

a. *There exists $\beta_* \in B^\circ$, where B° means the interior of B, satisfying*

$$\mathbb{E}_{P_*}\left[\frac{\partial L(f(\mathbf{X}, \beta), Y)}{\partial \beta}\right]\Bigg|_{\beta=\beta_*} = \mathbf{0},$$

and the inequalities

$$C(\beta_*) := \mathbb{E}_{P_*}\left[\frac{\partial^2 L(f(\mathbf{X}, \beta), Y)}{\partial \beta^2}\right]\Bigg|_{\beta=\beta_*} \succ 0, \tag{6}$$

$$\mathbb{E}_{P_*}\left[\left\|\frac{\partial L(f(\mathbf{X}, \beta), Y)}{\partial \beta}\right\|_2^2\right]\Bigg|_{\beta=\beta_*} < \infty$$

hold, where $C(\beta_*) \succ 0$ means the matrix $C(\beta_*)$ is a positive definite matrix.

b. *P_* is non-degenerate in the sense that*

$$P_*\left(\frac{\partial L(f(\mathbf{X}, \beta), Y)}{\partial \mathbf{X}} \neq \mathbf{0}\right)\Bigg|_{\beta=\beta_*} > 0,$$

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L(f(\mathbf{X}, \beta), Y)}{\partial \mathbf{X} \partial \beta}\left(\frac{\partial^2 L(f(\mathbf{X}, \beta), Y)}{\partial \mathbf{X} \partial \beta}\right)^\top\right]\Bigg|_{\beta=\beta_*} \succ 0,$$

where $\frac{\partial^2 L}{\partial \mathbf{x} \partial \beta}$ means taking the gradient first w.r.t. β and then w.r.t. $\mathbf{x}$.

Next, we obtain the associated convergence of the WDRO estimator β_n^{DRO} in problem (4) under Assumption 5, 7, and 8, which is shown in the following theorem.

Theorem 9 (Extension of Theorem 1 in Blanchet et al. (2022b)) *Suppose that Assumption 5, 7 and 8 are satisfied, $\Omega = \mathbb{R}^d$ and $\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\right] < \infty$, the WDRO estimator β_n^{DRO} in problem* (4) *has the following convergence in distribution:*

$$\sqrt{n}(\beta_n^{DRO} - \beta_*) \Rightarrow \mathcal{N}\left(-C(\beta_*)^{-1}H\left(\beta_*\right), D(\beta_*)\right), \tag{7}$$

where

$$H(\beta_*) = \tau \frac{\partial \sqrt{\mathbb{E}_{P_*}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial \mathbf{X}}\right\|_2^2\right]}}{\partial \beta}\Bigg|_{\beta=\beta_*}, \tag{8}$$

τ is the coefficient in the Wasserstein radius $\rho_n = \tau/\sqrt{n}$,

$$D(\beta_*) = C(\beta_*)^{-1}\mathrm{Cov}\left(\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial \beta}\right)\Bigg|_{\beta=\beta_*} C(\beta_*)^{-1}, \tag{9}$$

and $C(\beta_*)$ is defined in (6).

Remark 10 *The assumption $\Omega = \mathbb{R}^d$ could be relaxed. If Ω is compact and could be expressed as $\Omega = \{\mathbf{x} \in \mathbb{R}^d : A\mathbf{x} \leq b\}$, where A is an $l \times d$ matrix with linearly independent rows and $b \in \mathbb{R}^l$, and $\mathbf{X}$ has a probability density which is absolutely continuous w.r.t. Lebesgue measure, then the convergence* (7) *still holds. This claim can be seen in Section 6 in Blanchet et al. (2022b).*

Remark 11 (Finite Sample Size) *We investigate the empirical distribution of β_n^{DRO} when n is not very large. The WDRO esitmator β_n^{DRO} is computed in the logistic regression model when $n = 200$, and we plot the histograms of $\sqrt{n}(\beta_n^{DRO} - \beta_*)$ in Figure 1. Two dimensions of β_n^{DRO} are plotted separately. We conclude from Figure 1 that β_n^{DRO} is approximately normally distributed with a nonzero mean, as asymptotic convergence* (7) *suggested. We further apply the Shapiro–Wilk test and the test result supports our claim that β_n^{DRO} is approximately normally distributed even though the sample size is not very large, indicating that the asymptotic behavior of β_n^{DRO} "comes early". Therefore, making the bias in asymptotic convergence* (7) *disappear is meaningful in the sense of both asymptotic and finite sample size.*

Theorem 9 indicates that the term $\sqrt{n}(\beta_n^{DRO} - \beta_*)$ converges in distribution to a normal distribution with nonzero mean $-C(\beta_*)^{-1}H(\beta_*)$. Recall that we perturb the samples to achieve robustification. As explained in Blanchet et al. (2021), the bias term $-C(\beta_*)^{-1}H(\beta_*)$ could be understood as pushing towards solutions with less variation resulting from data perturbation. However, this nonzero bias term may imply that the WDRO estimator is not an accurate estimator for the ground-truth parameter β_*. We may consider transforming the WDRO estimator β_n^{DRO} to remove the bias term using the adjustment technique mentioned in Section 2.

4. Proposed Adjusted WDRO Estimator

This section introduces the formal formulation of our adjusted WDRO estimator and investigates the relevant properties, including unbiasedness, possible simplification, and the out-of-sample guarantee.

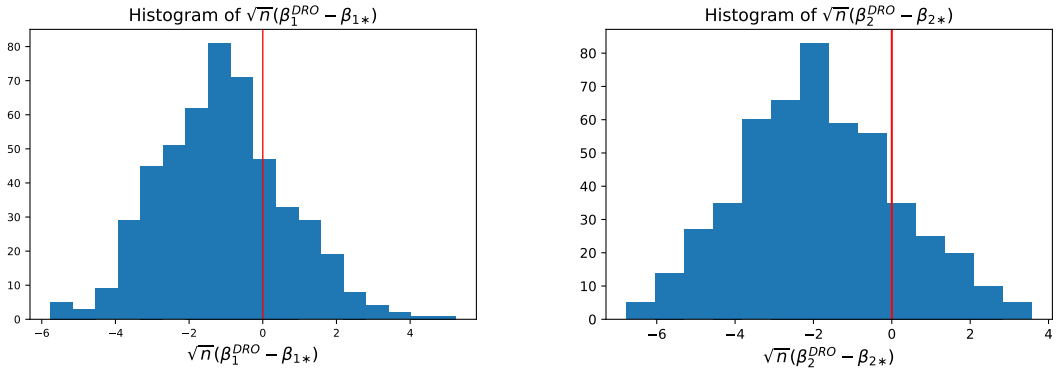

Figure 1: Histogram of β_n^{DRO}

4.1 Definition and Existence

The adjusted WDRO estimator is based on the asymptotic distribution obtained in Section 3.2 and the adjustment technique introduced in Section 2. Recall the WDRO estimator has the following convergence:

$$\sqrt{n}(\beta_n^{\mathrm{DRO}} - \beta_*) \Rightarrow \mathcal{N}\left(-C(\beta_*)^{-1}H\left(\beta_*\right), D(\beta_*)\right),$$

where

$$C(\beta_*) = \mathbb{E}_{P_*}\left[\frac{\partial^2 L(f(\mathbf{X},\beta),Y)}{\partial\beta^2}\right]\Bigg|_{\beta=\beta_*}, \quad H(\beta_*) = \tau\frac{\partial\sqrt{\mathbb{E}_{P_*}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}{\partial\beta}\Bigg|_{\beta=\beta_*}.$$

Notice that the asymptotic bias $f(\beta_*) = -C(\beta_*)^{-1}H\left(\beta_*\right)$ depends on the unknown underlying data-generating distribution P_*, but we can use the associated empirical distribution to approximate f. Applying the adjusted technique proposed in Theorem 3, we define the adjusted WDRO estimator in the following.

Definition 12 (Adjusted WDRO Estimator) *In the WDRO problem* (4)*, under Assumption 5, 7, and 8, the adjusted WDRO estimator is defined by*

$$\beta_n^{ADRO} = \mathcal{A}_n(\beta_n^{DRO}), \tag{10}$$

where

$$\mathcal{A}_n(\mathbf{z}) = \mathbf{z} + \frac{C_n(\mathbf{z})^{-1}H_n(\mathbf{z})}{\sqrt{n}},$$

$$H_n(\mathbf{z}) = \tau\frac{\partial\sqrt{\mathbb{E}_{\mathbb{P}_n}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}{\partial\beta}\Bigg|_{\beta=\mathbf{z}}, \tag{11}$$

$$C_n(\mathbf{z}) = \mathbb{E}_{\mathbb{P}_n}\left[\frac{\partial^2 L(f(\mathbf{X},\beta),Y)}{\partial\beta^2}\right]\Bigg|_{\beta=\mathbf{z}}. \tag{12}$$

To promise the existence of the adjusted WDRO estimator, we need additional conditions to let the matrix $C_n(\beta_n^{DRO})$ be invertible and the vector $H_n(\beta_n^{DRO})$ be well-defined. The conditions are shown in the following proposition.

Proposition 13 (Existence of Adjusted WDRO Estimator I) *In the WDRO problem* (4)*, under Assumption 5, 7, and 8, for the empirical distribution $\mathbb{P}_n$, the loss function $L(f(\mathbf{x},\beta),y)$ and the WDRO estimator β_n^{DRO}, if*

$$\mathbb{P}_n\left(\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial \mathbf{X}}\right\|_2^2 \neq 0\right)\Bigg|_{\beta=\beta_n^{DRO}} > 0, \quad \mathbb{E}_{\mathbb{P}_n}\left[\frac{\partial^2 L(f(\mathbf{X},\beta),Y)}{\partial \beta^2}\right]\Bigg|_{\beta=\beta_n^{DRO}} \succ 0$$

hold, then the adjusted WDRO estimator β_n^{ADRO} defined in (10) *exists.*

If the hypothesis function is linear, i.e., $f(\mathbf{x},\beta)=\langle \mathbf{x},\beta\rangle$, the existence conditions demonstrated in Proposition 13 could be further simplified as shown in the following proposition.

Proposition 14 (Existence of Adjusted WDRO Estimator II) *In the WDRO problem* (4)*, under Assumption 5, 7, and 8, for the empirical distribution $\mathbb{P}_n$, the loss function $L(\langle \mathbf{x},\beta\rangle,y)$ and the WDRO estimator β_n^{DRO}, if*

$$\beta_n^{DRO} \neq \mathbf{0}, \quad \frac{\partial^2 L(f,y)}{\partial f^2} > 0, \quad \mathbb{P}_n\left(\frac{\partial L(\langle \mathbf{X},\beta_n^{DRO}\rangle,Y)}{\partial f} \neq 0\right) > 0,$$

hold, where $\frac{\partial L}{\partial f}$ means taking the gradient of L w.r.t. the first argument, and there does not exist nonzero vector α such that $\mathbb{P}_n(\alpha^\top \mathbf{X}=0)=1$, then the adjusted WDRO estimator β_n^{ADRO} defined in (10) *exists.*

The conditions in Proposition 13 and 14 are mild. For example, for the nonzero WDRO estimator β_n^{DRO} and non-degenerate loss L with positive second-order derivative, if the feature variable $\mathbf{X}$ does not lie in any linear subspace of $\mathbb{R}^d$, the conditions in Proposition 14 can hold. One may check that the existence conditions could be satisfied by multiple statistical models, including linear regression and logistic regression, among many others.

4.2 Simplification of the Adjusted WDRO Estimator

In this subsection, we discuss under which conditions the expression of the adjusted WDRO estimator β_n^{ADRO} could be further simplified.

Recall that, in the definition of the adjusted WDRO estimator, seeing Definition 12, the term $H_n(\mathbf{z})$ appears complicated at first glance. The following proposition shows that the function $H_n(\mathbf{z})$ can be simplified under certain conditions.

Proposition 15 (Simplification) *If the hypothesis function in problem* (4) *is a linear function, i.e., $f(\mathbf{x},\beta)=\langle \mathbf{x},\beta\rangle$, and the equation*

$$\mathbb{E}_{P_*}\left[\frac{\partial L(\langle \mathbf{X},\beta\rangle,Y)}{\partial f}\frac{\partial L^2(\langle \mathbf{X},\beta\rangle,Y)}{\partial f^2}\mathbf{X}\right]\Bigg|_{\beta=\beta_*} = 0 \tag{13}$$

holds, then the function $H(\beta_*)$ *defined in* (8) *can be rewritten as*

$$H(\beta_*) = \tau \sqrt{\mathbb{E}_{P_*}\left[\left(\frac{\partial L(\langle \mathbf{X}, \beta\rangle, Y)}{\partial f}\right)^2\right]}\Bigg|_{\beta=\beta_*} \frac{\beta_*}{\|\beta_*\|_2}.$$

Proposition 15 implies that the linearity of the hypothesis function and the equation (13) can promise that $H(\mathbf{z})$ is a rescaling of $\mathbf{z}$. The associated function $H_n(\mathbf{z})$ is defined by

$$H_n(\mathbf{z}) = \tau \sqrt{\mathbb{E}_{\mathbb{P}_n}\left[\left(\frac{\partial L(\langle \mathbf{X}, \beta\rangle, Y)}{\partial f}\right)^2\right]}\Bigg|_{\beta=\mathbf{z}} \frac{\mathbf{z}}{\|\mathbf{z}\|_2}.$$

In this way, the expression of the adjusted WDRO estimator could be simplified. In particular, the conditions in Proposition 15 can be satisfied by multiple statistical models, e.g., linear regression, logistic regression, and Poisson regression. The details can be found in Section 5.

4.3 Asymptotically Unbiased

We establish the asymptotic distribution of the adjusted WDRO estimator β_n^{ADRO}.

Theorem 16 (Unbiasedness) *Under Assumption 5, 7, and 8, if the adjusted WDRO estimator* β_n^{ADRO} *defined in* (10) *exists, and* $\frac{\partial L(f(\mathbf{x},\beta),y)}{\partial \mathbf{x}\partial\beta}$, $\frac{\partial^2 L(f(\mathbf{x},\beta),y)}{\partial \beta^2}$ *are continuously differentiable w.r.t.* β*, then the adjusted WDRO estimator* β_n^{ADRO} *converges in distribution:*

$$\sqrt{n}(\beta_n^{ADRO} - \beta_*) \Rightarrow \mathcal{N}(0, D(\beta_*)),$$

where $D(\beta_*)$ *is defined in* (9)*.*

Theorem 16 indicates that our proposed estimator β_n^{ADRO} is asymptotically unbiased and the asymptotic mean squared error is $\mathrm{tr}(D(\beta_*))$. Recall the asymptotic distribution of the classic WDRO estimator β_n^{DRO} is

$$\sqrt{n}(\beta_n^{\mathrm{DRO}} - \beta_*) \Rightarrow \mathcal{N}\left(-C(\beta_*)^{-1}H\left(\beta_*\right), D(\beta_*)\right),$$

indicating that the asymptotic mean squared error of the classic WDRO estimator β_n^{DRO} is $\mathrm{tr}(D(\beta_*)) + f(\beta_*)^\top f(\beta_*)$, where $f(\beta_*) = -C(\beta_*)^{-1}H(\beta_*)$ might not be zero. In this way, our proposed estimator has a smaller asymptotic mean squared error.

4.4 Out-of-sample Performance Guarantee

This subsection discusses the out-of-sample performance guarantee for the adjusted WDRO estimator β_n^{ADRO}.

Informally, the out-of-sample performance guarantee for the WDRO estimator β_n^{DRO} reads that, with a high probability, the following inequality holds:

$$\mathbb{E}_{P_*}\left[L(f(\mathbf{X}, \beta_n^{DRO}), Y)\right] \leq \sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)} \mathbb{E}_P\left[L(f(\mathbf{X}, \beta_n^{DRO}), Y)\right] + \epsilon_n, \tag{14}$$

where the left-hand side is the generalization error of β_n^{DRO}, and the first term on the right-hand side is called Wasserstein robust loss of β_n^{DRO}. Inequality (14) implies that the ground-truth loss of β_n^{DRO} is upper bounded by the Wasserstein robust loss up to a higher order residual ϵ_n.

Recall that our proposed adjusted estimator β_n^{ADRO} is transformed from the WDRO estimator β_n^{DRO}. As the WDRO estimator β_n^{DRO} enjoys the out-of-sample performance guarantee (14), similar arguments can be established towards the adjusted WDRO estimator β_n^{ADRO}.

Corollary 17 (Performance Guarantee) *Suppose the generalization bound* (14) *holds for the WDRO estimator β_n^{DRO} for some residual term ϵ_n with probability $1-\alpha$. If the loss function $L(f(\mathbf{x},\beta),y)$ is h-Lipschitz continuous w.r.t. β, and the adjusted WDRO estimator β_n^{ADRO} exists, then the following inequality:*

$$\mathbb{E}_{P_*}\left[L(f(\mathbf{X},\beta_n^{ADRO}),Y)\right] \leq \sup_{P\in\mathcal{U}_{\rho n}(\mathbb{P}_n)} \mathbb{E}_P\left[L(f(\mathbf{X},\beta_n^{ADRO}),Y)\right] + \frac{h}{\sqrt{n}}\mathcal{R}_n + \epsilon_n,$$

where $\mathcal{R}_n = \mathbb{E}_{P_*}\left[\|C_n(\beta_n^{DRO})^{-1}H_n(\beta_n^{DRO})\|_2\right] + \sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)}\mathbb{E}_P\left[\|C_n(\beta_n^{DRO})^{-1}H_n(\beta_n^{DRO})\|_2\right]$, *holds with probability* $1-\alpha$.

Notably, Gao (2022) derives the generalization bound based on a novel variance-based concentration inequality for the empirical loss for the radius of the order $\mathcal{O}(1/\sqrt{n})$, where $\epsilon_n = \widetilde{\mathcal{O}}(1/n)$ ($\widetilde{\mathcal{O}}$ is used to suppress the logarithmic dependence). In this sense, Corollary 17 indicates that the generalization error of the adjusted WDRO estimator β_n^{ADRO} can be upper bounded by the Wasserstein robust loss of the adjusted WDRO estimator β_n^{ADRO} up to a new residual term, $h\mathcal{R}_n/\sqrt{n}+\epsilon_n$, which is of order $\mathcal{O}(1/\sqrt{n})$. The new residual order of the out-of-sample guarantee for the adjusted WDRO estimator may have a lower order than that of the classic WDRO estimator shown in Gao (2022). To further improve the residual order for the adjusted WDRO estimator could be considered as our future work.

5. Adjusted WDRO in the Generalized Linear Model

In this section, the generalized linear model is considered since several well-known regression models can be covered, including logistic regression, Poisson regression, and linear regression. We introduce how to develop the associated adjusted WDRO estimators in the generalized linear model.

5.1 Formulation of the Generalized Linear Model

In the generalized linear model, the response variable Y is generated from a particular distribution from the exponential family, including the Bernoulli distribution on $Y\in\{-1,1\}$ in the logistic regression, the Poisson distribution on $Y\in\{0,1,2,...\}$ in the Poisson regression, the normal distribution on $Y\in\mathbb{R}$ in the linear regression, etc. The expectation of the response variable Y conditional on the feature variable $\mathbf{X}$ is determined by the link function. With a little abuse of notation, if we denote the nonzero ground-truth parameter by β_* and the link function by G, we have $G(\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]) = \langle \mathbf{x},\beta_*\rangle$, where the link functions G is chosen as the logit function in the logistic regression, the log function in the Poisson

regression, the identity function in the linear regression, etc. If we denote the logit function, the log function, and the identity function by G^1, G^2, and G^3, respectively, we have that

$$G^1(t) = \log\left(\frac{t}{1-t}\right), \quad G^2(t) = e^t, \quad G^3(t) = t.$$

In the generalized linear model, the ground-truth parameter β_* is estimated by the maximum likelihood estimation method, and the associated loss function can be denoted by $L(f(\mathbf{x}, \beta), y) = L(\langle \mathbf{x}, \beta \rangle, y)$. If we denote the loss function in the logistic regression, the Poisson regression and the linear regression by L^1, L^2, and L^3, respectively, we have that

$$\begin{aligned} L^1\left(\langle \mathbf{x}, \beta \rangle, y\right) &= \log(1 + e^{-y\langle \mathbf{x}, \beta \rangle}), \\ L^2\left(\langle \mathbf{x}, \beta \rangle, y\right) &= e^{\langle \mathbf{x}, \beta \rangle} - y\langle \mathbf{x}, \beta \rangle, \\ L^3(\langle \mathbf{x}, \beta \rangle, y) &= \frac{1}{2}\left(\langle \mathbf{x}, \beta \rangle - y\right)^2, \end{aligned}$$

where $\beta \in B$, B is a compact convex subset of $\mathbb{R}^d$, $\beta_* \in B^\circ$, $\mathbf{x} \in \Omega$, and Ω is a convex subset of $\mathbb{R}^d$.

5.2 Asymptotic Convergence of the WDRO Estimator

This subsection derives the convergence of the WDRO estimator β_n^{DRO} in the linear regression, logistic regression, and Poisson regression.

Suppose that our choice of hyperparameters follows Assumption 5. As demonstrated in Section 3.2, we check Assumption 7 and Assumption 8 in the following lemmas.

Lemma 18 *The loss function $L^1(\langle \mathbf{x}, \beta \rangle, y)$ satisfies the conditions in Assumption 7.*

Lemma 19 *If Ω is bounded, the loss function $L^2(\langle \mathbf{x}, \beta \rangle, y)$ satisfies the conditions Assumption 7.*

Lemma 20 *The loss function $L^3(\langle \mathbf{x}, \beta \rangle, y)$ satisfies the conditions Assumption 7.*

Lemma 21 *In the logistic regression, if there does not exist nonzero vector α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, and $\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\right] < \infty$, Assumption 8 is satisfied.*

Lemma 22 *In the Poisson regression, if there does not exist nonzero vector α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, and $\mathbb{E}_{P_*}[e^{\langle \mathbf{X}, \beta_* \rangle}\|\mathbf{X}\|_2^2] < \infty$, Assumption 8 is satisfied.*

Lemma 23 *In the linear regression, if there does not exist nonzero vector α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, $\mathrm{Var}_{P_*}(Y|\mathbf{X}) < \infty$, and $\mathbb{E}_{P_*}[\|\mathbf{X}\|_2^2] < \infty$, Assumption 8 is satisfied.*

Lemma 18-20 imply that the loss functions satisfy the conditions in Assumption 7 while Lemma 21-23 show that Assumption 8 can be simplified in the logistic regression, Poisson regression, and linear regression.

Equipped with Lemma 18-23, the convergence in distribution of the WDRO estimator β_n^{DRO} can be established due to Theorem 9. The following three propositions give the explicit expression of the asymptotic distribution of the WDRO estimator for the logistic regression, Poisson regression, and linear regression.

Proposition 24 (Convergence of β_n^{DRO} in the logistic regression) *In the logistic regression, under Assumption 5, if $\Omega = \mathbb{R}^d$ and $\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\right] < \infty$, and there does not exist nonzero vector α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, the WDRO estimator β_n^{DRO} converges in distribution:*

$$\sqrt{n}(\beta_n^{DRO} - \beta_*) \Rightarrow \mathcal{N}(-C(\beta_*)^{-1}H(\beta_*), D(\beta_*)),$$

where

$$D(\beta_*) = \left(\mathbb{E}_{P_*}\left[\frac{e^{\langle \mathbf{X}, \beta_* \rangle}\mathbf{X}\mathbf{X}^\top}{\left(1 + e^{\langle \mathbf{X}, \beta_* \rangle}\right)^2}\right]\right)^{-1}, \tag{15}$$

and

$$C(\beta_*) = \mathbb{E}_{P_*}\left[\frac{e^{\langle \mathbf{X}, \beta_* \rangle}\mathbf{X}\mathbf{X}^\top}{\left(1 + e^{\langle \mathbf{X}, \beta_* \rangle}\right)^2}\right], \quad H(\beta_*) = \tau\sqrt{\mathbb{E}_{P_*}\left[\frac{e^{\langle \mathbf{X}, \beta_* \rangle}}{\left(1 + e^{\langle \mathbf{X}, \beta_* \rangle}\right)^2}\right]}\frac{\beta_*}{\|\beta_*\|_2}. \tag{16}$$

Proposition 25 (Convergence of β_n^{DRO} in the Poisson regression) *In the Poission regression, under Assumption 5, if Ω is compact and can be expressed as $\Omega = \{\mathbf{x} \in \mathbb{R}^d : A\mathbf{x} \leq b\}$, where A is an $l \times d$ matrix with linearly independent rows and $b \in \mathbb{R}^l$, $\mathbb{E}_{P_*}[\|\mathbf{X}\|_2^2 e^{\langle \mathbf{X}, \beta_* \rangle}] < \infty$, there does not exist nonzero vector α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, and $\mathbf{X}$ has a probability density which is absolutely continuous w.r.t. Lebesgue measure, the WDRO estimator β_n^{DRO} converges in distribution:*

$$\sqrt{n}(\beta_n^{DRO} - \beta_*) \Rightarrow \mathcal{N}(-C(\beta_*)^{-1}H(\beta_*), D(\beta_*)),$$

where

$$D(\beta_*) = \left(\mathbb{E}_{P_*}\left[e^{\langle \mathbf{X}, \beta_* \rangle}\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}, \tag{17}$$

and

$$C(\beta_*) = \mathbb{E}_{P_*}\left[e^{\langle \mathbf{X}, \beta_* \rangle}\mathbf{X}\mathbf{X}^\top\right], \quad H(\beta_*) = \tau\sqrt{\mathbb{E}_{P_*}[e^{\langle \mathbf{X}, \beta_* \rangle}]}\frac{\beta_*}{\|\beta_*\|_2}. \tag{18}$$

Proposition 26 (Convergence of β_n^{DRO} in the linear regression) *In the linear regression, under Assumption 5, if $\Omega = \mathbb{R}^d$, and $\mathbb{E}_{P_*}[\|\mathbf{X}\|_2^2] < \infty$, and there does not exist nonzero vector α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, the WDRO estimator β_n^{DRO} converges in distribution:*

$$\sqrt{n}(\beta_n^{DRO} - \beta_*) \Rightarrow \mathcal{N}(-C^{-1}H(\beta_*), D),$$

where

$$D = \sigma^2\left(\mathbb{E}_{P_*}\left[\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}, \tag{19}$$

$$C = \mathbb{E}_{P_*}\left[\mathbf{X}\mathbf{X}^\top\right], \quad H(\beta_*) = \tau\sigma\frac{\beta_*}{\|\beta_*\|_2}, \tag{20}$$

and $\mathrm{Var}_{P_*}(Y|\mathbf{X}) = \sigma^2, \sigma > 0$

We could obtain the associated adjusted WDRO estimators based on the convergence results derived in Proposition 24-26, and the details will be clarified in the next subsection.

Also, the proofs of Proposition 24-26 are relegated to Appendix A. The proofs show that the conditions in Proposition 15 are satisfied, which enables us to simplify the function H, seeing (16), (18) and (20).

5.3 Adjusted WDRO Estimator in the Generalized Linear Model

This subsection gives the formulations of the adjusted WDRO estimator for logistic regression, Poisson regression, and linear regression by plugging the expressions of the function C and H in (16), (18) and (20) into the definition of the adjusted WDRO estimator (10).

Definition 27 *Under assumptions in Proposition 24-26, for the nonzero WDRO estimator β_n^{DRO}, we define the adjusted WDRO estimator β_n^{ADRO} as follows,*

$$\beta_n^{ADRO} = \beta_n^{DRO} + \frac{\tau}{\sqrt{n}}\sqrt{\mathbb{E}_{\mathbb{P}_n}\left[\frac{e^{\langle \mathbf{X},\beta_n^{DRO}\rangle}}{\left(1+e^{\langle \mathbf{X},\beta_n^{DRO}\rangle}\right)^2}\right]}\left(\mathbb{E}_{\mathbb{P}_n}\left[\frac{e^{\langle \mathbf{X},\beta_n^{DRO}\rangle}\mathbf{X}\mathbf{X}^\top}{\left(1+e^{\langle \mathbf{X},\beta_n^{DRO}\rangle}\right)^2}\right]\right)^{-1}\frac{\beta_n^{DRO}}{\|\beta_n^{DRO}\|_2}, \tag{21}$$

$$\beta_n^{ADRO} = \beta_n^{DRO} + \frac{\tau}{\sqrt{n}}\sqrt{\mathbb{E}_{\mathbb{P}_n}[e^{\langle \mathbf{X},\beta_n^{DRO}\rangle}]}\left(\mathbb{E}_{\mathbb{P}_n}\left[e^{\langle \mathbf{X},\beta_n^{DRO}\rangle}\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}\frac{\beta_n^{DRO}}{\|\beta_n^{DRO}\|_2},$$

$$\beta_n^{ADRO} = \beta_n^{DRO} + \frac{\tau\sigma}{\sqrt{n}}\left(\mathbb{E}_{\mathbb{P}_n}\left[\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}\frac{\beta_n^{DRO}}{\|\beta_n^{DRO}\|_2}, \tag{22}$$

for the logistic regression, Poisson regression, and linear regression, respectively.

As we discussed in Proposition 14, one could check that the adjusted WDRO estimators defined in Definition 27 are well-defined. Then, it is easy to check that the conditions in Theorem 16, i.e., the smoothness of the loss function, hold for the logistic regression, Poisson regression, and linear regression, indicating the proposed adjustment technique could de-bias the associated adjusted WDRO estimators successfully in the logistic regression, Poisson regression, and linear regression. We conclude this result in the following proposition.

Proposition 28 *For the adjusted WDRO estimator β_n^{ADRO} defined in Definition 27, we have the following*

$$\sqrt{n}\left(\beta_n^{ADRO} - \beta_*\right) \Rightarrow \mathcal{N}(0, D(\beta_*)),$$

where $D(\beta_*)$ is defined by (15), (17), *and* (19) *in the logistic regression, Poisson regression, and linear regression, respectively.*

6. Numerical Experiments

In this section, we investigate the empirical performance of the adjusted WDRO estimator β_n^{ADRO}, compared with the classic WDRO estimator β_n^{DRO}.

6.1 Experiment Setting

The WDRO algorithmic framework of the logistic regression model and linear regression model with quadratic loss has been established in Blanchet et al. (2022c). Therefore, the adjusted WDRO estimators in the logistic regression model and the linear regression model are implemented as examples to evaluate the practical performance of our adjustment technique.

6.1.1 Logistic Regression

Suppose $\mathbf{X}$ follows 2-dimensional standard normal distribution, and the response variable Y follows the Bernoulli distribution, where $P_*(Y=1|\mathbf{X}=\mathbf{x}) = 1/(1+e^{-\langle \mathbf{x},\beta_*\rangle})$ and $\beta_* = (1/\sqrt{17}, 4/\sqrt{17})$. Data is generated 5 times for each sample size $n \in \{500, 700, 1000, 1500, 1800, 2000\}$. The WDRO estimator β_n^{DRO} is computed by the iterative algorithm in Blanchet et al. (2022c). The adjusted WDRO estimator β_n^{ADRO} is computed via equation (21). Per the iterative algorithm, we set the learning rate as 0.3 and the maximum number of iterations as 50000, respectively. Moreover, since the value of τ, which is the coefficient in the Wasserstein radius $\rho_n = \tau/\sqrt{n}$, should be determined, we let $\tau \in \{1.5, 2, 2.5, 3\}$.

6.1.2 Linear regression

Assume the feature variable $\mathbf{X}$ follows the 2-dimensional standard normal distribution, and the response variable Y follows the normal distribution, where $Y|\mathbf{X}=\mathbf{x} \sim \mathcal{N}(\langle \mathbf{x}, \beta_*\rangle, \sigma)$, $\beta_* = (3/\sqrt{10}, -1/\sqrt{10})$. We set $\sigma = 0.1$. Data is generated 5 times for each sample size $n \in \{500, 700, 1000, 1500, 1800, 2000\}$. The WDRO estimator β_n^{DRO} is computed by the iterative algorithm in Blanchet et al. (2022c). The adjusted WDRO estimator β_n^{ADRO} is computed via equation (22). Per the iterative algorithm, we set the learning rate as 0.01 and the maximum number of iterations as 50000, respectively. Then, we set the value of τ as $\tau \in \{1.5, 2, 2.5, 3\}$.

6.2 Experiment Results

The experimental results of the logistic regression are reported in Figure 2-5, and the results of the linear regression are reported in Figure 6-9.

The estimation accuracy of the estimators is evaluated by the squared error. The squared error of the estimator $\widehat{\beta}$ is defined by $\|\widehat{\beta}-\beta_*\|_2^2$. We plot the mean squared error of β_n^{DRO} and β_n^{ADRO} versus the logarithm of the sample size n, respectively. From the figures, we observe that the line of mean squared error of β_n^{DRO} is always above that of β_n^{ADRO}, illustrating that the proposed adjusted estimator has a smaller mean squared error. Recall that the adjusted WDRO estimator has a better asymptotic mean squared error in theory, while our empirical results show that the proposed estimator outperforms even when the sample size is finite. Moreover, we compute the difference of the squared error between β_n^{DRO} and β_n^{ADRO} for each run. This quantity helps evaluate the improvement achieved by the adjustment technique for each run. To visualize the improvement, we plot the boxplots for each sample size and each value of τ. The figures show that most parts of the boxplots are located above $y=0$ in the logistic regression, and all of the boxplots are located above $y=0$ in the linear regression. These observations indicate that the adjustment technique can generate a more accurate estimator for the ground-truth parameter β_*.

In addition to the squared error, we investigate the loss, i.e., the log-likelihood, of the estimators in the linear regression and the logistic regression. Similar to how we analyze the squared error, we plot the mean loss and the case-wise loss improvement. The figures show that the adjustment technique could help reduce the loss.

Overall, the adjusted WDRO estimator has better empirical performance than the classic WDRO estimator. When people plan to estimate parameters in statistical learning under the WDRO framework, the proposed adjusted estimator can be considered.

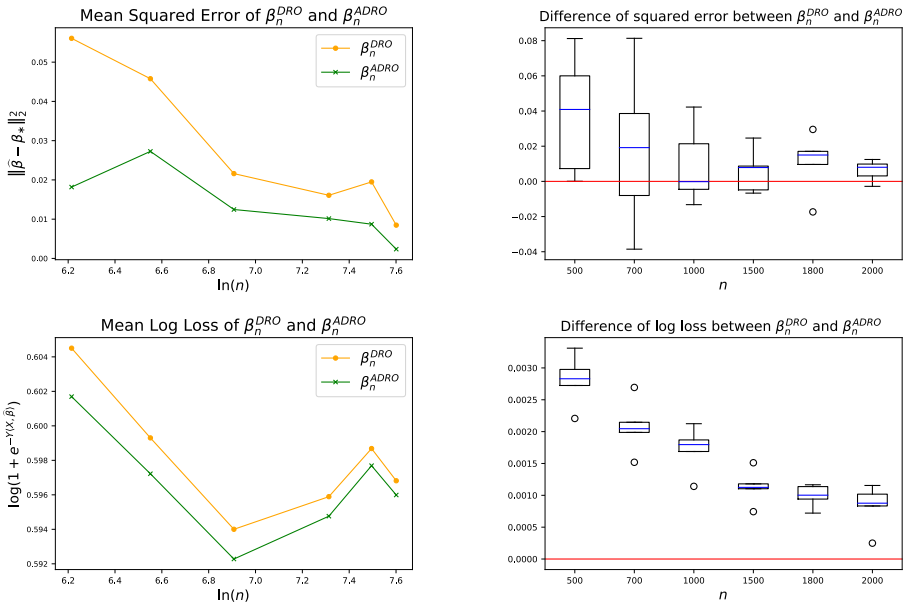

Figure 2: Squared error and log loss plots of the logistic regression, $\tau = 1.5$.

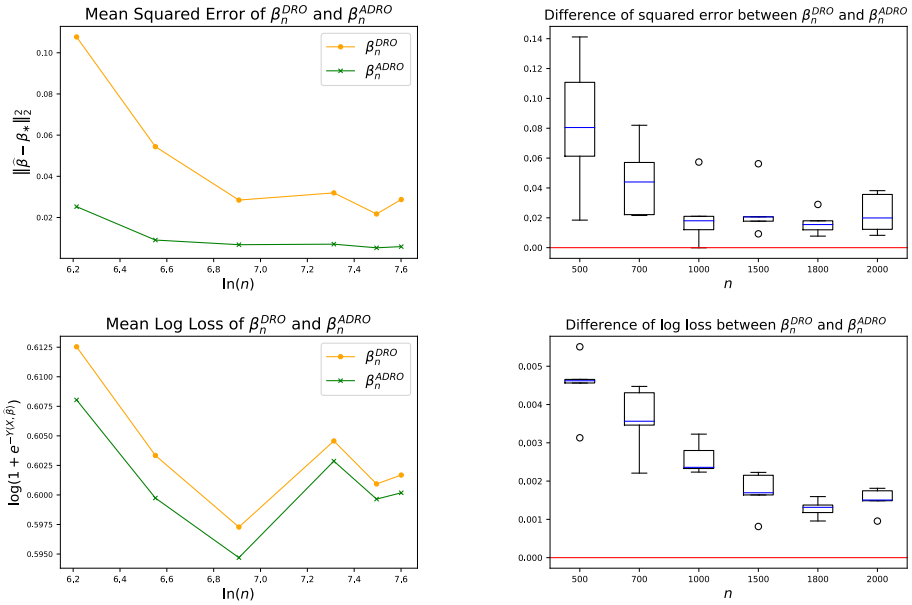

Figure 3: Squared error and log loss plots of the logistic regression, $\tau = 2$.

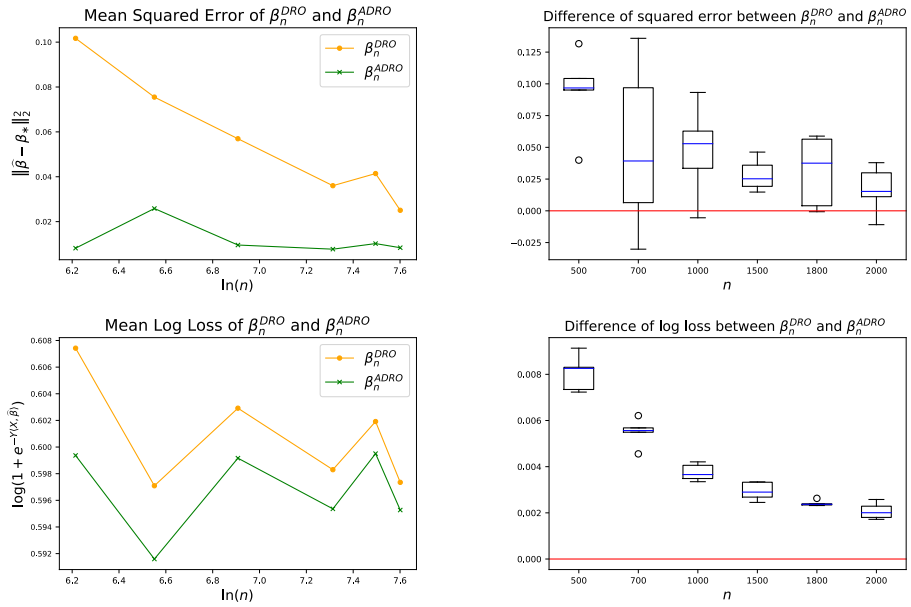

Figure 4: Squared error and log loss plots of the logistic regression, $\tau = 2.5$.

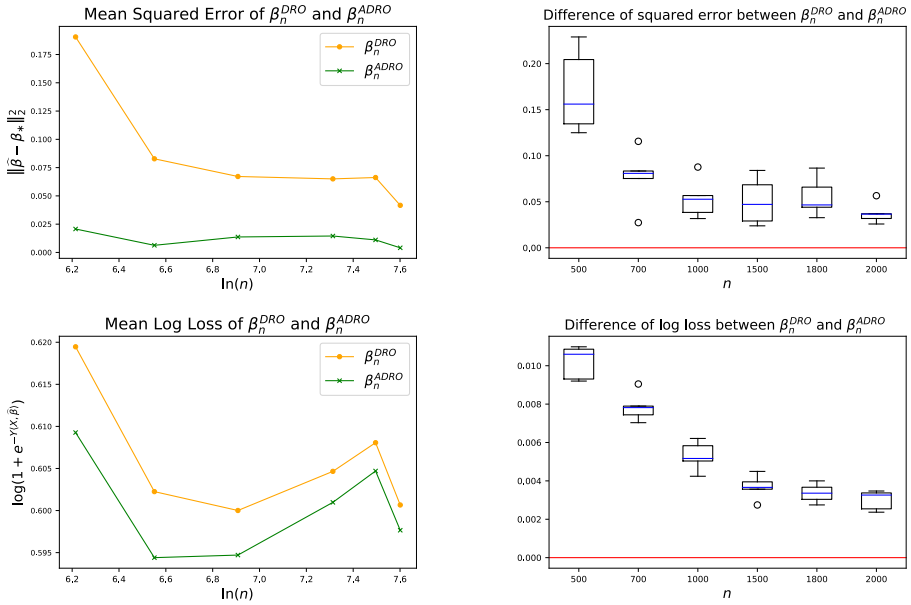

Figure 5: Squared error and log loss plots of the logistic regression, $\tau = 3$.

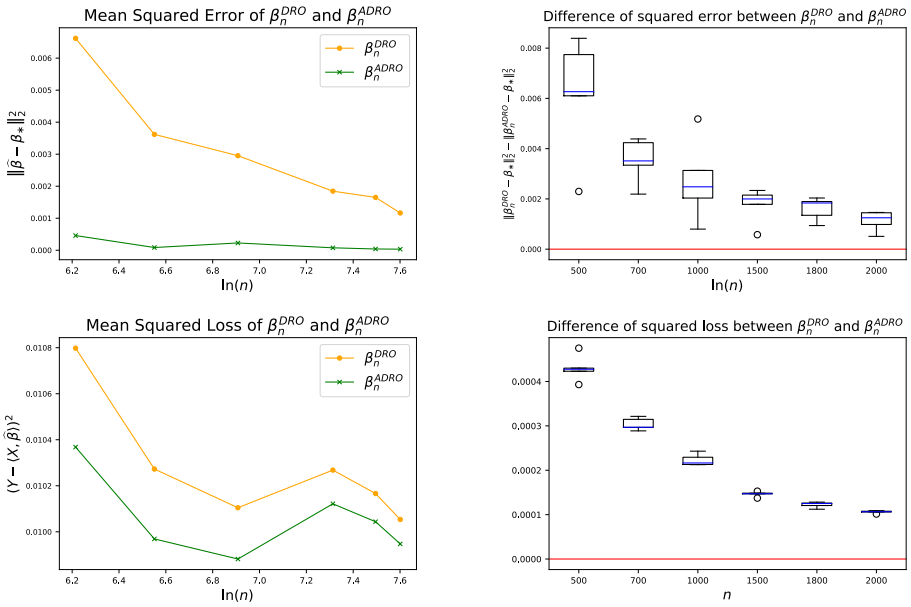

Figure 6: Squared error and squared loss plots of the linear regression, $\tau = 1.5$.

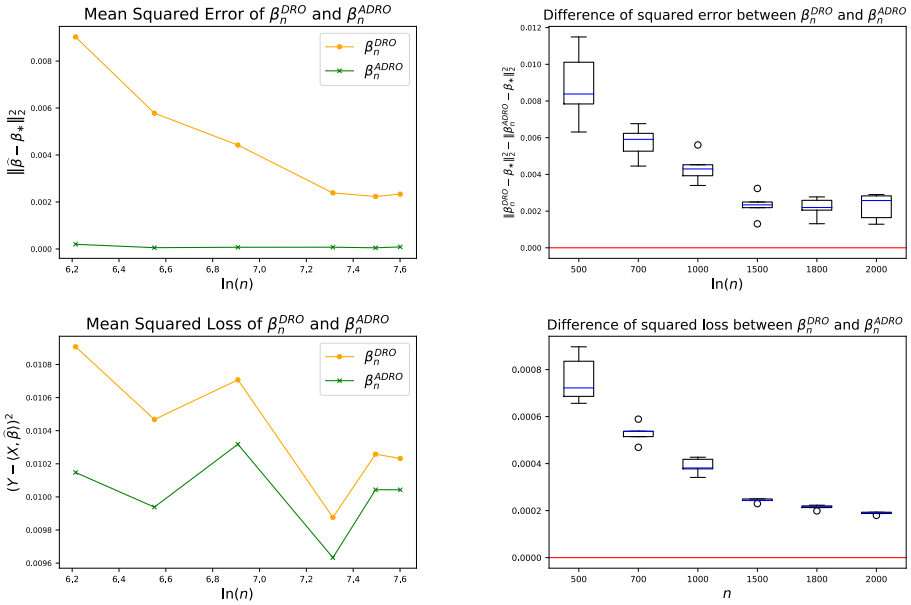

Figure 7: Squared error and squared loss plots of the linear regression, $\tau = 2$.

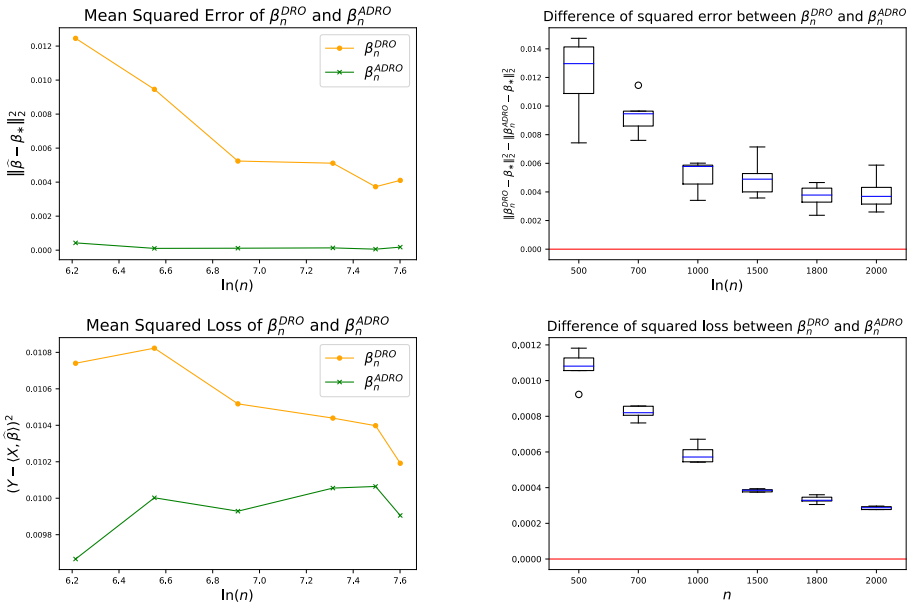

Figure 8: Squared error and squared loss plots of the linear regression, $\tau = 2.5$.

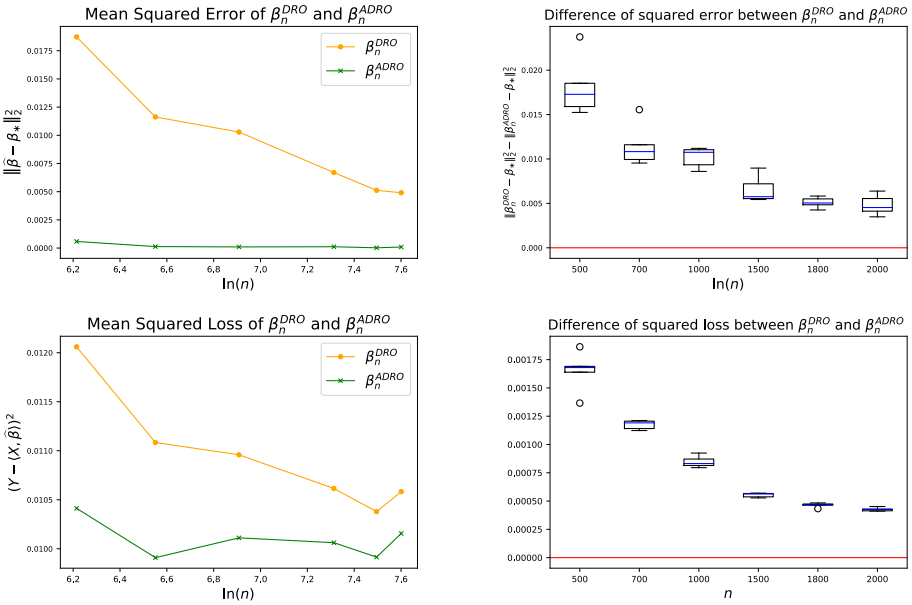

Figure 9: Squared error and squared loss plots of the linear regression, $\tau = 3$.

7. Discussion

This paper improves the performance of the WDRO estimator through the lens of the statistical asymptotics of the WDRO estimator. To the best of our knowledge, we are the first to propose transformations to de-bias the WDRO estimator asymptotically. The proposed adjusted WDRO estimator is asymptotically unbiased with a smaller asymptotic mean squared error. In addition, the adjusted WDRO estimator is easy to compute as long as the classic WDRO estimator is known. Also, we observe the superior empirical performance of the adjusted WDRO estimator over the classic WDRO estimator.

Notably, we carefully clarify and check the corresponding assumptions in the development of our theory and methodology, providing a rigorous scheme for applying and generalizing our adjustment technique.

Acknowledgments

The authors would like to thank the Action Editor and anonymous reviewers for their detailed and constructive comments, which helped greatly enhance the quality and presentation of the manuscript. The authors are partially sponsored by NSF grants CCF-1740776, DMS 2015363, and IIS-2229876. They are also partially supported by the A. Russell Chandler III Professorship at Georgia Tech.

Appendix A. Proof

A.1 Proof of Proposition 1

Proof Due to the sequential delta method, seeing Theorem 4, we have that

$$\sqrt{n}\left(\phi_n(\beta_n)-\phi_n(\beta_*)\right)\Rightarrow\mathcal{N}(\phi'(\beta_*)f(\beta_*),\phi'(\beta_*)D\phi'(\beta_*)^\top),$$

which is equivalent to

$$\sqrt{n}\left(\phi_n(\beta_n)-\beta_*\right)+\sqrt{n}\left(\beta_*-\phi_n(\beta_*)\right)\Rightarrow\mathcal{N}(\phi'(\beta_*)f(\beta_*),\phi'(\beta_*)D\phi'(\beta_*)^\top). \tag{23}$$

To make the distribution of $\sqrt{n}\left(\phi_n(\beta_n)-\beta_*\right)$, i.e., the first term in the left-hand side of (23), converge, we should require $\sqrt{n}\left(\beta_*-\phi_n(\beta_*)\right)$, i.e., the second term in the left-hand side of (23), has a finite limit. That is to say, the following holds:

$$\phi_n(\beta_*)=\beta_*+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \tag{24}$$

Since (24) holds and ϕ_n is differentiable at $\mathcal{B}(\beta_*)$, we can rewrite $\phi_n(\beta_*)$ as follows:

$$\phi_n(\beta_*)=\beta_*-\frac{1}{\sqrt{n}}g(\beta_*)+o\left(\frac{1}{\sqrt{n}}\right),$$

where $g(\beta)$ is differentiable at $\mathcal{B}(\beta_*)$.

In this way, we have that

$$\sqrt{n}\left(\phi_n(\beta_*)-\beta_*\right)=-g(\beta_*)+o(1) \tag{25}$$

In addition, (24) indicates $\phi_n'(\beta_*)\to I$, resulting in the following equivalent reformulation of (23):

$$\sqrt{n}\left(\phi_n(\beta_n)-\beta_*\right)+\sqrt{n}\left(\beta_*-\phi_n(\beta_*)\right)\Rightarrow\mathcal{N}(f(\beta_*),D). \tag{26}$$

It follows from (25), (26) and Slutsky's lemma that

$$\sqrt{n}\left(\phi_n(\beta_n)-\beta_*\right)\Rightarrow\mathcal{N}\left(f(\beta_*)-g(\beta_*),D\right).$$

In this way, the associated asymptotic mean squared error is

$$\mathrm{tr}(D)+\left(f(\beta_*)-g(\beta_*)\right)^\top\left(f(\beta_*)-g(\beta_*)\right),$$

implying that the least asymptotic mean squared error is $\mathrm{tr}(D)$ if and only $f(\beta_*)=g(\beta_*)$. ■

A.2 Proof of Theorem 3

Proof To prove (3), due to Slutsky's lemma, it suffices to show that

$$f_n(\beta_n)-f(\beta_*)\to_p 0,$$

which could be guaranteed if

$$f_n(\beta_*)-f(\beta_*)\to_p 0, \tag{27}$$

$$f_n(\beta_n) - f_n(\beta_*) \to_p 0,$$

where (27) is our assumption. Thus, it suffices to show $f_n(\beta_n) - f_n(\beta_*) \to_p 0$ holds.

Since $\sqrt{n}(\beta_n - \beta_*)$ converges to some distribution, β_n converges to β_* in probability. Since f_n is differentiable at $\mathcal{B}(\beta_*)$, it follows from the mean value theorem (or Taylor's expansion) that

$$\|f_n(\beta_n) - f_n(\beta_*)\| \leq \sup_{\beta \in \mathcal{B}(\beta_*)} \|f_n'(\beta)\| \|\beta_n - \beta_*\|,$$

It follows from $\beta_n - \beta_* \to_p 0$ and $\sup_{\beta \in \mathcal{B}(\beta_*)} \|f_n'(\beta)\|$ is bounded in probability that $f_n(\beta_n) - f_n(\beta_*) \to_p 0$. ■

A.3 Proof of Theorem 4

Proof The proof of the sequential delta method is based on the proof of the classic delta method, seeing Theorem 3.1 in van der Vaart (2000).

By the differentiablity of ϕ and ϕ_n, we have the following Taylor's expansions of ϕ_n and ϕ at ϑ:

$$\phi_n(\theta) - \phi_n(\vartheta) = \phi_n'(\vartheta)(\theta - \vartheta) + R_n,$$
$$\phi(\theta) - \phi(\vartheta) = \phi'(\vartheta)(\theta - \vartheta) + R,$$

where $\theta \in \mathcal{B}(\vartheta)$, and R_n, R are associated remainders. Note that it follows from the stated conditions that $\phi_n(\theta) \to \phi(\theta)$, $\phi_n(\vartheta) \to \phi(\vartheta)$ and $\phi_n'(\vartheta) \to \phi'(\vartheta)$. In this way, we have that $R_n \to R$, indicating that there exist N such that $|R_n| \leq 2|R|$ holds for $\forall n \geq N$. Since we have that $R = o(\|\theta - \vartheta\|)$, then $R_n = o(\|\theta - \vartheta\|)$ holds uniformly for $n \geq N$.

Since the sequence $r_n(T_n - \vartheta)$ converges in distribution, we have that $T_n - \vartheta$ converges to 0 in probability and $r_n(T_n - \vartheta)$ is uniformly tight. Then, according to the aforementioned Taylor's expansion, we have that

$$\phi_n(T_n) - \phi_n(\vartheta) = \phi_n'(\vartheta)(T_n - \vartheta) + o_p(\|T_n - \vartheta\|)$$

holds uniformly for $n \geq N$, where $o_p(1)$ means "converge to 0 in probability". Then, it follows from the uniform tightness of $r_n(T_n - \vartheta)$ that $o_p(r_n\|T_n - \vartheta\|) = o_p(1)$. That is to say,

$$r_n\left(\phi_n(T_n) - \phi_n(\vartheta)\right) = r_n\phi_n'(\vartheta)(T_n - \vartheta) + o_p(1), \tag{28}$$

holds uniformly for $n \geq N$.

Because matrix multiplication is continuous and we have $\phi_n'(\vartheta) \to \phi'(\vartheta)$, taking advantage of the extended continuous-mapping theorem, seeing Theorem 1.11.1 in Van der Vaart and Wellner (1996), we could obtain that

$$r_n\phi_n'(\vartheta)(T_n - \vartheta) \Rightarrow \mathcal{N}(\phi'(\vartheta)\mu, \phi'(\vartheta)\Sigma\phi'(\vartheta)^\top). \tag{29}$$

Further, it follows from (28), (29) and Slutsky's lemma that

$$r_n\left(\phi_n(T_n) - \phi_n(\vartheta)\right) \Rightarrow \mathcal{N}(\phi'(\vartheta)\mu, \phi'(\vartheta)\Sigma\phi'(\vartheta)^\top).$$

■

A.4 Proof of Theorem 9

Proof We denote the inner maximization of the WDRO problem (4), i.e.,

$$\max_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)} \mathbb{E}_P[L(f(\mathbf{X},\beta),Y)],$$

by $\Psi_n(\beta)$.

Then, we have

$$\Psi_n(\beta) = \inf_{\lambda\geq 0}\left[\lambda\rho_n^2 + \mathbb{E}_{(\mathbf{X},Y)\sim\mathbb{P}_n}\left[\sup_{\mathbf{x}\in\mathbb{R}^d}\left[L(f(\mathbf{x},\beta),Y) - \lambda\|\mathbf{x}-\mathbf{X}\|_2^2\right]\right]\right]. \tag{30}$$

Note that Assumption 5, 7, and 8 are extracted from Assumption 1 and 2 in Blanchet et al. (2022b), and problem (30) can be reduced to the problem in Lemma A.1 in Blanchet et al. (2022b). Following the same technique, one could derive the convergence in distribution of β_n^{DRO}:

$$\sqrt{n}(\beta_n^{DRO} - \beta_*) \Rightarrow C(\beta_*)^{-1}E - C(\beta_*)^{-1}H(\beta_*),$$

where

$$E \sim \mathcal{N}\left(0, \mathrm{Cov}\left(\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\beta}\right)\Big|_{\beta=\beta_*}\right),$$

$$H(\beta_*) = \tau\frac{\partial\sqrt{\mathbb{E}_{P_*}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}{\partial\beta}\Bigg|_{\beta=\beta_*}.$$

It follows from the matrix $C(\beta_*)$ is positive definite that

$$\sqrt{n}(\beta_n^{DRO} - \beta_*) \Rightarrow \mathcal{N}\left(-C(\beta_*)^{-1}H(\beta_*), C(\beta_*)^{-1}\mathrm{Cov}\left(\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\beta}\right)\Big|_{\beta=\beta_*} C(\beta_*)^{-1}\right).$$

■

A.5 Proof of Proposition 14

Proof Notice we have that

$$\frac{\partial^2 L(\langle\mathbf{x},\beta\rangle,y)}{\partial\beta^2} = \frac{\partial^2 L(\langle\mathbf{x},\beta\rangle,y)}{\partial f^2}\mathbf{x}\mathbf{x}^\top.$$

Since $\frac{\partial^2 L(f,y)}{\partial f^2} > 0$ and there does not exit nonzero α such that $\mathbb{P}_n(\alpha^\top\mathbf{X}=0)=1$, we have that

$$\mathbb{E}_{\mathbb{P}_n}\left[\frac{\partial^2 L(f(\mathbf{X},\beta),Y)}{\partial\beta^2}\right]\Big|_{\beta=\beta_n^{DRO}} \succ 0.$$

Notice that

$$\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial \mathbf{X}}\right\|_2^2\Bigg|_{\beta=\beta_n^{DRO}} = \left\|\frac{\partial L(\langle \mathbf{X},\beta\rangle,Y)}{\partial \mathbf{X}}\right\|_2^2\Bigg|_{\beta=\beta_n^{DRO}} = \left(\frac{\partial L(\langle \mathbf{X},\beta_n^{DRO}\rangle,Y)}{\partial f}\right)^2 \|\beta_n^{DRO}\|_2^2.$$

Since we have $\beta_n^{DRO} \neq \mathbf{0}$ and $\mathbb{P}_n\left(\frac{\partial L(\langle \mathbf{X},\beta_n^{DRO}\rangle,Y)}{\partial f} \neq 0\right) > 0$, we have that

$$\mathbb{P}_n\left(\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial \mathbf{X}}\right\|_2^2 \neq 0\right)\Bigg|_{\beta=\beta_n^{DRO}} > 0.$$

■

A.6 Proof of Proposition 15

Proof Since $f(\mathbf{x},\beta) = \langle \mathbf{x},\beta\rangle$ holds, we have that

$$\begin{aligned} H(\beta_*) &= \tau \frac{\partial \sqrt{\mathbb{E}_{P_*}\left[\left\|\frac{\partial L(\langle \mathbf{X},\beta\rangle,Y)}{\partial \mathbf{X}}\right\|_2^2\right]}}{\partial \beta}\Bigg|_{\beta=\beta_*} \\ &= \tau \frac{\partial\left(\|\beta\|_2 \sqrt{\mathbb{E}_{P_*}\left[\left(\frac{\partial L(\langle \mathbf{X},\beta\rangle,Y)}{\partial f}\right)^2\right]}\right)}{\partial \beta}\Bigg|_{\beta=\beta_*} \\ &= \tau\left(\sqrt{\mathbb{E}_{P_*}\left[\left(\frac{\partial L(\langle \mathbf{X},\beta_*\rangle,Y)}{\partial f}\right)^2\right]}\frac{\beta_*}{\|\beta_*\|_2} + \|\beta_*\|_2 \frac{\mathbb{E}_{P_*}\left[\frac{\partial L(\langle \mathbf{X},\beta_*\rangle,Y)}{\partial f}\frac{\partial^2 L(\langle \mathbf{X},\beta_*\rangle,Y)}{\partial f^2}\mathbf{X}\right]}{\sqrt{\mathbb{E}_{P_*}\left[\left(\frac{\partial L(\langle \mathbf{X},\beta_*\rangle,Y)}{\partial f}\right)^2\right]}}\right). \end{aligned} \tag{31}$$

Further, if

$$\mathbb{E}_{P_*}\left[\frac{\partial L(\langle \mathbf{X},\beta_*\rangle,Y)}{\partial f}\frac{\partial^2 L(\langle \mathbf{X},\beta_*\rangle,Y)}{\partial f^2}\mathbf{X}\right] = 0$$

holds, the second term in the equation (31) equals to 0.

Then, we have

$$H(\beta_*) = \tau\sqrt{\mathbb{E}_{P_*}\left[\left(\frac{\partial L(\langle \mathbf{X},\beta\rangle,Y)}{\partial f}\right)^2\right]}\Bigg|_{\beta=\beta_*}\frac{\beta_*}{\|\beta_*\|_2}.$$

■

A.7 Proof of Theorem 16

Proof Note we have that

$$\frac{\partial\sqrt{\mathbb{E}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}{\partial\beta}=\frac{\mathbb{E}\left[\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}\partial\beta}\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right]}{\sqrt{\mathbb{E}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}.$$

In this way, we have that

$$f(\mathbf{z})=-C(\mathbf{z})^{-1}H(\mathbf{z})$$
$$f_n(\mathbf{z})=-C_n(\mathbf{z})^{-1}H_n(\mathbf{z}),$$

where

$$C(\mathbf{z})=\mathbb{E}_{P_*}\left[\frac{\partial^2 L(f(\mathbf{X},\beta),Y)}{\partial\beta^2}\right]\Bigg|_{\beta=\mathbf{z}},\quad H(\mathbf{z})=\tau\frac{\mathbb{E}_{P_*}\left[\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}\partial\beta}\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right]}{\sqrt{\mathbb{E}_{P_*}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}\Bigg|_{\beta=\mathbf{z}},$$

$$C_n(\mathbf{z})=\mathbb{E}_{\mathbb{P}_n}\left[\frac{\partial^2 L(f(\mathbf{X},\beta),Y)}{\partial\beta^2}\right]\Bigg|_{\beta=\mathbf{z}},\quad H_n(\mathbf{z})=\tau\frac{\mathbb{E}_{\mathbb{P}_n}\left[\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}\partial\beta}\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right]}{\sqrt{\mathbb{E}_{\mathbb{P}_n}\left[\left\|\frac{\partial L(f(\mathbf{X},\beta),Y)}{\partial\mathbf{X}}\right\|_2^2\right]}}\Bigg|_{\beta=\mathbf{z}}.$$

It follows from Theorem 3 that it suffices to show f_n satisfies Assumption 2.

It follows from Assumption 7 that $L(f(\mathbf{x},\beta),y)$ is twice differentiable, $\frac{\partial L(f(\mathbf{x},\beta),y)}{\partial\mathbf{x}\partial\beta}$ and $\frac{\partial^2 L(f(\mathbf{x},\beta),y)}{\partial\beta^2}$ are differentiable w.r.t. β, indicating that both $f_n(\mathbf{z})=-C_n(\mathbf{z})^{-1}H_n(\mathbf{z})$ and $f(\mathbf{z})=-C(\mathbf{z})^{-1}H(\mathbf{z})$ are differentiable at $\mathcal{B}(\beta_*)$. The first item in Assumption 2 is satisfied.

Notably, since $L(f(\mathbf{x},\beta),y)$ is twice continuously differentiable, and $\frac{\partial L(f(\mathbf{x},\beta),y)}{\partial\mathbf{x}\partial\beta}$, $\frac{\partial^2 L(f(\mathbf{x},\beta),y)}{\partial\beta^2}$ are continuously differentiable w.r.t β, then the the gradient of $f(\mathbf{z})=-C(\mathbf{z})^{-1}H(\mathbf{z})$, i.e., $f'(\mathbf{z})$, is continuous at $\mathcal{B}(\beta_*)$. In this way, we have that $\sup_{\beta\in\mathcal{B}(\beta_*)}\|f'(\beta)\|$ is bounded. In addition, the law of large numbers implies $f_n'(\mathbf{z})\to_p f'(\mathbf{z})$ holds for every $\mathbf{z}$ at $\mathcal{B}(\beta_*)$. This convergence promises that $\sup_{\beta\in\mathcal{B}(\beta_*)}\|f_n'(\beta)\|$ is bounded in probability. The second item in Assumption 2 is satisfied.

Since $C_n(\mathbf{z})$ and $H_n(\mathbf{z})$ are defined in terms of the empirical distribution, $f_n(\beta_*)\to_p f(\beta_*)$ holds due to the law of large numbers. The third item in Assumption 2 is satisfied. ■

A.8 Proof of Corollary 17

Since the loss function $L(f(\mathbf{x},\beta),y)$ is h-Lipschitz continuous w.r.t. β, we have that

$$|L(f(\mathbf{x},\beta_n^{DRO}),y)-L(f(\mathbf{x},\beta_n^{ADRO}),y)|\le h\|\beta_n^{DRO}-\beta_n^{ADRO}\|_2,$$

indicating

$$\mathbb{E}_{P_*}\left[L(f(\mathbf{X},\beta_n^{ADRO}),Y)\right]-h\mathbb{E}_{P_*}\left[\|\beta_n^{DRO}-\beta_n^{ADRO}\|_2\right]\leq\mathbb{E}_{P_*}\left[L(f(\mathbf{X},\beta_n^{DRO}),Y)\right],$$

and

$$\begin{aligned}&\sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)}\mathbb{E}_P\left[L(f(\mathbf{X},\beta_n^{DRO}),Y)\right]\\&\leq\sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)}\mathbb{E}_P\left[L(f(\mathbf{X},\beta_n^{ADRO}),Y)\right]+h\sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)}\mathbb{E}_P\left[\|\beta_n^{DRO}-\beta_n^{ADRO}\|_2\right].\end{aligned}$$

Since we have the following definition of β_n^{ADRO}:

$$\beta_n^{ADRO}=\beta_n^{DRO}+\frac{C_n(\beta_n^{DRO})^{-1}H_n(\beta_n^{DRO})}{\sqrt{n}},$$

together with (14), we have that

$$\begin{aligned}&\mathbb{E}_{P_*}\left[L(f(\mathbf{X},\beta_n^{ADRO}),Y)\right]\leq\sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)}\mathbb{E}_P\left[L(f(\mathbf{X},\beta_n^{ADRO}),Y)\right]\\&+\frac{h}{\sqrt{n}}\left(\mathbb{E}_{P_*}\left[\|C_n(\beta_n^{DRO})^{-1}H_n(\beta_n^{DRO})\|_2\right]+\sup_{P\in\mathcal{U}_{\rho_n}(\mathbb{P}_n)}\mathbb{E}_P\left[\|C_n(\beta_n^{DRO})^{-1}H_n(\beta_n^{DRO})\|_2\right]\right)+\epsilon_n,\end{aligned}$$

holds with probability $1-\alpha$.

A.9 Proof of Lemma 18

Proof a. The loss function $L^1\left(\langle\mathbf{x},\beta\rangle,y\right)=\log(1+e^{-y\langle\mathbf{x},\beta\rangle})$ is twice continuously differentiable w.r.t. $\mathbf{x}$ and β.

b. Since we have that

$$\frac{\partial^2L^1\left(\langle\mathbf{x},\beta\rangle,y\right)}{\partial\beta^2}=\frac{e^{y\langle\mathbf{x},\beta\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{y\langle\mathbf{x},\beta\rangle}\right)^2}\succeq 0,$$

where $\succeq$ means the matrix is positive semidefinite, the function $L^1\left(\langle\mathbf{x},\beta\rangle,y\right)$ is convex w.r.t. β.

c. Note we have that

$$\begin{aligned}\left\|\frac{\partial^2L^1\left(\langle\mathbf{x},\beta\rangle,y\right)}{\partial\mathbf{x}^2}\right\|_2&=\left\|\frac{\beta\beta^\top e^{y\langle\mathbf{x},\beta\rangle}}{\left(1+e^{y\langle\mathbf{x},\beta\rangle}\right)^2}\right\|_2\\&=\|\beta\|_2^2\,\frac{e^{y\langle\mathbf{x},\beta\rangle}}{\left(1+e^{y\langle\mathbf{x},\beta\rangle}\right)^2}<M(\beta)=\|\beta\|_2^2.\end{aligned}$$

Further, we have that

$$\frac{\partial\left\|\frac{\partial^2L^1(\langle\mathbf{x},\beta\rangle,y)}{\partial\mathbf{x}^2}\right\|_2}{\partial\mathbf{x}}=\|\beta\|_2^2\,\frac{ye^{y\langle\mathbf{x},\beta\rangle}\left(1-e^{y\langle\mathbf{x},\beta\rangle}\right)}{\left(1+e^{y\langle\mathbf{x},\beta\rangle}\right)^3}\beta.$$

We know that $\frac{e^{y\langle \mathbf{x},\beta\rangle}\left(1-e^{y\langle \mathbf{x},\beta\rangle}\right)}{\left(1+e^{y\langle \mathbf{x},\beta\rangle}\right)^3}$ is bounded. Since $\beta \in B$ and B is bounded, we have that $\frac{\partial\left\|\frac{\partial^2 L^1(\langle \mathbf{x},\beta\rangle,y)}{\partial \mathbf{x}^2}\right\|_2}{\partial \mathbf{x}}$ is bounded, implying $\left\|\frac{\partial^2 L^1(\langle \mathbf{x},\beta\rangle,y)}{\partial \mathbf{x}^2}\right\|_2$ is uniformly continuous w.r.t. $\mathbf{x}$. ■

A.10 Proof of Lemma 19

Proof **a.** The loss function $L^2\left(\langle \mathbf{x},\beta\rangle, y\right) = e^{\langle \mathbf{x},\beta\rangle} - y\langle \mathbf{x},\beta\rangle$ is twice continuously differentiable w.r.t. $\mathbf{x}$ and β.

b. Because we have

$$\frac{\partial^2 L^2(\langle \mathbf{x},\beta\rangle, y)}{\partial \beta^2} = e^{\langle \mathbf{x},\beta\rangle}\mathbf{x}\mathbf{x}^\top \succeq 0,$$

the function $L^2\left(\langle \mathbf{x},\beta\rangle, y\right)$ is convex w.r.t. β.

c. We have

$$\left\|\frac{\partial^2 L^2(\langle \mathbf{x},\beta\rangle, y)}{\partial \mathbf{x}^2}\right\|_2 = \left\|\beta\beta^\top\right\|_2 e^{\langle \mathbf{x},\beta\rangle} = \|\beta\|_2^2 e^{\langle \mathbf{x},\beta\rangle}.$$

Since $\mathbf{x} \in \Omega, \beta \in B$, where both Ω and B are bounded, $\|\frac{\partial^2 L^2(\langle \mathbf{x},\beta\rangle,y)}{\partial \mathbf{x}^2}\|_2$ is bounded by a function of β and uniformly continuous w.r.t. $\mathbf{x}$. ■

A.11 Proof of Lemma 20

Proof **a.** The loss function $L^3(\langle \mathbf{x},\beta\rangle, y) = \frac{1}{2}\left(\langle \mathbf{x},\beta\rangle - y\right)^2$ is twice continuously differentiable w.r.t. $\mathbf{x}$ and β.

b. The loss function $L^3(\langle \mathbf{x},\beta\rangle, y) = \frac{1}{2}\left(\langle \mathbf{x},\beta\rangle - y\right)^2$ is convex w.r.t. β.

c. We have

$$\left\|\frac{\partial^2 L(f(\mathbf{x},\beta), y)}{\partial \mathbf{x}^2}\right\|_2 = \|2\beta\beta^\top\|_2 = 2\|\beta\|_2^2.$$

Since $\beta \in B$ and B is bounded, $\|\frac{\partial^2 L(f(\mathbf{x},\beta),y)}{\partial \mathbf{x}^2}\|_2$ is bounded by function of $2\|\beta\|_2^2$ and uniformly continuous w.r.t. $\mathbf{x}$. ■

A.12 Proof of Lemma 21

Proof **a.** From the equation

$$\frac{\partial L^1\left(\langle \mathbf{x},\beta\rangle, y\right)}{\partial \beta} = \frac{-y\mathbf{x}}{1+e^{y\langle \mathbf{x},\beta\rangle}},$$

and the assumption $\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\right] < \infty$, we have that

$$\begin{aligned}\mathbb{E}_{P_*}\left[\left\|\frac{\partial L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right\|_2^2\right]\Bigg|_{\beta=\beta_*} &= \mathbb{E}_{P_*}\left[\frac{\|\mathbf{X}\|_2^2}{(1+e^{Y\langle \mathbf{X},\beta_*\rangle})^2}\right] \\ &< \mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\right] < \infty.\end{aligned}$$

Since we have that

$$\frac{\partial^2 L^1(\langle \mathbf{x}, \beta\rangle, y)}{\partial \beta^2} = \frac{e^{y\langle \mathbf{x},\beta\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{y\langle \mathbf{x},\beta\rangle}\right)^2},$$

where

$$e^{y\langle \mathbf{x},\beta\rangle}/(1+e^{y\langle \mathbf{x},\beta\rangle})^2 > 0,$$

and there does not exist nonzero α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, then we could conclude

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L^1(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta^2}\right]\Bigg|_{\beta=\beta_*} \succ 0.$$

In addition, we have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[\frac{\partial L^1(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta}\right]\Bigg|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[\frac{-Y\mathbf{X}}{1+e^{Y\langle \mathbf{X},\beta_*\rangle}}\right]\\
=&\int P_*(Y=1|\mathbf{X}=\mathbf{x})\frac{\mathbf{x}}{1+e^{\langle \mathbf{x},\beta_*\rangle}}dF_*(\mathbf{x}) + \int P_*(Y=-1|\mathbf{X}=\mathbf{x})\frac{\mathbf{x}}{1+e^{-\langle \mathbf{x},\beta_*\rangle}}dF_*(\mathbf{x})\\
=&\int \frac{\mathbf{x}}{1+e^{-\langle \mathbf{x},\beta_*\rangle}}\frac{-1}{1+e^{\langle \mathbf{x},\beta_*\rangle}}dF_*(\mathbf{x}) + \int \frac{\mathbf{x}}{1+e^{\langle \mathbf{x},\beta_*\rangle}}\frac{1}{1+e^{-\langle \mathbf{x},\beta_*\rangle}}dF_*(\mathbf{x})\\
=&0,
\end{aligned}$$

where F_* is the distribution function of P_*.

b. Notice we have that

$$\frac{\partial L^1(\langle \mathbf{x}, \beta\rangle, y)}{\partial \mathbf{x}}\Bigg|_{\beta=\beta_*} = \frac{-y\beta_*}{1+e^{y\langle \mathbf{x},\beta_*\rangle}},$$

where

$$\beta_* \neq \mathbf{0}, y \neq 0, 1+e^{y\langle \mathbf{x},\beta_*\rangle} > 0,$$

then we can conclude that

$$P_*\left(\frac{\partial L^1(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \mathbf{X}} \neq 0\right)\Bigg|_{\beta=\beta_*} > 0.$$

Then, we have that

$$\frac{\partial^2 L^1(\langle \mathbf{x}, \beta\rangle, Y)}{\partial \mathbf{x}\partial \beta}\Bigg|_{\beta=\beta_*} = \frac{-yI_d}{1+e^{y\langle \mathbf{x},\beta_*\rangle}} + \frac{e^{y\langle \mathbf{x},\beta_*\rangle}\beta_*\mathbf{x}^\top}{\left(1+e^{y\langle \mathbf{x},\beta_*\rangle}\right)^2}.$$

Since the kernel space of the matrix $\frac{\partial^2 L^1(\langle \mathbf{x}, \beta\rangle, Y)}{\partial \mathbf{x}\partial \beta}\big|_{\beta=\beta_*}$ is different for different $\mathbf{x}, y$, we can conclude that

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L^1(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \mathbf{X}\partial \beta}\left(\frac{\partial^2 L^1(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \mathbf{X}\partial \beta}\right)^\top\right]\Bigg|_{\beta=\beta_*} \succ 0.$$

■

A.13 Proof of Lemma 22

Proof a. From the equation

$$\frac{\partial L^2\left(\langle \mathbf{x}, \beta\rangle, y\right)}{\partial \beta} = \mathbf{x}e^{\langle \mathbf{x}, \beta\rangle} - y\mathbf{x},$$

we have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[\left\|\frac{\partial L^2(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta}\right\|_2^2\right]\Bigg|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\left(e^{\langle \mathbf{X}, \beta_*\rangle} - Y\right)^2\right]\\
=&\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X}, \beta_*\rangle} - Y\right)^2\Big|\mathbf{X}\right]\right]
\end{aligned}$$

Since $Y|\mathbf{X} = \mathbf{x}$ follows the Poisson distribution with parameter $e^{\langle \mathbf{x}, \beta_*\rangle}$, we have that

$$\begin{aligned}
\mathbb{E}_{P_*}\left[\left\|\frac{\partial L^2(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta}\right\|_2^2\right]\Bigg|_{\beta=\beta_*} &= \mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\mathrm{Var}_{P_*}(Y|\mathbf{X})\right]\\
&= \mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2 e^{\langle \mathbf{X}, \beta_*\rangle}\right] < \infty.
\end{aligned}$$

Since we have that

$$\frac{\partial^2 L^2(\langle \mathbf{x}, \beta\rangle, y)}{\partial \beta^2} = e^{\langle \mathbf{x}, \beta\rangle}\mathbf{x}\mathbf{x}^\top,$$

where $e^{\langle \mathbf{x}, \beta\rangle} > 0$, and there does not exist nonzero α such that $P_*(\alpha^\top \mathbf{X} = 0) = 1$, we could conclude that

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L^2(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta^2}\right]\Bigg|_{\beta=\beta_*} \succ 0.$$

In addition, we have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[\frac{\partial L^2(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta}\right]\Bigg|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[\mathbf{X}e^{\langle \mathbf{X}, \beta_*\rangle} - Y\mathbf{X}\right]\\
=&\mathbb{E}_{P_*}\left[e^{\langle \mathbf{X}, \beta_*\rangle}\mathbf{X} - \mathbb{E}_{P_*}\left[Y|\mathbf{X}\right]\mathbf{X}\right]\\
=&0.
\end{aligned}$$

b. Notice we have that

$$\frac{\partial L^2(\langle \mathbf{x}, \beta\rangle, y)}{\partial \mathbf{x}}\Bigg|_{\beta=\beta_*} = (e^{\langle \mathbf{x}, \beta_*\rangle} - y)\beta_*,$$

where $\beta_* \neq \mathbf{0}$,

$$P_*\left(e^{\langle \mathbf{X}, \beta_*\rangle} - Y \neq 0\right) > 0,$$

then we can conclude that

$$P_*\left(\frac{\partial L^2(\langle \mathbf{X},\beta\rangle, Y)}{\partial \mathbf{X}} \neq 0\right)\Big|_{\beta=\beta_*} > 0.$$

Then, we have that

$$\frac{\partial^2 L^2(\langle \mathbf{x},\beta\rangle, Y)}{\partial \mathbf{x}\partial\beta}\Big|_{\beta=\beta_*} = (e^{\langle \mathbf{x},\beta_*\rangle} - y)I_d + e^{\langle \mathbf{x},\beta_*\rangle}\beta_*\mathbf{x}^\top.$$

Since the kernel space of the matrix $\frac{\partial^2 L^2(\langle \mathbf{x},\beta\rangle, Y)}{\partial \mathbf{x}\partial\beta}\big|_{\beta=\beta_*}$ is different for different $\mathbf{x}, y$, then we can conclude that

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L^2(\langle \mathbf{X},\beta\rangle, Y)}{\partial \mathbf{X}\partial\beta}\left(\frac{\partial^2 L^2(\langle \mathbf{X},\beta\rangle, Y)}{\partial \mathbf{X}\partial\beta}\right)^\top\right]\Big|_{\beta=\beta_*} \succ 0.$$

■

A.14 Proof of Lemma 23

Proof a. From the equation

$$\frac{\partial L^3\left(\langle \mathbf{x},\beta\rangle, y\right)}{\partial \beta} = (\langle \mathbf{x},\beta\rangle - y)\mathbf{x},$$

we have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[\left\|\frac{\partial L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right\|_2^2\right]\Big|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\left(\langle \mathbf{X},\beta_*\rangle - Y\right)^2\right]\\
=&\mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\mathbb{E}_{P_*}\left[\left(\langle \mathbf{X},\beta_*\rangle - Y\right)^2\Big|\mathbf{X}\right]\right]
\end{aligned}$$

Notice that $Y|\mathbf{X}=\mathbf{x}$ follows the normal distribution with a mean value of $\langle \mathbf{x},\beta_*\rangle$. Thus, we have that

$$\mathbb{E}_{P_*}\left[\left\|\frac{\partial L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right\|_2^2\right]\Big|_{\beta=\beta_*} = \mathbb{E}_{P_*}\left[\|\mathbf{X}\|_2^2\mathrm{Var}_{P_*}(Y|\mathbf{X})\right] < \infty. \tag{32}$$

Since we have that

$$\frac{\partial^2 L^3(\langle \mathbf{x},\beta\rangle, y)}{\partial \beta^2} = \mathbf{x}\mathbf{x}^\top,$$

and there does not exist nonzero α such that $P_*(\alpha^\top\mathbf{X}=0)=1$, we could conclude that

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta^2}\right]\Big|_{\beta=\beta_*} \succ 0.$$

In addition, we have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[\frac{\partial L^3(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \beta}\right]\Bigg|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[\langle \mathbf{X}, \beta_*\rangle \mathbf{X} - Y\mathbf{X}\right]\\
=&\mathbb{E}_{P_*}\left[\langle \mathbf{X}, \beta_*\rangle \mathbf{X} - \mathbb{E}_{P_*}\left[Y|\mathbf{X}\right]\mathbf{X}\right]\\
=&0.
\end{aligned}$$

b. Notice that,

$$\frac{\partial L^3(\langle \mathbf{x}, \beta\rangle, y)}{\partial \mathbf{x}}\Bigg|_{\beta=\beta_*} = \left(\langle \mathbf{x}, \beta_*\rangle - y\right)\beta_*,$$

where $\beta_* \neq \mathbf{0}$,

$$P_*\left(\langle \mathbf{X}, \beta_*\rangle - Y \neq 0\right) > 0,$$

then we can conclude that

$$P_*\left(\frac{\partial L^3(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \mathbf{X}} \neq 0\right)\Bigg|_{\beta=\beta_*} > 0.$$

Then, we have that

$$\frac{\partial^2 L^3(\langle \mathbf{x}, \beta\rangle, Y)}{\partial \mathbf{x}\partial \beta}\Bigg|_{\beta=\beta_*} = \left(\langle \mathbf{x}, \beta_*\rangle - y\right) I_d + \beta_* \mathbf{x}^\top.$$

Since the kernel space of the matrix $\frac{\partial^2 L^3(\langle \mathbf{x}, \beta\rangle, Y)}{\partial \mathbf{x}\partial \beta}\big|_{\beta=\beta_*}$ is different for different $\mathbf{x}, y$, we can conclude that

$$\mathbb{E}_{P_*}\left[\frac{\partial^2 L^3(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \mathbf{X}\partial \beta}\left(\frac{\partial^2 L^3(\langle \mathbf{X}, \beta\rangle, Y)}{\partial \mathbf{X}\partial \beta}\right)^\top\right]\Bigg|_{\beta=\beta_*} \succ 0.$$

■

A.15 Proof of Proposition 24

Proof Regarding the asymptotic covariance matrix, since we have that

$$
\begin{aligned}
&\mathrm{Cov}_{P_*}\left(\frac{\partial L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)\Bigg|_{\beta=\beta_*} \\
&= \mathbb{E}_{P_*}\left[\frac{\partial L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\left(\frac{\partial L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)^\top\right] \\
&= \mathbb{E}_{P_*}\left[\frac{\mathbf{X}\mathbf{X}^\top}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^2}\right] \\
&= \int P_*(Y=1|\mathbf{X}=\mathbf{x})\frac{\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) + \int P_*(Y=-1|\mathbf{X}=\mathbf{x})\frac{\mathbf{x}\mathbf{x}^\top}{\left(1+e^{-\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) \\
&= \int \frac{1}{1+e^{-\langle \mathbf{x},\beta_*\rangle}}\frac{\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) + \int \frac{1}{1+e^{\langle \mathbf{x},\beta_*\rangle}}\frac{\mathbf{x}\mathbf{x}^\top}{\left(1+e^{-\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) \\
&= \int \frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) + \int \frac{e^{2\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) \\
&= \int \frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) \\
&= \mathbb{E}_{P_*}\left[\frac{e^{\langle \mathbf{X},\beta_*\rangle}\mathbf{X}\mathbf{X}^\top}{\left(1+e^{\langle \mathbf{X},\beta_*\rangle}\right)^2}\right],
\end{aligned}
$$

and

$$
\begin{aligned}
C(\beta_*) &= \mathbb{E}_{P_*}\left[\frac{\partial^2 L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta^2}\right] \\
&= \mathbb{E}_{P_*}\left[\frac{\mathbf{X}\mathbf{X}^\top e^{Y\langle \mathbf{X},\beta_*\rangle}}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^2}\right] \\
&= \int P(Y=1|\mathbf{X}=\mathbf{x})\frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) + \int P(Y=-1|\mathbf{X}=\mathbf{x})\frac{e^{-\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{-\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) \\
&= \int \frac{1}{1+e^{-\langle \mathbf{x},\beta_*\rangle}}\frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) + \int \frac{1}{1+e^{\langle \mathbf{x},\beta_*\rangle}}\frac{e^{-\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{-\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) \\
&= \int \frac{e^{2\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) + \int \frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) \\
&= \int \frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}\mathbf{x}^\top}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x}) \\
&= \mathbb{E}_{P_*}\left[\frac{e^{\langle \mathbf{X},\beta_*\rangle}\mathbf{X}\mathbf{X}^\top}{\left(1+e^{\langle \mathbf{X},\beta_*\rangle}\right)^2}\right],
\end{aligned}
$$

then we could derive that

$$\begin{aligned} D(\beta_*) &= C(\beta_*)^{-1}\mathrm{Cov}_{P_*}\left(\frac{\partial L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)\Big|_{\beta=\beta_*} C(\beta_*)^{-1} \\ &= \left(\mathbb{E}_{P_*}\left[\frac{e^{\langle \mathbf{X},\beta_*\rangle}\mathbf{X}\mathbf{X}^\top}{\left(1+e^{\langle \mathbf{X},\beta_*\rangle}\right)^2}\right]\right)^{-1}. \end{aligned} \tag{33}$$

Regarding the asymptotic mean of β_n^{ADRO}, we have that

$$H(\beta_*) = \tau\left(\sqrt{\mathbb{E}_{P_*}\left[\frac{1}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^2}\right]}\frac{\beta_*}{\|\beta_*\|_2} - \frac{\|\beta_*\|_2\mathbb{E}_{P_*}\left[\frac{Ye^{Y\langle \mathbf{X},\beta_*\rangle}\mathbf{X}}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^3}\right]}{\sqrt{\mathbb{E}_{P_*}\left[\frac{1}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^2}\right]}}\right). \tag{34}$$

Notice we have that

$$\begin{aligned} &\mathbb{E}_{P_*}\left[\frac{Ye^{Y\langle \mathbf{X},\beta_*\rangle}\mathbf{X}}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^3}\right] \\ =&\int P_*(Y=1|\mathbf{X}=\mathbf{x})\frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) - \int P_*(Y=-1|\mathbf{X}=\mathbf{x})\frac{e^{-\langle \mathbf{x},\beta_*\rangle}\mathbf{x}}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) \\ =&\int \frac{1}{1+e^{-\langle \mathbf{x},\beta_*\rangle}}\frac{e^{\langle \mathbf{x},\beta_*\rangle}\mathbf{x}}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) - \int \frac{1}{1+e^{\langle \mathbf{x},\beta_*\rangle}}\frac{e^{-\langle \mathbf{x},\beta_*\rangle}\mathbf{x}}{\left(1+e^{-\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) \\ =&\int \frac{1}{1+e^{\langle \mathbf{x},\beta_*\rangle}}\frac{e^{2\langle \mathbf{x},\beta_*\rangle}\mathbf{x}}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) - \int \frac{1}{1+e^{\langle \mathbf{x},\beta_*\rangle}}\frac{e^{2\langle \mathbf{x},\beta_*\rangle}\mathbf{x}}{\left(1+e^{\langle \mathbf{x},\beta_*\rangle}\right)^3}dF_*(\mathbf{x}) \\ =&\mathbf{0}, \end{aligned}$$

which indicates that the equation (13) holds and the second term in (34) equals to 0.

Then, we obtain that

$$H(\beta_*) = \tau\sqrt{\mathbb{E}_{P_*}\left[\frac{1}{\left(1+e^{Y\langle \mathbf{X},\beta_*\rangle}\right)^2}\right]}\frac{\beta_*}{\|\beta_*\|_2}.$$

Notice we have that

$$
\begin{aligned}
&\mathbb{E}_{P_*}\left[\frac{1}{\left(1+e^{Y\langle\mathbf{X},\beta_*\rangle}\right)^2}\right]\\
=&\int P_*(Y=1|\mathbf{X}=\mathbf{x})\frac{1}{\left(1+e^{\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})+\int P_*(Y=-1|\mathbf{X}=\mathbf{x})\frac{1}{\left(1+e^{-\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})\\
=&\int\frac{1}{1+e^{-\langle\mathbf{x},\beta_*\rangle}}\frac{1}{\left(1+e^{\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})+\int\frac{1}{1+e^{\langle\mathbf{x},\beta_*\rangle}}\frac{1}{\left(1+e^{-\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})\\
=&\int\frac{1}{1+e^{\langle\mathbf{x},\beta_*\rangle}}\frac{e^{\langle\mathbf{x},\beta_*\rangle}}{\left(1+e^{\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})+\int\frac{1}{1+e^{\langle\mathbf{x},\beta_*\rangle}}\frac{e^{2\langle\mathbf{x},\beta_*\rangle}}{\left(1+e^{\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})\\
=&\int\frac{1}{1+e^{\langle\mathbf{x},\beta_*\rangle}}\frac{e^{\langle\mathbf{x},\beta_*\rangle}+e^{2\langle\mathbf{x},\beta_*\rangle}}{\left(1+e^{\langle\mathbf{x},\beta_*\rangle}\right)^2}dF_*(\mathbf{x})\\
=&\mathbb{E}_{P_*}\left[\frac{e^{\langle\mathbf{X},\beta_*\rangle}}{\left(1+e^{\langle\mathbf{X},\beta_*\rangle}\right)^2}\right].
\end{aligned}
$$

Then, $H(\beta_*)$ can be simplified as

$$H(\beta_*)=\tau\sqrt{\mathbb{E}_{P_*}\left[\frac{e^{\langle\mathbf{X},\beta_*\rangle}}{\left(1+e^{\langle\mathbf{X},\beta_*\rangle}\right)^2}\right]}\frac{\beta_*}{\|\beta_*\|_2}.$$

■

A.16 Proof of Proposition 25

Proof Regarding the asymptotic covariance matrix, since we have that

$$
\begin{aligned}
&\mathrm{Cov}_{P_*}\left(\frac{\partial L^2(\langle\mathbf{X},\beta\rangle,Y)}{\partial\beta}\right)\Big|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[\frac{\partial L^2(\langle\mathbf{X},\beta\rangle,Y)}{\partial\beta}\left(\frac{\partial L^2(\langle\mathbf{X},\beta\rangle,Y)}{\partial\beta}\right)^\top\right]\Big|_{\beta=\beta_*}\\
=&\mathbb{E}_{P_*}\left[(e^{\langle\mathbf{X},\beta_*\rangle}-Y)^2\mathbf{X}\mathbf{X}^\top\right],\\
=&\mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[\left(e^{\langle\mathbf{X},\beta_*\rangle}-Y\right)^2\Big|\mathbf{X}\right]\mathbf{X}\mathbf{X}^\top\right]\\
=&\mathbb{E}_{P_*}\left[\mathrm{Var}_{P_*}(Y|\mathbf{X})\mathbf{X}\mathbf{X}^\top\right]\\
=&\mathbb{E}_{P_*}\left[e^{\langle\mathbf{X},\beta_*\rangle}\mathbf{X}\mathbf{X}^\top\right],
\end{aligned}
\tag{35}
$$

and

$$C(\beta_*)=\mathbb{E}_{P_*}\left[\frac{\partial^2L^2(\langle\mathbf{X},\beta\rangle,Y)}{\partial\beta^2}\right]\Big|_{\beta=\beta_*}=\mathbb{E}_{P_*}\left[e^{\langle\mathbf{X},\beta_*\rangle}\mathbf{X}\mathbf{X}^\top\right],$$

then we could derive that

$$\begin{aligned} D(\beta_*) &= C(\beta_*)^{-1}\mathrm{Cov}_{P_*}\left(\frac{\partial L^2(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)\Big|_{\beta=\beta_*} C(\beta_*)^{-1} \\ &= \left(\mathbb{E}_{P_*}\left[e^{\langle \mathbf{X},\beta_*\rangle}\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}. \end{aligned} \tag{36}$$

Regarding the asymptotic mean of β_n^{ADRO}, we have that

$$H(\beta_*) = \tau\left(\sqrt{\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X},\beta_*\rangle}-Y\right)^2\right]}\frac{\beta_*}{\|\beta_*\|_2} - \|\beta_*\|_2\frac{\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X},\beta_*\rangle}-Y\right)e^{\langle \mathbf{X},\beta_*\rangle}\mathbf{X}\right]}{\sqrt{\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X},\beta_*\rangle}-Y\right)^2\right]}}\right), \tag{37}$$

For the second term, we have that

$$\begin{aligned} &\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X},\beta_*\rangle}-Y\right)e^{\langle \mathbf{X},\beta_*\rangle}\mathbf{X}\right] \\ =&\mathbb{E}_{P_*}\left[e^{\langle \mathbf{X},\beta_*\rangle}\mathbb{E}_{P_*}\left[e^{\langle \mathbf{X},\beta_*\rangle}-Y\middle|\mathbf{X}\right]\mathbf{X}\right] \\ =&\mathbf{0}, \end{aligned}$$

which indicates that the equation (13) holds and the second term in (37) equals to 0.

Further, we have that

$$\begin{aligned} &\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X},\beta_*\rangle}-Y\right)^2\right] \\ =&\mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[\left(e^{\langle \mathbf{X},\beta_*\rangle}-Y\right)^2\middle|\mathbf{X}\right]\right] \\ =&\mathbb{E}_{P_*}\left[\mathrm{Var}_{P_*}(Y|\mathbf{X})\right] \\ =&\mathbb{E}_{P_*}\left[e^{\langle \mathbf{X},\beta_*\rangle}\right]. \end{aligned}$$

Hence, we have that

$$H(\beta_*) = \tau\sqrt{\mathbb{E}_{P_*}[e^{\langle \mathbf{X},\beta_*\rangle}]}\frac{\beta_*}{\|\beta_*\|_2}.$$

■

A.17 Proof of Proposition 26

Proof Regarding the asymptotic covariance matrix, since we have that

$$\begin{aligned}
\mathrm{Cov}_{P_*}\left(\frac{\partial L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)\Big|_{\beta=\beta_*} &= \mathbb{E}_{P_*}\left[\frac{\partial L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\left(\frac{\partial L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)^\top\right]\Big|_{\beta=\beta_*} \\
&= \mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2\mathbf{X}\mathbf{X}^\top\right] \\
&= \mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2|\mathbf{X}\right]\mathbf{X}\mathbf{X}^\top\right] \\
&= \mathbb{E}_{P_*}\left[\mathrm{Var}_{P_*}(Y|\mathbf{X})\mathbf{X}\mathbf{X}^\top\right] \\
&= \sigma^2\mathbb{E}_{P_*}\left[\mathbf{X}\mathbf{X}^\top\right],
\end{aligned}$$

and

$$C = \mathbb{E}_{P_*}\left[\frac{\partial^2 L^3(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta^2}\right] = \mathbb{E}_{P_*}\left[\mathbf{X}\mathbf{X}^\top\right],$$

then we could derive that

$$\begin{aligned}
D &= C^{-1}\mathrm{Cov}_{P_*}\left(\frac{\partial L^1(\langle \mathbf{X},\beta\rangle, Y)}{\partial \beta}\right)\Big|_{\beta=\beta_*} C^{-1} \\
&= \sigma^2\left(\mathbb{E}_{P_*}\left[\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}.
\end{aligned}$$

Regarding the asymptotic mean of β_n^{ADRO}, it follows from (31) that

$$H(\beta_*) = \tau\left(\sqrt{\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2\right]}\frac{\beta_*}{\|\beta_*\|_2} - \frac{\|\beta_*\|_2\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)\mathbf{X}\right]}{\sqrt{\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2\right]}}\right). \tag{38}$$

For the second term, we have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)\mathbf{X}\right] \\
=&\mathbb{E}_{P_*}\left[\langle \mathbf{X},\beta_*\rangle\mathbf{X} - \mathbb{E}_{P_*}\left[Y|\mathbf{X}\right]\mathbf{X}\right] \\
=&\mathbf{0},
\end{aligned}$$

indicating that the equation (13) holds and the second term in (38) equals to 0.

Then, we obtain that

$$H(\beta_*) = \tau\sqrt{\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2\right]}\frac{\beta_*}{\|\beta_*\|_2}.$$

Notice we also have that

$$\begin{aligned}
&\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2\right] \\
=&\mathbb{E}_{P_*}\left[\mathbb{E}_{P_*}\left[(\langle \mathbf{X},\beta_*\rangle - Y)^2|\mathbf{X}\right]\right] \\
=&\mathbb{E}_{P_*}\left[\mathrm{Var}_{P_*}(Y|\mathbf{X})\right] \\
=&\sigma^2.
\end{aligned}$$

Thus, we have that

$$H(\beta_*) = \tau\sigma \frac{\beta_*}{\|\beta_*\|_2}.$$

■

References

Liviu Aolaritei, Soroosh Shafieezadeh-Abadeh, and Florian Dörfler. The performance of Wasserstein distributionally robust M-estimators in high dimensions. *arXiv preprint arXiv:2206.13269*, 2022.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

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.

Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of Wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 227–254. INFORMS, 2021.

Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Science*, 68(9):6382–6410, 2022a.

Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, 109(2):295–315, 2022b.

Jose Blanchet, Karthyek Murthy, and Fan Zhang. Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529, 2022c.

John Gunnar Carlsson, Mehdi Behroozi, and Kresimir Mihic. Wasserstein distance and the distributionally robust TSP. *Operations Research*, 66(6):1603–1624, 2018.

Ruidi Chen and Ioannis C Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19 (13), 2018.

Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 2022.

Rui Gao, Xi Chen, and Anton J Kleywegt. Distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

Jiajin Li, Sen Huang, and Anthony Man-Cho So. A first-order algorithmic framework for distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Operations Research*, 70(1):490–515, 2022.

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

A. Van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN 9780387946405. URL `https://books.google.com/books?id=0CenCW9qmp4C`.

AW van der Vaart. Asymptotic statistics. *Cambridge Books*, 2000.

Cheng Wang, Rui Gao, Wei Wei, Miadreza Shafie-khah, Tianshu Bi, and Joao PS Catalao. Risk-based distributionally robust optimal gas-power flow with Wasserstein distance. *IEEE Transactions on Power Systems*, 34(3):2190–2204, 2018.

Yiling Xie and Xiaoming Huo. Asymptotic behavior of adversarial training estimator under ℓ_∞-perturbation. *arXiv preprint arXiv:2401.15262*, 2024.

Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870, 2020.