On Generalization and Regularization via Wasserstein Distributionally Robust Optimization

Qinyu Wu* Jonathan Yu-Meng Li† Tiantian Mao‡

December 12, 2022

Abstract

Wasserstein distributionally robust optimization (DRO) has found success in operations research and machine learning applications as a powerful means to obtain solutions with favourable out-of-sample performances. Two compelling explanations for the success are the generalization bounds derived from Wasserstein DRO and the equivalency between Wasserstein DRO and the regularization scheme commonly applied in machine learning. Existing results on generalization bounds and the equivalency to regularization are largely limited to the setting where the Wasserstein ball is of a certain type and the decision criterion takes certain forms of an expected function. In this paper, we show that by focusing on Wasserstein DRO problems with affine decision rules, it is possible to obtain generalization bounds and the equivalency to regularization in a significantly broader setting where the Wasserstein ball can be of a general type and the decision criterion can be a general measure of risk, i.e., nonlinear in distributions. This allows for accommodating many important classification, regression, and risk minimization applications that have not been addressed to date using Wasserstein DRO. Our results are strong in that the generalization bounds do not suffer from the curse of dimensionality and the equivalency to regularization is exact. As a byproduct, our regularization results broaden considerably the class of Wasserstein DRO models that can be solved efficiently via regularization formulations.

Key-words: distributionally robust optimization, Wasserstein metrics, finite-sample guarantees, regularization

*Department of Statistics and Finance, University of Science and Technology of China, China. E-mail: wu051555@mail.ustc.edu.cn

†Telfer School of Management, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada. E-mail: jonathan.li@telfer.uottawa.ca

‡Department of Statistics and Finance, University of Science and Technology of China, China. E-mail: tmao@ustc.edu.cn

1 Introduction

Stochastic optimization problems of the following form

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) \tag{1}$$

arise naturally from many machine learning (ML) and operations research (OR) applications, where F represents the joint distribution of random variables $(Y,\mathbf{X})\in\{-1,1\}\times\mathbb{R}^n$, and $\boldsymbol{\beta}^\top\mathbf{X}$ is a random variable linearly depending on the decision variable $\boldsymbol{\beta}$ and can be interpreted generally as an affine decision rule. The binary random variable Y, taking values from $\{-1,1\}$, occurs most often in ML to represent binary outcomes for a classification problem. In the setup where Y is constant, the formulation (1) covers a wide array of regression problems in ML and risk minimization problems in OR. The function ρ^F generally represents a measure of risk that maps a given random variable to a real value that quantifies the riskiness of the random variable. The notion of risk in this paper is broadly defined as the undesirability of a random variable Z, i.e., Z with a larger value of $\rho^F(Z)$ is less preferable. In ML, the measure ρ^F typically takes the form of an expected function, i.e. $\rho^F(Z)=\mathbb{E}^F[\ell(Z)]$, where ℓ represents a loss function used to penalize regression or classification errors and the expected penalty $\mathbb{E}^F[\ell(Z)]$ is considered as the risk of prediction errors. In OR applications, the function ℓ often represents a disutility function and the expected disutility is considered as a measure of risk. In some important ML and OR applications, however, the measure ρ^F can go beyond an expected function. This is the case for example of ν-support vector machine (Schölkopf et al. (2000)) in ML, or Conditional Value-at-Risk (CVaR) minimization in portfolio optimization, where the measure ρ^F is a non-expected function, i.e. nonlinear in distributions.

In this paper, we study the distributionally robust counterpart of (1), i.e.

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}(F_0,\varepsilon)} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}), \tag{2}$$

for a broad class of measures ρ^F, where $\mathcal{B}(F_0,\varepsilon)$ denotes a ball of distributions centred at a reference distribution F_0 with a radius ε. In particular, we consider the case where the distance between a distribution F and the reference distribution F_0 is measured according to the optimal transportation cost of moving the probability mass from F_0 to F, also known as the Wasserstein distance (Kantorovich (1942), Villani (2009)). The transportation cost of moving a unit mass between any two points $\xi_1,\xi_2\in\mathbb{R}^n$ is most often calculated by the norm $\|\xi_1-\xi_2\|^p$ for some order $p\geqslant 1$, and the corresponding optimal transportation cost is called the type-p Wasserstein distance (Kuhn et

al. (2019)). Throughout this paper, a ball $\mathcal{B}_p(F_0, \varepsilon)$ defined based on the type-p Wasserstein distance is called type-p Wasserstein ball. While there are several other ways to measure the distance between two distributions and define the ball $\mathcal{B}(F_0, \varepsilon)$, such as ϕ-divergence (Ben-Tal et al. (2013), Hu and Hong (2018), Jiang and Guan (2016)), the optimal transportation cost has become an increasingly popular distance measure in both OR and ML, given its many desirable theoretical properties (see e.g. Gao and Kleywegt (2016), Esfahani and Kuhn (2018)). The distributionally robust optimization problem (2) with a type-p Wasserstein ball $\mathcal{B}_p(F_0, \varepsilon)$ has been studied extensively for the case where ρ^F is an expected function, i.e. $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ (see e.g. Kuhn et al. (2019), Shafieezadeh-Abadeh et al. (2019), Gao et al. (2017), Gao (2022)).

There are two key findings in this stream of works. First, in the case where the empirical distribution is chosen as the reference distribution F_0, the Wasserstein distributionally robust optimization model (2), with $\mathcal{B}(F_0, \varepsilon) := \mathcal{B}_p(F_0, \varepsilon)$ and $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$, can enjoy in some settings generalization bounds, i.e. upper confidence bounds on the out-of-sample performances (see e.g. Shafieezadeh-Abadeh et al. (2019), Gao (2022)). In particular, Gao (2022) recently shows that in the case where the Wasserstein ball $\mathcal{B}_p(F_0, \varepsilon)$ is of type-1 or 2, i.e. $p = 1, 2$, the bounds can be established with the radius ε of the Wasserstein ball $\mathcal{B}_p(F_0, \varepsilon)$ chosen in the square-root order $N^{-1/2}$, where N denotes the sample size. These bounds are in sharp contrast with, and much less conservative than, the bounds derived directly from the measure concentration property of Wasserstein distance (Esfahani and Kuhn (2018)). The latter require the radius ε to be chosen in the order of $N^{-1/\max\{2,n\}}$, where n denotes the dimension of the random vector in (2), and is suffered from the curse of dimensionality (Shafieezadeh-Abadeh et al. (2019)). Second, in the case where the Wasserstein ball is of type-1, i.e. $p = 1$, the Wasserstein distributionally robust optimization model (2) with $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ has been found equivalent to the nominal problem (1) with the addition of a regularizer on the decision variable $\boldsymbol{\beta}$ when the loss function ℓ is Lipschitz continuous (Shafieezadeh-Abadeh et al. (2019)). Blanchet et al. (2019) shows a similar equivalence relation holds when the Wasserstein ball is of type-2, i.e. $p = 2$, and the loss function ℓ is a square function. This relation to the classical regularization scheme, commonly applied in ML, provides a powerful interpretation of the distributionally robust optimization model and has stimulated considerable interest in its applications in ML and OR (see e.g. Blanchet and Kang (2021), Blanchet et al. (2019), Chen and Paschalidis (2018), Gao et al. (2017), Gao et al. (2022)).

It remains largely unclear, however, whether these findings, particularly the generalization bounds, can be carried over to the general setting of (2), where the ball $\mathcal{B}(F_0, \varepsilon)$ is a general type-p Wasserstein ball $\mathcal{B}_p(F_0, \varepsilon)$, $p \in [1, \infty]$ and ρ^F is a general measure of risk, i.e. potentially

a non-expected function. The question of how to bound out-of-sample risk for a general measure of risk ρ^F arises naturally from many risk minimization problems (see e.g. Postek et al. (2016), Li (2018), and references therein) or regression/classification problems that entail the use of a risk measure (see e.g. Rockafellar and Uryasev (2013), Rockafellar et al. (2008), Gotoh and Uryasev (2017)). To date, however, even the notion of generalization bounds has not been well defined in the literature for a general measure of risk ρ^F. It is perhaps doubtful also whether generalization bounds can be established for a general measure of risk ρ^F without suffering from the curse of dimensionality, given that such bounds are only known for the special case of type-1 and 2 Wasserstein ball and an expected function $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ (Gao (2022)), and that a general measure ρ^F can have distinctly different properties than an expected function. In fact, even in the literature of ML, the question of how to obtain generalization bounds, possibly through the classical regularization scheme, for the nominal problem (1) in general has been largely left open (see e.g. Shalev-Shwartz and Ben-David (2014)). This may have to do with the fact that existing analyses in Wasserstein DRO or ML for building generalization bounds rely heavily on exploiting the properties of expected function. As a key contribution of this paper, we show how to establish generalization bounds for the general model (2) and bypass the curse of dimensionality by leveraging the structure of affine decision rules. Our result may be viewed as a generalization of the result in Shafieezadeh-Abadeh et al. (2019), who study the special case of (2) with the type-1 Wasserstein ball and an expected function as the measure ρ^F. Their approach to derive generalization bounds with the $N^{-1/2}$-rate, i.e. with the radius ε chosen in the square-root order, is fundamentally different from ours and can hardly be applicable beyond this special case, since it relies heavily on the equivalency of this special case to a regularized model. Gao (2022) is a recent effort on studying how to break the curse of dimensionality in studying generalization bounds for a more general setting of Wasserstein DRO. While the setting in Gao (2022) is more general than ours in that it does not impose the structure of affine decision rules, it is more restrictive in that the measure ρ^F is largely limited to an expected function. The approach of Gao (2022) not only relies on the properties of expected function, but also applicable only to type-1 and 2 Wasserstein balls. Given the reliance of existing approaches on the properties of expected function, one may suspect the availability of $N^{-1/2}$-rate bounds may hinge heavily on the functional form of ρ^F. Yet, as shown in this paper, it turns out that the bounds can be established (almost) independently from the form of ρ^F and thus for the model (2) in great generality. Our finding that the Wasserstein DRO model (2) can break the curse of dimensionality virtually for any problem with affine decision rules offers a solid theoretical underpinning for justifying the power of Wasserstein DRO in general out-of-sample tests.

Another focus of this work is on investigating whether the other key finding of Wasserstein DRO, namely its equivalency to the classical regularization scheme in ML, can also be carried over to a more general setting in (2). To answer this, we pay particular attention first to the setting where the Wasserstein ball $\mathcal{B}(F_0, \varepsilon)$ is of any type, i.e. $p \in [1, \infty]$, and the measure ρ^F is an expected function $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ with a loss function ℓ having a growth order p. The setting includes, as a special case, the distributionally robust least square regression problem studied in Blanchet et al. (2019) when $p = 2$ and the loss function ℓ is a square function, and many other importance instances of classification, regression, and risk minimization problems (see Section 2.2). We show that for many loss functions ℓ arising from practical applications, it is possible to build an exact relation between the Wasserstein DRO model (2) and the classical regularization scheme in more general terms. To demonstrate the generality of our result, we prove something stronger - there is no loss function ℓ, other than the ones we identify, under which the exact relation can hold. This gives a definite answer as to how far one can interpret the Wasserstein DRO model from a classical regularization perspective. Our regularization results reveal also the tractability of solving the Wasserstein DRO model for many non-Lipschitz continuous loss functions ℓ and higher-order Wasserstein balls, i.e. $p > 1$. It came to our attention that the recent work of Sim et al. (2021) proposes to solve Wasserstein distributionally robust classification and regression problems for non-Lipschitz continuous loss functions ℓ and higher order Wasserstein balls. Yet, their focus is on solving the problems by an approximation approach, given the challenge of identifying tractable reformulations for the Wasserstein DRO problems. Our results, on the other hand, show the cases that can be solved exactly via regularization reformulations. The results also enable us to discover the equivalency relation in the setting where the measure ρ^F is a non-expected function. To the best of our knowledge, in this setting, the exact relation between the Wasserstein DRO model (2) and classical regularization has been known only for the case where the measure ρ^F is variance (Blanchet et al. (2022)) or a distortion risk measure (Wozabal (2014)). We show that the equivalency relation exists for a larger family of measures, including for instance higher-order risk measures (Krokhmal (2007)) and other measures emerging more recently from the literature of OR and ML (Rockafellar and Uryasev (2013), Rockafellar et al. (2008), Gotoh and Uryasev (2017)).

Related work

From the perspective of generalization bounds. A series of works done by Blanchet et al. (2019), Blanchet and Kang (2021), Blanchet et al. (2022) take a different approach to tackle the curse of dimensionality. They study the classical setting of Wasserstein DRO, where the measure ρ^F is an

expected function, and propose a radius selection rule for Wasserstein balls. They show the rule can be applied to build a confidence region of the optimal solution, and the radius can be chosen in the square-root order as the sample size goes to infinity. Although this allows for bypassing the curse of dimensionality, the bounds derived from the rule are only valid in the asymptotic sense. Blanchet et al. (2022) also takes this approach to obtain generalization bounds for mean-variance portfolio selection problems. On the other hand, the generalization bounds established in this paper, like those in Shafieezadeh-Abadeh et al. (2019), Chen and Paschalidis (2018), and Gao (2022), break the curse of dimensionality in a non-asymptotic sense, i.e. applicable to any finite sample size and dimension.

From the perspective of the equivalency between Wasserstein DRO and regularization. While the focus of this work is on studying the exact equivalency between Wasserstein DRO and regularization, there is an active stream of works studying the asymptotic equivalence in the setting where the measure ρ^F is an expected function (see Gao et al. (2017); Blanchet et al. (2019); Blanchet et al. (2022); Volpi et al. (2018); Bartl et al. (2020)). In particular, Gao et al. (2022) introduce the notion of variation regularization and show that for a broad class of loss functions, Wasserstein DRO is asymptotically equivalent to a variation regularization problem.

The rest of the paper is organized as follows. In Section 2, we start with some preliminaries regarding Wasserstein distributionally robust optimization and then provide a list of examples for different measures ρ^F. In Section 3, we provide generalization bounds for the general formulation (2). We perform in Section 4 a systematic study of the equivalency between the distributionally robust optimization model (2) and the regularized versions of the problem (1). Section 5 concludes the paper.

2 Wasserstein Distributionally Robust Optimization Model

2.1 Preliminaries

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an atomless probability space. A random vector ξ is a measurable mapping from Ω to $\mathbb{R}^{n+1}$, $n \in \mathbb{N}$. Denote by F_ξ the distribution of ξ under $\mathbb{P}$. For $p \geqslant 1$, let $\mathcal{M}_p := \mathcal{M}_p(\Xi)$ be the set of all distributions on $\Xi \subseteq \mathbb{R}^{n+1}$ with finite pth moment in each component. Denote q as the Hölder conjugate of p, i.e. $1/p + 1/q = 1$. Recall that given any two distributions $F_1 \in \mathcal{M}_p$ and $F_2 \in \mathcal{M}_p$, the type-p Wasserstein metric is defined as

$$W_{d,p}(F_1, F_2) := \left(\inf_{\pi \in \Pi(F_1, F_2)} \mathbb{E}^\pi [d(\xi_1, \xi_2)^p] \right)^{1/p}, \tag{3}$$

where $d(\cdot,\cdot) : \Xi \times \Xi \to \mathbb{R}_{\geqslant 0} \cup \{+\infty\}$ is a metric on Ξ. The set $\Pi(F_1, F_2)$ denotes the set of all joint distributions of $\xi_1 \in \Xi$ and $\xi_2 \in \Xi$ with marginals F_1 and F_2 respectively. The metric is often interpreted as the minimal transportation cost of moving the mass from the distribution F_1 to the distribution F_2 with the cost calculated according to the chosen function $d(\xi_1, \xi_2)^p$.

Consider now the distributionally robust optimization problem (2), where a ball of distributions needs to be defined for random variables $\xi = (Y, \mathbf{X}) \in \Xi$ with $\Xi = \{-1, 1\} \times \mathbb{R}^n \subseteq \mathbb{R}^{n+1}$. We apply the type-$p$ Wasserstein metric (3) with $d(\xi_1, \xi_2) = d((Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2))$, defined by the following additively separable form

$$d((Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2)) := \|\mathbf{X}_1 - \mathbf{X}_2\| + \Theta(Y_1 - Y_2), \tag{4}$$

where $\|\cdot\|$ is any given norm on $\mathbb{R}^n$ with its dual norm $\|\cdot\|_*$ defined by $\|\mathbf{y}\|_* = \sup_{\|\mathbf{x}\| \leqslant 1} \mathbf{x}^\top \mathbf{y}$, and $\Theta : \mathbb{R} \to \{0, \infty\}$ satisfies $\Theta(s) = 0$ if $s = 0$ and $\Theta(s) = \infty$ otherwise. That is, the function (4) assigns an infinitely large cost to any discrepancy in Y, i.e. $Y_1 - Y_2 \neq 0$, and reduces to a general norm on $\mathbf{X}$ when there is no discrepancy in Y, i.e. $Y_1 - Y_2 = 0$. With this form of norm, we define the ball of distribution $\overline{\mathcal{B}}_p(F_0, \varepsilon)$ by

$$\overline{\mathcal{B}}_p(F_0, \varepsilon) = \{F \in \mathcal{M}_p : W_{d,p}(F, F_0) \leqslant \varepsilon\}, \tag{5}$$

and call it type-p Wasserstein ball throughout this paper. In the remainder of this paper, we will show that the distributionally robust optimization problem (2) with the above definition of Wasserstein ball, i.e.

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \overline{\mathcal{B}}_p(F_0, \varepsilon)} \rho^F(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}) \tag{6}$$

can enjoy generalization guarantees for a general measure ρ^F and any type-p Wasserstein ball. Thereafter, we will shed light on this general guarantee by drawing the connection between the distributionally robust optimization model (6) and the classical regularization scheme in Section 4.

We first highlight in the next section the generality of the model (6) to accommodate a wide array of measures ρ^F arising from machine learning and operations research applications.

2.2 Classification, Regression, and Risk Minimization

2.2.1 Classification

The distributionally robust optimization model (6) can be directly applied to classification problems in ML, where the random variable $Y \in \{-1, 1\}$ is often termed the output and the random variables $\mathbf{X} \in \mathbb{R}^n$ are considered as the input. The use of a linear scoring function $\boldsymbol{\beta}^\top \mathbf{X}$ remains the most popular approach in classification to predict the sign of the output Y, and the product $Y \cdot \boldsymbol{\beta}^\top \mathbf{X}$ would be positive if the prediction is correct and negative otherwise. The set $\mathcal{D}$ encodes the prior knowledge of the decision variables $\boldsymbol{\beta}$, also called the weight variables. Distributionally robust classification problems (6) that minimize the worst-case expected prediction error, i.e. $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$, have been studied in Shafieezadeh-Abadeh et al. (2019) for the case where $p = 1$, i.e. the type-1 Wasserstein ball, and the loss function ℓ is convex Lipschitz continuous. We present several examples of ρ^F below for classification, where either the loss function ℓ is not Lipschitz continuous or the measure ρ^F is not an expected function.

Example (i): Higher-order hinge loss $\rho(Z) = \mathbb{E}[(1 - Z)_+^s]$, $s \geqslant 1$.

In the case $s = 2$, the classification model is also known as smooth support vector machine (SSVM), first proposed in Lee and Mangasarian (2001).

Example (ii): Higher-order SVM $\rho(Z) = \mathbb{E}[|1 - Z|^s]$, $s \geqslant 1$.

In the case $s = 2$, the classification model is widely known as least-squares support vector machine (LS-SVM) (Suykens and Vandewalle (1999)).

Example (iii): Sum-Exp $\rho(Z) = \frac{1}{t}\left(\mathbb{E}[e^{-tZ}]\right)$, $t > 0$.

This measure appears in the popular classification model, AdaBoost (Freund and Schapire (1997)).

Example (iv): ν-support vector machine $\rho(Z) = \mathrm{CVaR}_\alpha(-Z)$, $\alpha \in (0, 1)$, where CVaR denotes the conditional value-at-risk (Rockafellar and Uryasev (2002)) defined by

$$\mathrm{CVaR}_\alpha(Z) = \frac{1}{1-\alpha}\int_\alpha^1 F_Z^{-1}(s)\mathrm{d}s,$$

where F_Z^{-1} is the quantile function of Z. It is known that ν-support vector machine, introduced in the seminal work of Schölkopf et al. (2000), employs CVaR as the measure (see e.g. Gotoh and Uryasev (2017)).

2.2.2 Regression

The distributionally robust optimization model (6) can also be applied to regression problems in ML. Namely, without loss of generality, the output in regression problems can be represented by the first random variable, denoted by X_1, in $\mathbf{X}$ given that the output in regression takes real values, whereas the input can be represented by the rest of the random variables, denoted by $\mathbf{X}_{(2,:)}$, in $\mathbf{X}$. By setting $\boldsymbol{\beta} := (1, -\boldsymbol{\beta}_r)$, $\boldsymbol{\beta}_r \in \mathcal{D}$, and F_0 be a reference distribution satisfying $F_0(\{Y = 1\}) = 1$ in (6), we arrive at the following distributionally robust regression model

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \sup_{F \in \overline{\mathcal{B}}_p(F_0, \varepsilon)} \rho^F((1, -\boldsymbol{\beta}_r)^\top \mathbf{X}). \tag{7}$$

The above model seeks a linear regressor $\boldsymbol{\beta}_r^\top \mathbf{X}_{(2,:)}$ to predict the output $X_1 \in \mathbb{R}$ from the input $X_{(2,:)}$. It has been studied in Shafieezadeh-Abadeh et al. (2019) for the case where $p = 1$, i.e. the type-1 Wasserstein ball, and $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ with a convex Lipschitz continuous loss function ℓ. Several other examples of ρ^F are presented below.

Example (i): Higher-order regression $\rho(Z) = \mathbb{E}[|Z|^s]$, $s \geqslant 1$.

In the case $s = 2$, the regression model is the well-known least-square regression.

Example (ii): Higher-order c-insensitive regression $\rho(Z) = \mathbb{E}[(|Z| - c)_+^s]$, $s \geqslant 1$ and $c \geqslant 0$.

The regression model is the well-known c-insensitive support vector regression (c-SVR) (Drucker et al. (1997)) in the case $s = 1$ and c-smooth support vector regression (c-SSVR) (Lee et al. (2005)) in the case $s = 2$.

Example (iii): ν-support vector regression $\rho(Z) = \mathrm{CVaR}_\alpha(|Z|)$, $\alpha \in (0, 1)$.

ν-support vector regression (Schölkopf et al. (1998)) is a popular alternative to the c-insensitive support vector regression. It allows for bypassing the difficulty of specifying the insensitivity parameter c in the c-insensitive support vector regression.

2.2.3 Risk minimization

The distributionally robust optimization model (6) can also accommodate a general risk minimization problem, taking the form of

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \overline{\mathcal{B}}_p(F_0, \varepsilon)} \rho^F(\boldsymbol{\beta}^\top \mathbf{X}). \tag{8}$$

Table 1: Risk functions analyzed in this paper.

Applications	**Risk function**	**Formulation**
Classification	Higher-order hinge loss	$\mathbb{E}[(1-Z)_+^s]$, $s \geqslant 1$
	Higher-order SVM	$\mathbb{E}[\lvert 1-Z\rvert^s]$, $s \geqslant 1$
	Sum-Exp	$\frac{1}{t}\mathbb{E}[e^{-tZ}]$, $t > 0$
	ν-support vector machine	$\mathrm{CVaR}_\alpha(-Z)$, $\alpha \in (0,1)$
Regression	Higher-order regression	$\mathbb{E}[\lvert Z\rvert^s]$, $s \geqslant 1$
	Higher-order c-insensitive	$\mathbb{E}[(\lvert Z\rvert - c)_+^s]$, $s \geqslant 1$, $c \geqslant 0$
	ν-support vector regression	$\mathrm{CVaR}_\alpha(\lvert Z\rvert)$, $\alpha \in (0,1)$
Risk minimization	Lower partial moments	$\mathbb{E}[(Z-c)_+^s]$, $s \geqslant 1$, $c \in \mathbb{R}$
	CVaR-Deviation	$\mathrm{CVaR}_\alpha(Z - \mathbb{E}[Z])$, $\alpha \in (0,1)$
	Higher moment risk measure	$\inf_{t\in\mathbb{R}}\{t + c(\mathbb{E}[(Z-t)_+^s])^{1/s}\}$, $s, c \geqslant 1$

The most classical example of a risk minimization problem is portfolio optimization, in which case the variable $\mathbf{X}$ represents a random vector of losses from d different financial assets and $\boldsymbol{\beta}$ denotes a portfolio vector.

Example (i): Lower partial moments (LPM) $\rho(Z) = \mathbb{E}[(Z-c)_+^s]$, $s \geqslant 1$ and $c \in \mathbb{R}$.

Lower partial moments represent an important class of downside risk measure, first introduced by Bawa (1975) and Fishburn (1977) and more recently studied by Chen et al. (2011) in DRO.

Example (ii): CVaR-Deviation $\rho(Z) = \mathrm{CVaR}_\alpha(Z - \mathbb{E}[Z])$, $\alpha \in (0,1)$.

This represents an important example of deviation measures built upon risk measures (Rockafellar et al. (2008)). Its generalization can be found in Section 4.

Example (iii): Higher moment coherent risk measures $\rho(Z) = \inf_{t\in\mathbb{R}}\{t + c(\mathbb{E}[(Z-t)_+^s])^{1/s}\}$, $s, c \geqslant 1$.

This measure is a well-known generalization of CVaR that is closely related to lower partial moments (LPM) and compatible with the second order stochastic dominance and utility theory (Krokhmal (2007)).

We summarize in Table 1 a list of measures ρ^F considered in this paper.

3 Generalization bounds

Wasserstein distributionally robust optimization is most commonly applied in a data-driven setting, where the joint distribution F of random variables $(Y, \mathbf{X}) \in \{-1, 1\} \times \mathbb{R}^n$ can only be partially observed through sample data $(\widehat{y}_i, \widehat{\mathbf{x}}_i)$, $i = 1, ..., N$, independently drawn from F. In this setting, the empirical distribution, i.e. $\widehat{F}_N := \frac{1}{N}\sum_{i=1}^N \delta_{(\widehat{y}_i, \widehat{\mathbf{x}}_i)}$, is chosen as the reference distribution F_0 in Wasserstein DRO where $\delta_{\mathbf{x}}$ denotes the point-mass at $\mathbf{x}$. The key question underlying Wasserstein DRO is whether there exist upper confidence bounds on the out-of-sample performances, i.e. generalization bounds, that can scale gracefully in the dimensionality of the random vector $(Y, \mathbf{X})$, i.e. breaking the curse of dimensionality. Such out-of-sample performance guarantees are only known for the case where the Wasserstein ball is of order $p \in [1, 2]$ and the measure ρ is an expected function, i.e. $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ for some ℓ (Gao (2022), Shafieezadeh-Abadeh et al. (2019) and Chen and Paschalidis (2018)). Existing analyses hinge on the specific choice of the order p and loss functions ℓ.

In this section, we seek to answer the question of whether generalization bounds exist for a more general setting of the Wasserstein DRO model (6). Let ρ^F be a distribution-invariant measure, i.e. $\rho^F(Z_1) = \rho^F(Z_2)$ for any $Z_1 \equiv_F Z_2$ and F^* denote the true distribution of $(Y, \mathbf{X})$. We say that generalization bounds exist if the following condition can hold with high probability

$$\rho^{F^*}(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}) \leqslant \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N)} \rho^F(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}) + \tau_N, \quad \forall\, \boldsymbol{\beta} \in \mathcal{D},$$

where the radius ε_N can decrease in the order of $1/\sqrt{N}$ (or $\sqrt{(\log N)/N}$) and the residual τ_N diminishes in the order of $1/N$. The bounds, if exist, would break the curse of dimensionality. The order $1/\sqrt{N}$ for the radius follows closely the square root law.

We take two steps to study generalization bounds. First, in Section 3.1 we study how to build a confidence bound on the out-of-sample performance for a single solution $\boldsymbol{\beta}$. Then in Section 3.2, we extend the result to all $\boldsymbol{\beta} \in \mathcal{D}$ by building a union bound on the out-of-sample performances for all $\boldsymbol{\beta} \in \mathcal{D}$, i.e. the generalization bound. The first step is critical. As shown below, quite surprisingly, a confidence bound can be built for any measure ρ and a Wasserstein ball of any type.

3.1 Confidence bounds on out-of-sample risks

The notion of out-of-sample performance has not been well defined and studied for a general measure ρ. Throughout this section, we adopt the following definitions.

Definition 1. Given a solution $\boldsymbol{\beta} \in \mathcal{D}$, its (Wasserstein) in-sample risk is defined by

$$\sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon)} \rho^F(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}), \tag{9}$$

whereas its out-of-sample risk is defined by

$$\rho^{F^*}(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}).$$

Let $\widehat{J}_N$ denote the in-sample risk achieved by the solution of the DRO model (6), i.e.

$$\widehat{J}_N = \inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon)} \rho^F(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}), \tag{10}$$

and J_{oos} denote the out-of-sample risk of the solution, i.e.

$$J_{oos} = \rho^{F^*}(Y \cdot \widehat{\boldsymbol{\beta}}_N^\top \mathbf{X}), \tag{11}$$

where $\widehat{\boldsymbol{\beta}}_N$ is the optimal solution of the problem (10).

We show that under the following assumptions, the out-of-sample risk of a solution $\boldsymbol{\beta} \in \mathcal{D}$ can be bounded by its in-sample risk with high probability.

Assumption 1 (Light-tailed distribution)**.** *There exists an exponent $a > p$ such that*

$$A := \mathbb{E}^{F^*}[\exp(\|\mathbf{X}\|^a)] < +\infty.$$

Assumption 2. *The feasible set $\mathcal{D} \subseteq \mathbb{R}^n$ is bounded, that is $U_{\mathcal{D}} := \sup_{\boldsymbol{\beta} \in \mathcal{D}} \|\boldsymbol{\beta}\|_* < +\infty$.*

Assumption 3. *The feasible set $\mathcal{D} \subseteq \mathbb{R}^n$ is away from the origin, that is $L_{\mathcal{D}} := \inf_{\boldsymbol{\beta} \in \mathcal{D}} \|\boldsymbol{\beta}\|_* > 0$.*

Assumption 1 requires the tail of the distribution F^* decays at an exponential rate, which is a common assumption in Wasserstein DRO. It enables invoking the measure concentration property of Wasserstein metrics (Fournier and Guillin (2015)). Assumptions 2 and 3 impose some restrictions on the decision space $\mathcal{D}$. It is interesting to note that these two assumptions appeared also in the earlier work of Shafieezadeh-Abadeh et al. (2019), even though the approach taken in Shafieezadeh-Abadeh et al. (2019) to build $1/\sqrt{N}$-rate confidence bounds is fundamentally different from, and more restrictive than, our approach. Their approach applies only to the case where the Wasserstein ball is of type-1 and the measure ρ^F is an expected function $\rho^F(Z) = \mathbb{E}^F[\ell(Z)]$ with the loss function

ℓ assumed to be convex Lipschitz continuous. Our approach, on the other hand, allows for building $1/\sqrt{N}$-rate confidence bounds for any type of Wasserstein ball and measure ρ.

Before delving into the details of our approach, we present first the confidence bounds obtained from our approach, as the main result of this section.

Theorem 1. *Suppose that Assumptions 1, 2 and 3 hold and $\eta \in (0,1)$. By setting the radius ε in problem (9) as $\varepsilon_{p,N}(\eta)/L_{\mathcal{D}}$, where*

$$\varepsilon_{p,N}(\eta) = \begin{cases} \left(\frac{\log(c_1\eta^{-1})}{c_2 N}\right)^{1/2}, & \text{if } N \geqslant \frac{\log(c_1\eta^{-1})}{c_2}, \\ \left(\frac{\log(c_1\eta^{-1})}{c_2 N}\right)^{p/a}, & \text{if } N < \frac{\log(c_1\eta^{-1})}{c_2}, \end{cases} \tag{12}$$

and c_1, c_2 are positive constants that only depend on $U_{\mathcal{D}}$, a, A, and p[1], we have

$$\mathbb{P}\left(\rho^{F^*}(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) \leqslant \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_{p,N}(\eta)/L_{\mathcal{D}})} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X})\right) \geqslant 1-\eta, \quad \forall \boldsymbol{\beta}\in\mathcal{D}.$$

In particular, letting $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_N$ in the above equation yields

$$\mathbb{P}(J_{oos} \leqslant \widehat{J}_N) \geqslant 1-\eta. \tag{13}$$

One can see that the above confidence bounds are essentially dimension-independent, and for any $N \geqslant \log(c_1\eta^{-1})/c_2$, the radius decreases in the order of $1/\sqrt{N}$. They provide also out-of-sample guarantee for the solution of the DRO model (10), i.e. (13), for any choice of the type-p Wasserstein ball and measure ρ.

The approach we take to obtain the above general dimension-independent bounds consists of two steps. First, we reduce the Wasserstein ball defined over $(Y,\mathbf{X}) \sim F_{(Y,\mathbf{X})}$, with dimensionality $n+1$, to a Wasserstein ball defined over $Y\cdot\boldsymbol{\beta}^\top\mathbf{X} \sim F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}}$, with dimensionality 1. Then, we derive confidence bounds by applying a measure concentration property of the Wasserstein metric to the one-dimensional Wasserstein ball. Our first step is formally summarized as follows. To state it, we

[1]Theorem 2 of Fournier and Guillin (2015) (in the one-dimension case) shows that for a distribution F on $\mathbb{R}$ satisfying $A := \mathbb{E}^F[\exp(\gamma\|X\|^a)] < +\infty$, $a > p$, it holds that

$$\mathbb{P}\left(W_p\left(\widehat{F}_N, F\right) \geqslant \varepsilon\right) \leqslant \begin{cases} c_1\exp(-c_2 N\varepsilon^2) & \text{if } \varepsilon \leqslant 1, \\ c_1\exp(-c_2 N\varepsilon^{a/p}) & \text{if } \varepsilon > 1, \end{cases}$$

where $\widehat{F}_N$ is the empirical distribution based on the independent sample drawn from F, and c_1 and c_2 are constants depending on γ, a, A, p; see more details in Theorem 2 of Fournier and Guillin (2015). The constants c_1 and c_2 in (12) are exactly the same as those in Theorem 2 of Fournier and Guillin (2015) in one-dimension case with $\gamma = U_{\mathcal{D}}^{-a}$. See the more details in Appendix.

introduce the Wasserstein ball on $\mathbb{R}^n$ with the metric $d(\cdot,\cdot)$ being a norm:

$$\mathcal{B}_p(F_0,\varepsilon)=\{F\in\mathcal{M}_p(\mathbb{R}^n):W_p(F,F_0)\leqslant\varepsilon\}, \tag{14}$$

where

$$W_p\left(F_1,F_2\right):=\left(\inf_{\pi\in\Pi(F_1,F_2)}\mathbb{E}^{\pi}[\|\xi_1-\xi_2\|^p]\right)^{1/p}, \tag{15}$$

and $\|\cdot\|$ is the norm of (4) on $\mathbb{R}^n$ with its dual norm $\|\cdot\|_*$. Without loss of generality, for $n=1$, i.e., on $\mathbb{R}$, assume that $\|\cdot\|=|\cdot|$ is the absolute-value norm.

Theorem 2. *For $p\in[1,+\infty]$, $\varepsilon\geqslant 0$, $\boldsymbol{\beta}\in\mathbb{R}^n$, a distribution F_0 on $\{-1,1\}\times\mathbb{R}^n$, and $(Y_0,\mathbf{X}_0)\sim F_0$, we have*

$$\{F_{Y\cdot\mathbf{X}}:F_{(Y,\mathbf{X})}\in\overline{\mathcal{B}}_p(F_0,\varepsilon)\}=\mathcal{B}_p(F_{Y_0\cdot\mathbf{X_0}},\varepsilon) \tag{16}$$

and

$$\{F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}}:F_{(Y,\mathbf{X})}\in\overline{\mathcal{B}}_p(F_0,\varepsilon)\}=\mathcal{B}_p\left(F_{Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X_0}},\varepsilon\|\boldsymbol{\beta}\|_*\right). \tag{17}$$

Note that the above theorem will be invoked also in Section 4 for studying the Wasserstein DRO model (6) from a regularization perspective. By Theorem 2, we know that

$$\{F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}}:F_{(Y,\mathbf{X})}\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon)\}=\mathcal{B}_p(\widehat{F}_{N,\boldsymbol{\beta}},\varepsilon\|\boldsymbol{\beta}\|_*),$$

where $\widehat{F}_{N,\boldsymbol{\beta}}=\frac{1}{N}\sum_{i=1}^N\delta_{\widehat{y}_i\cdot\boldsymbol{\beta}^\top\widehat{\mathbf{x}}_i}$. If the one-dimensional Wasserstein ball $\mathcal{B}_p(\widehat{F}_{N,\boldsymbol{\beta}},\varepsilon\|\boldsymbol{\beta}\|_*)$ can contain the distribution $F^*_{\boldsymbol{\beta}}:=F_{Y^*\cdot\boldsymbol{\beta}^\top\mathbf{X}^*}$, i.e. the distribution of a random variable projected from $(Y^*,\mathbf{X}^*)\sim F^*$, with high probability, this will then allow us to derive the confidence bounds in Theorem 1. Our second step achieves this by applying the following measure concentration property of one-dimensional Wasserstein metric, built upon Theorem 2 of Fournier and Guillin (2015).

Lemma 1. *If Assumptions 1 and 2 hold, then for any $\boldsymbol{\beta}\in\mathcal{D}$, $\eta\in(0,1)$ and $N\geqslant 1$, we have*

$$\mathbb{P}\left(W_p\left(\widehat{F}_{N,\boldsymbol{\beta}},F^*_{\boldsymbol{\beta}}\right)\leqslant\varepsilon_{p,N}(\eta)\right)\geqslant 1-\eta,$$

where $\widehat{F}_{N,\boldsymbol{\beta}}=\frac{1}{N}\sum_{i=1}^N\delta_{\widehat{y}_i\cdot\boldsymbol{\beta}^\top\widehat{\mathbf{x}}_i}$, $F^*_{\boldsymbol{\beta}}:=F_{Y^*\cdot\boldsymbol{\beta}^\top\mathbf{X}^*}$ with $(Y^*,\mathbf{X}^*)\sim F^*$, and $\varepsilon_{p,N}$ is defined in (12).

Combining the two steps, we obtain the confidence bounds in Theorem 1.

Proof of Theorem 1 Denote by $\varepsilon_N := \varepsilon_{p,N}(\eta)/L_{\mathcal{D}}$. The result follows by noting that

$$\begin{aligned}
\mathbb{P}\left(\rho^{F^*}(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) \leqslant \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N)} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X})\right) &= \mathbb{P}\left(\rho^{F^*_{\boldsymbol{\beta}}}(Z) \leqslant \sup_{F\in\{F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}}:F_{(Y,\mathbf{X})}\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N)\}} \rho^F(Z)\right)\\
&\geqslant \mathbb{P}\left(F^*_{\boldsymbol{\beta}} \in \{F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}} : F_{(Y,\mathbf{X})} \in \overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N)\}\right)\\
&= \mathbb{P}\left(F^*_{\boldsymbol{\beta}} \in \mathcal{B}_p\left(\widehat{F}_{N,\boldsymbol{\beta}}, \|\boldsymbol{\beta}\|_*\varepsilon_{p,N}(\eta)/L_{\mathcal{D}}\right)\right)\\
&= \mathbb{P}\left(W_p\left(\widehat{F}_{N,\boldsymbol{\beta}}, F^*_{\boldsymbol{\beta}}\right) \leqslant \|\boldsymbol{\beta}\|_*\varepsilon_{p,N}(\eta)/L_{\mathcal{D}}\right)\\
&\geqslant \mathbb{P}\left(W_p\left(\widehat{F}_{N,\boldsymbol{\beta}}, F^*_{\boldsymbol{\beta}}\right) \leqslant \varepsilon_{p,N}(\eta)\right)\\
&\geqslant 1-\eta,
\end{aligned}$$

where the second equality and the last inequality follow from Theorem 2 and Lemma 1, respectively. □

It should be clear from above that one can also obtain naive confidence bounds by applying measure concentration results (Fournier and Guillin (2015)) directly to the probability $\mathbb{P}(F^* \in \overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N))$. This naive approach, unfortunately, suffers from the curse of dimensionality, with the radius scaled in the order of $O(N^{-\min\{p/n,1/2\}})$ (see e.g., Esfahani and Kuhn (2018)).

Remark 1. Suppose that the metric $d(\cdot,\cdot)$ on $\{-1,1\}\times\mathbb{R}^n$ is defined by

$$d((Y_1,\mathbf{X}_1),(Y_2,\mathbf{X}_2)) := \|\mathbf{X}_1-\mathbf{X}_2\| + \theta|Y_1-Y_2| \tag{18}$$

for some $\theta\in(0,\infty]$, and we use the convention that $0\cdot\infty = 0$. Obviously, the metric in (4) is a special case of (18) with $\theta=\infty$. Note that the size of $\overline{\mathcal{B}}_p(F_0,\varepsilon)$ is smaller when θ is larger. Therefore, the result of finite sample guarantee in Theorem 1 is also valid if one chooses to apply the metric (18) with $\theta\in(0,\infty)$.

3.2 Union bounds

With Theorem 1, we are now ready to move on to building generalization bounds. The confidence bounds in Theorem 1 apply to any fixed decision $\boldsymbol{\beta}\in\mathcal{D}$. Generalization bounds require further establishing that the in-sample risk can bound the out-of-sample risk uniformly for all $\boldsymbol{\beta}\in\mathcal{D}$ with high probability. In the case where the set $\mathcal{D}$ is finite, one can easily build generalization bounds for any type of Wasserstein ball and measure ρ by applying the union bound to Theorem 1. Clearly, these union bounds remain to be dimension-independent and the radius required to reach any probability level remains in the same order.

The focus of this section is on the case where the set $\mathcal{D}$ is not finite. We build generalization bounds from Theorem 1 by applying the standard covering number argument (see e.g., Gao (2022)). Recall that for $\tau > 0$, an τ-cover of $\mathcal{D}$, denoted by $\mathcal{D}_\tau$, is a subset of $\mathcal{D}$ which satisfies that for each $\boldsymbol{\beta} \in \mathcal{D}$, there exists $\widetilde{\boldsymbol{\beta}} \in \mathcal{D}_\tau$ such that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\mathcal{D}} \leqslant \tau$, where $\|\cdot\|_{\mathcal{D}}$ is a norm on $\mathcal{D}$. The covering number $\mathcal{N}(\tau; \mathcal{D}, \|\cdot\|_{\mathcal{D}})$ of $\mathcal{D}$ with respect to $\|\cdot\|_{\mathcal{D}}$ is defined as the smallest cardinality of an τ-cover of $\mathcal{D}$.

We show that under the following assumption on the measure ρ, one can obtain $1/\sqrt{N}$-rate generalization bounds for any type of Wasserstein ball and measure ρ.

Assumption 4. *Let $p \in [1, +\infty]$ and $\rho : L^p \to \mathbb{R}$. Assume that there exist $M > 0$ and $k \in [1, p]$ such that*

$$|\rho(Z_1) - \rho(Z_2)| \leqslant M \left(\mathbb{E}[|Z_1 - Z_2|^k]\right)^{1/k}, \quad \forall Z_1, Z_2 \in L^p.$$

Theorem 3. *Given that Assumptions 1-4 hold, $\eta \in (0, 1)$, and*

$$\varepsilon_N := \varepsilon_{p,N}\left(\frac{\eta}{\mathcal{N}(1/N; \mathcal{D}, \|\cdot\|_*)}\right) / L_{\mathcal{D}}, \tag{19}$$

with $\varepsilon_{p,N}(\cdot)$ defined in (12), then

$$\mathbb{P}\left(\rho^{F^*}(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}) \leqslant \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N)} \rho^F(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}) + \tau_N, \quad \forall\, \boldsymbol{\beta} \in \mathcal{D}\right) \geqslant 1 - \eta - \frac{1}{N},$$

where $\tau_N = M[2\left(\mathbb{E}^{F^*}[\|\mathbf{X}\|^k]\right)^{1/k} + \left(\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)\right)^{1/2k} + \varepsilon_N]/N$.

It is known that the covering number in (19) satisfies $\log(\mathcal{N}(\tau; \mathcal{D}, \|\cdot\|_*)) \leqslant n \log(1 + 2B/\tau)$, where B is the diameter of $\mathcal{D}$, i.e. $B = \sup_{\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathcal{D}} \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_*$ (see Example 5.8 of Wainwright (2019)). Moreover, we have $B \leqslant 2U_{\mathcal{D}}$, which follows from Assumption 2. With these facts, Theorem 3 implies that the radius ε_N can be chosen in the order of $\sqrt{(\log N)/N}$ such that the in-sample risk can be an upper bound of the out-of-sample risk, up to a residual in the order of $1/N$, for all $\boldsymbol{\beta} \in \mathcal{D}$. That is, $1/\sqrt{N}$-rate generalization bounds exist. Assumption 4 holds for a large class of measures ρ but does not hold for expected functions $\rho(Z) = \mathbb{E}[\ell(Z)]$ when the loss function ℓ is not Lipschitz continuous. We show below that if only expected functions are considered, one can still obtain $1/\sqrt{N}$-rate generalization bounds even if the loss function is not Lipschitz continuous, by relaxing Assumption 4 to the following.

Assumption 5. *Let $p \in [1, +\infty]$ and $\rho(Z) = \mathbb{E}[\ell(Z)]$, $Z \in L^p$ for some $\ell : \mathbb{R} \to \mathbb{R}$. Assume*

that there exists a function $f : \mathbb{R}^n \to \mathbb{R}_+$ *and constant* $a_1, a_2 \geqslant 0$ *and* $k \in [1, p]$ *satisfying* $f(\mathbf{x}) \leqslant a_1\|\mathbf{x}\|^k + a_2$ *for all* $\mathbf{x} \in \mathbb{R}^n$*, such that*

$$\left|\ell(\boldsymbol{\beta}^\top \mathbf{x}) - \ell(\widetilde{\boldsymbol{\beta}}^\top \mathbf{x})\right| \leqslant f(\mathbf{x})\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_*, \quad \forall \boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathcal{D}.$$

Theorem 4. *Given that Assumptions 1-3 and 5 hold and* $\eta \in (0, 1)$*, then*

$$\mathbb{P}\left(\mathbb{E}^{F^*}[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] \leqslant \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N)} \mathbb{E}^F[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] + \tau_N, \ \forall\ \boldsymbol{\beta} \in \mathcal{D}\right) \geqslant 1 - \eta - \frac{1}{N},$$

where $\tau_N = [a_1(2^{k-1}+1)\mathbb{E}^{F^*}[\|\mathbf{X}\|^k] + a_1 2^{k-1}(\sqrt{\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)} + \varepsilon_N^k) + 2a_2]/N$ *and* ε_N *is defined by* (19)*.*

By the same reasoning as for Theorem 3, we can conclude from above that $1/\sqrt{N}$-rate generalization bounds exist.

4 A Regularization Perspective

Regularization generally refers to any means that can be applied to avoid overfitting and enhance generalization. In this regard, with the generalization guarantees established in the previous section, the Wasserstein DRO model (6) can be particularly well justified as a general regularization model. In this section, we seek to provide further insights as to the regularization effect of the model (6) on the decision variable $\boldsymbol{\beta}$, which offers an alternative, and practically useful, interpretation to the model. In particular, it is known that in some settings Wasserstein DRO is equivalent to a regularized empirical optimization model in ML, with a norm regularizer capturing the regularization effect on the decision variable $\boldsymbol{\beta}$. We show that a similar equivalence relationship can be established more generally for many instances of the model (6), such as those discussed in Section 2.2. In addition to explaining the regularization effect, the equivalence relationship reveals also how these instances, previously unknown for their tractability, can be solved efficiently via regularized empirical optimization models. Our results are general in that not only do they unify all the previously known equivalence relationships, they also provide necessary conditions under which such an equivalence relationship can exist. Theorem 2 greatly facilitates our analysis. Note first that by (16) in the theorem, the Wasserstein DRO model (6) can be recast as

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0, \varepsilon)} \rho^F(\boldsymbol{\beta}^\top \mathbf{Z}). \tag{20}$$

In the case where ρ^F is an expected function, the Wasserstein distributionally robust optimization model (20), i.e.

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \mathbb{E}^F[\ell(\boldsymbol{\beta}^\top \mathbf{Z})], \tag{21}$$

is known to be equivalent to a regularized model of the form

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \left\{ \mathbb{E}^{F_0}[\ell(\boldsymbol{\beta}^\top \mathbf{Z})] + \mathrm{Lip}(\ell)\varepsilon\|\boldsymbol{\beta}\|_* \right\}, \tag{22}$$

when $p = 1$, i.e. the type-1 Wasserstein ball, and $\mathrm{Lip}(\ell)$ is the Lipschitz constant of ℓ.

In the case where the Wasserstein ball is of a higher order, i.e. $p > 1$, and/or the loss function ℓ is not Lipschitz continuous, the relationship between the model (21) and a regularized model is largely unknown except the special case of $p = 2$ and $\ell(y) = y^2$ (Blanchet et al. (2019)). It remains largely open also whether (21) in this case can be tractably solved. Higher-order Wasserstein balls can be attractive from a practical standpoint, given that they are less conservative than the type-1 Wasserstein ball. It is natural to wonder whether there exists an equivalence relationship between the model (21) and a regularized model of the form (22) in the case of a higher-order Wasserstein ball, i.e. $p > 1$. We show below that such an equivalence relationship exists if and only if the loss function ℓ is linear or takes the form of an absolute function.

Theorem 5. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a convex function. For $p \in (1, \infty]$, suppose that $\mathbb{E}[|\ell(Z)|] < +\infty$ for all $Z \in L^p$. Then the following statements are equivalent.*

(i) For any $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$, $\varepsilon \geqslant 0$ and $\mathcal{D} \subseteq \mathbb{R}^n$, we have

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \mathbb{E}^F[\ell(\boldsymbol{\beta}^\top \mathbf{Z})] = \inf_{\boldsymbol{\beta}\in\mathcal{D}} \left\{ \mathbb{E}^{F_0}[\ell(\boldsymbol{\beta}^\top \mathbf{Z})] + C\varepsilon\|\boldsymbol{\beta}\|_* \right\}$$

with some $C > 0$.

(ii) ℓ has one of the following two forms:

-$\ell_1(x) = Cx + b$ or $\ell_1(x) = -Cx + b$ with some $b \in \mathbb{R}$;

-$\ell_2(x) = C|x - m| + b$ with some $m, b \in \mathbb{R}$.

The above result, which applies to any type-p Wasserstein ball, is perhaps surprising in that the Wasserstein DRO model is equivalent to the same regularized model, regardless of the order p. This turns out to be the case when the slope of the loss function ℓ takes values only from a constant and its negative. The result indicates also that it would not be possible to obtain a regularized

model in the form of (22) for any other loss function ℓ. In other words, if there does exist a certain equivalence relationship between the Wasserstein DRO model (21) and a regularized model for some other loss function ℓ, the regularized model must take some form other than (22). Before moving on to discussing other forms of regularization, it is worth noting first the following application of the above result in regression.

Example 1. (Regression)

- (Least absolute deviation (LAD) regression) Applying $\ell_2(x) = C|x-m|+b$ in Theorem 5 and setting $m = 0$, $b = 0$, and $C = 1$, we arrive at the distributionally robust counterpart of the least absolute deviation regression, i.e.

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0,\varepsilon)} \mathbb{E}^F[|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}|]. \tag{23}$$

It is equivalent to

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \left\{ \mathbb{E}^{F_0}[|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}|] + \varepsilon \|(1, -\boldsymbol{\beta}_r)\|_* \right\},$$

for any $p \geqslant 1$.

We now turn our attention to exploring whether there exists any other form of regularization equivalent to the Wasserstein DRO model (21). In particular, we seek to address the case where the loss function ℓ is not Lipschitz continuous, such as those loss functions of a high order arising from the applications discussed in Section 2.2. It is known that the worst-case expectation problem defined based on the type-1 Wasserstein ball can be unbounded if the loss function is not Lipschitz continuous. To study the case that goes beyond Lipschitz continuous functions, we consider the following formulation of the Wasserstein DRO model (21)

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0,\varepsilon)} \mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top \mathbf{Z})], \tag{24}$$

where ℓ^p denotes a Lipschitz continuous function ℓ raised to the power of $p > 1$, with the same order as that of the type-p Wasserstein ball.

As the main result, we show below that for several loss functions ℓ of practical interest, the Wasserstein DRO model (24) is equivalent to an alternative form of regularization.

Theorem 6. *Suppose that ℓ has one of the following four forms:*

-$\ell_1(x) = (x-m)_+$ with some $m \in \mathbb{R}$;

-$\ell_2(x) = (x-m)_-$ *with some* $m \in \mathbb{R}$;

-$\ell_3(x) = (|x-m_1| - m_2)_+$ *with some* $m_1 \in \mathbb{R}$ *and* $m_2 \geqslant 0$;

-$\ell_4(x) = |x-m| + b$ *with some* $m \in \mathbb{R}$ *and* $b > 0$.

Then for any $p \in (1,\infty)$, $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$, $\varepsilon \geqslant 0$ *and* $\mathcal{D} \subseteq \mathbb{R}^n$, *we have*

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top \mathbf{Z})] = \inf_{\boldsymbol{\beta}\in\mathcal{D}} \left(\left(\mathbb{E}^{F_0}[\ell^p(\boldsymbol{\beta}^\top \mathbf{Z})] \right)^{1/p} + \varepsilon ||\boldsymbol{\beta}||_* \right)^p.$$

The above result covers, quite remarkably, all the instances of expected functions in Section 2.2, except Sum-Exp in classification.

Example 2. (Classification)

- (Higher-order hinge loss) Applying $\ell_2(x) = (x-m)_-$ and setting $m = 1$, we have that the following classification problem with a higher-order hinge loss

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\overline{\mathcal{B}}_p(F_0,\varepsilon)} \mathbb{E}^F[(1 - Y\cdot\boldsymbol{\beta}^\top \mathbf{X})_+^p]$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \left(\left(\mathbb{E}^{F_0}[(1 - Y\cdot\boldsymbol{\beta}^\top \mathbf{X})_+^p] \right)^{1/p} + \varepsilon ||\boldsymbol{\beta}||_* \right)^p.$$

- (Higher-order SVM) Applying $\ell_3(x) = (|x-m_1| - m_2)_+$ and setting $m_1 = 1$ and $m_2 = 0$, we have that the higher-order SVM classification problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\overline{\mathcal{B}}_p(F_0,\varepsilon)} \mathbb{E}^F[|1 - Y\cdot\boldsymbol{\beta}^\top \mathbf{X}|^p]$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \left((\mathbb{E}^{F_0}[|1 - Y\cdot\boldsymbol{\beta}^\top \mathbf{X}|^p])^{1/p} + \varepsilon ||\boldsymbol{\beta}||_* \right)^p.$$

Example 3. (Regression)

- (Higher-order error measure) Applying $\ell_3(x) = (|x-m_1| - m_2)_+$ and setting $m_1 = 0$ and $m_2 = 0$, we have that the regression with a higher order error measure

$$\inf_{\boldsymbol{\beta}_r\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \mathbb{E}^F[|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}|^p],$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \left(\left(\mathbb{E}^{F_0}[|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}|^p] \right)^{1/p} + \varepsilon \|(1, -\boldsymbol{\beta}_r)\|_* \right)^p.$$

- (Higher order c-insensitive measure) Applying $\ell_3(x) = (|x - m_1| - m_2)_+$ and setting $m_1 = 0$ and $m_2 = c$, we have that the following regression problem with a higher order c-insensitive measure

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0, \varepsilon)} \mathbb{E}^F[(|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}| - c)_+^p]$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \left(\left(\mathbb{E}^{F_0}[(|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}| - c)_+^p] \right)^{1/p} + \varepsilon ||(1, -\boldsymbol{\beta}_r)||_* \right)^p.$$

Example 4. (Risk minimization)

- (Lower partial moments) Applying $\ell_1(x) = (x - m)_+$ and setting $m = c$, we have that the risk minimization with lower partial moments

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0, \varepsilon)} \mathbb{E}^F[(\boldsymbol{\beta}^\top \mathbf{X} - c)_+^p]$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \left(\left(\mathbb{E}^{F_0}[(\boldsymbol{\beta}^\top \mathbf{X} - c)_+^p] \right)^{1/p} + \varepsilon ||\boldsymbol{\beta}||_* \right)^p.$$

With these many examples covered by the theorem, it is perhaps tempting to conjecture that the equivalence relationship may hold more broadly for other loss functions ℓ. We provide a negative answer below.

Theorem 7. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a nonnegative, Lipschitz continuous and convex function. For an integer $p \in (1, \infty)$, suppose that for any $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$, $\varepsilon \geqslant 0$ and $\mathcal{D} \subseteq \mathbb{R}^n$, we have*

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0, \varepsilon)} \mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top \mathbf{Z})] = \inf_{\boldsymbol{\beta} \in \mathcal{D}} \left(\left(\mathbb{E}^{F_0}[\ell^p(\boldsymbol{\beta}^\top \mathbf{Z})] \right)^{1/p} + C\varepsilon ||\boldsymbol{\beta}||_* \right)^p \tag{25}$$

with some $C > 0$. Then ℓ must be one of the four forms in Theorem 6 multiplying a constant C.

This is an "impossibility" theorem, which puts to rest any effort attempting to draw the equivalence relation for other convex Lipschitz continuous functions $\ell : \mathbb{R} \to \mathbb{R}_{\geqslant 0}$. This theorem should

be of fundamental importance to the study of Wasserstein DRO, given the continuous interests and efforts in motivating Wasserstein DRO from a classical regularization perspective. It shows *exactly* how far one can take this perspective.

Even though Theorem 7 points out the impossibility to draw the equivalence relation (25) for a more general loss function ℓ, Theorem 6 can in fact be applied more broadly, as a powerful basis, to derive alternative equivalence relations for a richer family of measure ρ. In particular, there is a large family of measures ρ that can be expressed generally by the following two forms

$$\rho^F(Z) = \inf_{t\in\mathbb{R}} \left\{ t + \left(\mathbb{E}^F[\ell^p(Z,t)]\right)^{1/p} \right\} \quad \text{and} \quad \mathcal{V}^F(Z) = \inf_{t\in\mathbb{R}} \mathbb{E}^F[\ell^p(Z,t)] \tag{26}$$

for some loss functions ℓ, such as the measure applied in ν-support vector regression, higher moment coherent risk measures in Section 2.2, and variance.

We show in the appendix (see Lemma 6) that for a wide range of loss functions ℓ, the following switching of sup and inf is valid

$$\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \pi_{i,\ell}(F,t) = \inf_{t\in\mathbb{R}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \pi_{i,\ell}(F,t), \quad i=1,2,$$

where

$$\pi_{1,\ell}(F,t) = t + \left(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t)]\right)^{1/p} \quad \text{and} \quad \pi_{2,\ell}(F,t) = \mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t)].$$

This, combined with Theorem 6, leads to the following.

Corollary 1. *For any $p\in[1,\infty)$ and $c>1$, let ρ^F be defined by (26), where $\ell(z,t) := c\ell(z-t)$, and ℓ is one of ℓ_1,ℓ_3,ℓ_4 in Theorem 6 or $\ell(z,t)=c(|z|-t)_+$. It holds that for $F_0\in\mathcal{M}_p(\mathbb{R}^n)$ and $\varepsilon\geqslant 0$*

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho^F(\boldsymbol{\beta}^\top\mathbf{Z}) = \inf_{\boldsymbol{\beta}\in\mathcal{D}} \left\{ \rho^{F_0}(\boldsymbol{\beta}^\top\mathbf{Z}) + c\varepsilon||\boldsymbol{\beta}||_* \right\}.$$

Example 5. (Classification) Setting $\ell(z,t)=(z-t)_+$ in Corollary 1 and $p=1$, we have that the following ν-support vector machine problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\overline{\mathcal{B}}_1(F_0,\varepsilon)} \mathrm{CVaR}_\alpha^F(-Y\cdot\boldsymbol{\beta}^\top\mathbf{X})$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \left\{ \mathrm{CVaR}_\alpha^{F_0}(-Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) + \frac{1}{1-\alpha}\varepsilon||\boldsymbol{\beta}||_* \right\}.$$

Example 6. (Regression) Setting $\ell(z,t) = (|z| - t)_+$ in Corollary 1 and $p = 1$, we have that the following ν-support vector regression problem

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \sup_{F \in \mathcal{B}_1(F_0,\varepsilon)} \mathrm{CVaR}_\alpha^F(|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}|)$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}_r \in \mathcal{D}} \left\{ \mathrm{CVaR}_\alpha^{F_0}(|(1, -\boldsymbol{\beta}_r)^\top \mathbf{X}|) + \frac{1}{1-\alpha} \varepsilon ||(1, -\boldsymbol{\beta}_r)||_* \right\}.$$

Example 7. (Risk minimization) Setting $\ell(z,t) = \ell_1(z-t) = (z-t)_+$ in Corollary 1, we have that the following problem of minimizing higher moment risk measures

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0,\varepsilon)} \inf_{t \in \mathbb{R}} \left\{ t + c \left(\mathbb{E}^F[(\boldsymbol{\beta}^\top \mathbf{X} - t)_+^p] \right)^{1/p} \right\}$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}, t \in \mathbb{R}} \left\{ t + c \left(\mathbb{E}^{F_0}[(\boldsymbol{\beta}^\top \mathbf{X} - t)_+^p] \right)^{1/p} + c\varepsilon ||\boldsymbol{\beta}||_* \right\}.$$

Corollary 2. *For any $p \in [1, \infty)$, let $\mathcal{V}^F$ be defined by (26), where $\ell(z,t) := c\ell(z-t)$ and ℓ is ℓ_3 or ℓ_4 in Theorem 6. It holds that for $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$ and $\varepsilon \geqslant 0$*

$$\inf_{\boldsymbol{\beta} \in \mathcal{D}} \sup_{F \in \mathcal{B}_p(F_0,\varepsilon)} \mathcal{V}^F(\boldsymbol{\beta}^\top \mathbf{Z}) = \inf_{\boldsymbol{\beta} \in \mathcal{D}} \left(\left(\mathcal{V}^{F_0}(\boldsymbol{\beta}^\top \mathbf{Z}) \right)^{1/p} + \varepsilon ||\boldsymbol{\beta}||_* \right)^p.$$

The above corollary can accommodate, for example, the case where variance is applied as the measure. Variance is not listed as an example in Section 2.2 because it has been studied in Blanchet et al. (2022). Yet, our approach to derive the equivalency relation is different from, and significantly more general than, the approach taken in Blanchet et al. (2022). We give the following example to highlight the generality of our approach.

Example 8. (Blanchet et al. (2022)) When $p = 2$ and $\ell(z,t) = \ell_3(z-t) = |z-t|$ (with $m_1 = 0$ and $m_2 = 0$), $\mathcal{V}^F = \mathbb{V}\mathrm{ar}^F$. That is,

$$\sup_{F \in \mathcal{B}_2(F_0,\varepsilon)} \mathbb{V}\mathrm{ar}^F(\boldsymbol{\beta}^\top \mathbf{Z}) = \sup_{F \in \mathcal{B}_2(F_0,\varepsilon)} \inf_{t \in \mathbb{R}} \mathbb{E}^F[(\boldsymbol{\beta}^\top \mathbf{Z} - t)^2].$$

Applying Corollary 2 yields

$$\sup_{F\in\mathcal{B}_2(F_0,\varepsilon)} \mathbb{V}\mathrm{ar}^F(\boldsymbol{\beta}^\top\mathbf{Z}) = \left(\sqrt{\mathbb{V}\mathrm{ar}^{F_0}(\boldsymbol{\beta}^\top\mathbf{Z})} + \varepsilon\|\boldsymbol{\beta}\|_*\right)^2.$$

The case of exponential functions

Sum-Exp for classification in Section 2.2, i.e. $\rho(Z) := \frac{1}{t}\mathbb{E}[e^{-tZ}]$, is a special case that cannot be covered by the theorems above. We proceed by assuming that the Wasserstein ball in (20) is of type-∞, since otherwise the problem may be unbounded. We can make the following observation and identify the regularization counterpart of the Sum-Exp classification problems.

Proposition 1. *For a monotonic function $\rho : L^\infty \to \mathbb{R}$, that is, $\rho(Z_1) \leqslant \rho(Z_2)$ whenever $Z_1 \leqslant Z_2$ a.s., we have*

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_\infty(F_0,\varepsilon)} \rho^F(\boldsymbol{\beta}^\top\mathbf{Z}) = \inf_{\boldsymbol{\beta}\in\mathcal{D}} \rho^{F_0}(\boldsymbol{\beta}^\top\mathbf{Z} + \varepsilon\|\boldsymbol{\beta}\|_*).$$

Example 9. (Classification)

- (Sum-Exp) Applying $\rho(Z) = \frac{1}{t}\mathbb{E}[e^{tZ}]$, we have that the Sum-Exp classification problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\overline{\mathcal{B}}_\infty(F_0,\varepsilon)} \frac{1}{t}\mathbb{E}^F[e^{-tY\cdot\boldsymbol{\beta}^\top\mathbf{X}}]$$

is equivalent to the regularization problem

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \frac{1}{t}\mathbb{E}^{F_0}\left[e^{-t(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}-\varepsilon\|\boldsymbol{\beta}\|_*)}\right].$$

The case of distortion functionals

ν-support vector machine (regression) and CVaR-deviation in Section 2.2 are two examples of how quantile-based risk measures such as CVaR play the role of building blocks for other measures. It is known in the literature of risk measures that CVaR is a special case of distortion functional, defined by

$$\rho_h^F(Z) = \int_0^1 F^{-1}(s)\mathrm{d}h(s),$$

where $h : [0,1] \to \mathbb{R}$ is called a distortion function and F^{-1} is left-quantile function of F, i.e. $F^{-1}(s) = \inf\{x : F(x) \geqslant s\}$ for $s \in (0,1]$, and $F^{-1}(0) = \inf\{x : F(x) > 0\}$. Distortion functionals allow for taking into account a spectrum of quantiles with respect to different probability levels.

In particular, when the distortion function h is a convex function, the distortion functional ρ_h^F is a mixture of CVaRs with the mixture weights determined by the derivative of h (see e.g. Rockafellar and Uryasev (2013)). The following generalization of CVaR-deviation appears in Rockafellar et al. (2008)

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho_h^F(\boldsymbol{\beta}^\top \mathbf{X} - \mathbb{E}^F[\boldsymbol{\beta}^\top \mathbf{X}]), \tag{27}$$

where h is convex. Similarly, one can consider the following generalization of ν-support vector regression

$$\inf_{\boldsymbol{\beta}_r\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho_h^F(|(1,-\boldsymbol{\beta}_r)^\top \mathbf{X}|), \tag{28}$$

where h is increasing and convex. We show below that for both cases, there exists an exact equivalence relation between the Wasserstein DRO model and regularization.

Proposition 2. *(i) Let $h:[0,1]\to\mathbb{R}$ be a convex function. For $p\in[1,\infty]$, the problem (27) is equivalent to*

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \left\{ \rho_h^{F_0}(\boldsymbol{\beta}^\top \mathbf{X} - \mathbb{E}^{F_0}[\boldsymbol{\beta}^\top \mathbf{X}]) + \|h'_- + h(0) - h(1)\|_q \varepsilon ||\boldsymbol{\beta}||_* \right\}, \tag{29}$$

where $\|g\|_q = (\int_0^1 |g(s)|^q \mathrm{d}s)^{1/q}$, and h'_- is the left derivative of h.

(ii) Let $h:[0,1]\to\mathbb{R}$ be an increasing and convex distortion function. For $p\in[1,\infty]$, the problem (28) is equivalent to

$$\inf_{\boldsymbol{\beta}_r\in\mathcal{D}} \left\{ \rho_h^{F_0}(|(1,-\boldsymbol{\beta}_r)^\top \mathbf{X}|) + \|h'_-\|_q \varepsilon ||(1,-\boldsymbol{\beta}_r)||_* \right\}. \tag{30}$$

One can easily apply the above proposition with $h(s)=(s-\alpha)_+/(1-\alpha)$, where $(s)_+=\max\{s,0\}$, to derive the regularization counterpart for the example of CVaR-deviation in risk minimization (in Section 2.2).

5 Conclusion

In this paper, we first show that Wasserstein DRO provides generalization guarantees for virtually any problem with affine decision rules - namely, for any type-p Wasserstein ball and (almost) any measure of risk. This justifies the use and reveals the strength of Wasserstein DRO beyond the classical expectation optimization setting. Our approach sheds light on why generalization guarantees can be obtained (almost) independently of the choice of a Wasserstein ball and the measure of risk. Although our approach is limited to problems with affine decision rules, such problems are very

common in practical ML and OR applications given the popularity of affine decision rules. Such problems also constitute an important basis for studying more sophisticated forms of problems. We further show how to draw exact equivalency relations between Wasserstein DRO and the classical regularization scheme for a wide range of regression, classification, and risk minimization problems that have not been considered to date using Wasserstein DRO. Our result broadens considerably the class of Wasserstein DRO problems that can be solved exactly via their regularization formulations. We also prove that our equivalency results are, in some sense, the most general that one can possibly obtain, in the case where the measure is an expected function (Theorems 5 and 7). This provides valuable insights into how far the classical regularization interpretation can be applied to Wasserstein DRO. The generalization and regularization results presented in this paper should be of high interest to ML and OR researchers/practitioners who seek to gain confidence in Wasserstein DRO for a broader set of applications.

References

Bartl, D., Drapeau, S., Obloj, J., and Wiesel, J. (2020). Robust uncertainty sensitivity analysis. *arXiv preprint arXiv:2006.12022*

Bawa, V. S. (1975) Optimal rules for ordering uncertain prospects. *Journal of Financial Economics* **2**(1) 95–1.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B. and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, **59**(2), 341–357.

Blanchet, J., Chen, L. and Zhou, X. (2022). Distributionally robust mean-variance portfolio selection with Wasserstein distances. *Management Science*, forthcoming.

Blanchet, J. and Kang, Y. (2021). Sample out-of-sample inference based on Wasserstein distance. *Operations Research*, **69**(3), 985–1013.

Blanchet, J., Kang, Y. and Murthy, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, **56**(3), 830–857.

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

Blanchet, J. Murthy, K. and Si, N. (2022). Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, **109**(2), 295–315.

Chen, R. and Paschalidis, I. C. (2018). A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization. *Journal of Machine Learning Research*, **19**, 1–48.

Chen, L., He, S., Zhang, S. (2011) Tight Bounds for Some Risk Measures, with Applications to Robust Portfolio Selection. *Operations Research*, **59**(4), 847–865.

Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. and Vyncke, D. (2002). The concept of comonotonicity in

actuarial science and finance: Theory. *Insurance: Mathematics and Economics*. **31**, 3–33.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. (1997) Support vector regression machines. *Advances in Neural Information Processing Systems* **9**, 155–161.

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

Fan, K. (1953). Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, **39**(1), 42.

Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below target returns. *American Economic Review* **67**(2), 116–126.

Föllmer, H. and Schied, A. (2016). *Stochastic Finance. An Introduction in Discrete Time*. Fourth Edition. Walter de Gruyter, Berlin.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, **162**(3), 707–738.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

Gao, R. (2022). Finite-Sample Guarantees for Wasserstein Distributionally Robust Optimization: Breaking the Curse of Dimensionality. *Operations Research*, forthcoming.

Gao, R., Chen, X. and Kleywegt, A. J. (2017). Distributional robustness and regularization in statistical learning. *arXiv preprint arXiv*:1712.06050.

Gao, R., Chen, X. and Kleywegt, A. J. (2022). Wasserstein Distributionally Robust Optimization and Variation Regularization. *Operations Research*, forthcoming.

Gao, R. and Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv*:1604.02199.

Gotoh, J. and Uryasev, S. (2017). Support vector machines based on convex risk functions and general norms. *Annals of Operations Research*, **249**, 301–328.

Hu, Z. and Hong, L. J. (2018). Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1695–1724.

Jiang, R. and Guan, Y. (2016). Data-driven chance constrained stochastic program. *Mathematical Programming*, **158**(1), 291–327.

Kantorovich, L.V. (1942). On the translocation of masses. *Doklady Akademii Nauk USSR*, 37:199–201.

Krokhmal, P. (2007). Higher moment coherent risk measures. *Quantitative Finance*, **7**, 373–387.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., Shafieezadeh-Abadeh, S. (2019) Wasserstein distributionally robust optimization: theory and applications in machine learning. *INFORMS TutORials in Operations Research*, 130–166.

Lee, Y.-J., and Mangasarian (2001). SSVM: A Smooth Support Vector Machine for Classification. *Computational Optimization and Applications* **20**, 5–22.

Lee, Y. -J., Hsieh, W. -F. , Huang, C.-M. , ε-SSVR: A smooth support vector machine for ε-insensitive

regression. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 5–22.

Li, J. Y.-M. (2018) Closed-Form Solutions for Worst-Case Law Invariant Risk Measures with Application to Robust Portfolio Optimization. *Operations Research*, **66**(6),1533–1541.

Mao, T., Wang, R. and Wu, Q. (2022). Model Aggregation for Risk Evaluation and Robust Optimization. *http://arxiv.org/abs/2201.06370*

Postek, K., Den Hertog, D., and Melenberg, B. (2016) Computationally tractable counterparts of distributionally robust constraints on risk measures *SIAM Review*, **58**(4), 603–650.

Qi, M., Cao, Y. and Shen, Z. J. (2022). Distributionally robust conditional quantile prediction with fixed design. *Management Science*, **68**(3), 1639–1658.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. Advances in neural information processing systems, 31, 5334–5344.

Rockafellar, R.T. (1970). *Convex analysis*. Princeton New Jersey: Princeton University Press.

Rockafellar, R.T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, **26**(7), 1443–1471.

Rockafellar, R.T., Uryasev, S., Zabarankin, M. (2008). Risk tuning with generalized linear regression. *Mathematics of Operations Research*, **33**(3), 712–729.

Rockafellar, R.T. and Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, **18**(1-2), 33–53.

Schölkopf, B., Bartlett, P., Smola, A., Williamson, R. (1998) Support vector regression with automatic accuracy control. In ICANN 98; Springer: Heidelberg, Germany, 111–116.

Schölkopf, B., Smola, A. J., Williamson, R. C., Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, **12** (5), 1207–1245.

Shafieezadeh-Abadeh, S., Esfahanim P. M. and Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 1576–1584.

Shafieezadeh-Abadeh, S., Kuhn, D. and Esfahani, P. M. (2019). Regularization via mass transportation. *Journal of Machine Learning Research*, **20**, 1–68.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms* Cambridge University Press, Cambridge.

Suykens, J. and Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Processing Letters 9*, 293–300.

Sim, M., Zhao, L., Zhou, M. (2021). Tractable Robust Supervised Learning Models. https://ssrn.com/abstract=3981205

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics*, **8**(1), 171–176.

Villani, C. (2009). *Optimal Transport: Old and New* (Vol. 338, p. 23). Berlin: springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Volume 48 (Cambridge University Press).

Wang, R., Wei, Y. and Willmot, G. E. (2020). Characterization, robustness and aggregation of signed Choquet

integrals. *Mathematics of Operations Research*, **45**(3), 993–1015.

Wozabal, D. (2014). Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Operations Research*, **62**(6), 1302–1315.

A Proofs of Section 3

Proof of Theorem 2. We first prove (16). Note that

$$\begin{aligned}\{F_{Y\cdot\mathbf{X}} : F_{(Y,\mathbf{X})} \in \overline{\mathcal{B}}_p(F_0,\varepsilon)\} &= \{F_{Y\cdot\mathbf{X}} : \mathbb{E}[d((Y,\mathbf{X}),(Y_0,\mathbf{X}_0))^p] \leqslant \varepsilon^p\} \\ &= \{F_{Y_0\cdot\mathbf{X}} : \mathbb{E}[\|\mathbf{X}-\mathbf{X}_0\|^p] \leqslant \varepsilon^p\} \\ &= \{F_{Y_0\cdot\mathbf{X}} : \mathbb{E}[\|Y_0\cdot\mathbf{X}-Y_0\cdot\mathbf{X}_0\|^p] \leqslant \varepsilon^p\} =: \mathcal{B}_{(1)},\end{aligned}$$

where the second equality follows from the definition in (4) which implies $Y = Y_0$ almost surely for any $(Y,\mathbf{X})$ satisfying $\mathbb{E}[d((Y,\mathbf{X}),(Y_0,\mathbf{X}_0))^p] \leqslant \varepsilon^p$, and the third equality follows from $\|\mathbf{X}-\mathbf{X}_0\| = \|Y_0\cdot\mathbf{X}-Y_0\cdot\mathbf{X}_0\|$ as $|Y_0|=1$. Noting that

$$\mathcal{B}_p(F_{Y_0\cdot\mathbf{X_0}},\varepsilon) = \{F_{\mathbf{Z}} : \mathbb{E}[\|\mathbf{Z}-Y_0\cdot\mathbf{X_0}\|^p] \leqslant \varepsilon^p\} =: \mathcal{B}_{(2)},$$

it holds that $\mathcal{B}_{(1)} \subseteq \mathcal{B}_{(2)}$ obviously. To see the converse inclusion, note that for any $F \in \mathcal{B}_{(2)}$, there exists $\mathbf{Z}$ such that $\mathbf{Z} \sim F$ and $\mathbb{E}[\|\mathbf{Z}-Y_0\cdot\mathbf{X}_0\|^p] \leqslant \varepsilon^p$. Take $(Y,\mathbf{X}) = (Y_0, \mathbf{Z}/Y_0)$. We have $Y = Y_0$ and

$$\mathbb{E}[\|Y_0\mathbf{X}-Y_0\mathbf{X}_0\|^p] = \mathbb{E}\left[\|\mathbf{Z}-Y_0\cdot\mathbf{X}_0\|^p\right] \leqslant \varepsilon^p.$$

Hence, we have $F \in \mathcal{B}_{(1)}$, and thus $\mathcal{B}_{(2)} \subseteq \mathcal{B}_{(1)}$. Therefore, we have $\mathcal{B}_{(1)} = \mathcal{B}_{(2)}$ which implies (16).

With the aid of (16), we can now prove (17). Note that

$$\begin{aligned}\{F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}} : F_{(Y,\mathbf{X})} \in \overline{\mathcal{B}}_p(F_0,\varepsilon)\} &= \left\{F_{Y\cdot\boldsymbol{\beta}^\top\mathbf{X}} : \mathbb{E}[d((Y,\mathbf{X}),(Y_0,\mathbf{X}_0))^p] \leqslant \varepsilon^p\right\} \\ &= \left\{F_{Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X}} : \mathbb{E}[\|\mathbf{X}-\mathbf{X}_0\|^p] \leqslant \varepsilon^p\right\} =: \mathcal{B},\end{aligned}$$

where the second equality follows from the definition in (4). It suffices to show $\mathcal{B} = \mathcal{B}_p(F_{Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X_0}}, \varepsilon\|\boldsymbol{\beta}\|_*)$. The case of $\|\boldsymbol{\beta}\| = 0$ is trivial, and we assume now $\|\boldsymbol{\beta}\| > 0$. On the one hand, for $F \in \mathcal{B}$, there exists $\mathbf{X}$ such that $Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X} \sim F$ and $\mathbb{E}[\|\mathbf{X}-\mathbf{X}_0\|^p] \leqslant \varepsilon^p$. Thus, we have

$$\mathbb{E}[|Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X}-Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X}_0|^p] = \mathbb{E}[|\boldsymbol{\beta}^\top\mathbf{X}-\boldsymbol{\beta}^\top\mathbf{X}_0|^p] \leqslant \|\boldsymbol{\beta}\|_*^p\mathbb{E}[\|\mathbf{X}-\mathbf{X}_0\|^p] \leqslant \varepsilon^p\|\boldsymbol{\beta}\|_*^p,$$

where we use the Hölder inequality in the second step. This implies $F \in \mathcal{B}_p(F_{Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X_0}}, \varepsilon\|\boldsymbol{\beta}\|_*)$, and hence, $\mathcal{B} \subseteq \mathcal{B}_p(F_{Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X_0}}, \varepsilon\|\boldsymbol{\beta}\|_*)$. On the other hand, for $F \in \mathcal{B}_p(F_{Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X_0}}, \varepsilon\|\boldsymbol{\beta}\|_*)$, there exists $Z \sim F$ such that $\mathbb{E}[|Z-Y_0\cdot\boldsymbol{\beta}^\top\mathbf{X_0}|^p] \leqslant \varepsilon^p\|\boldsymbol{\beta}\|_*^p$. It follows from the definition of dual norm that

there exists $\boldsymbol{\beta}_0 \in \mathbb{R}^n$ such that $\|\boldsymbol{\beta}_0\| = 1$ and $\|\boldsymbol{\beta}\|_* = \boldsymbol{\beta}^\top \boldsymbol{\beta}_0$. Define

$$T = Z - Y_0 \cdot \boldsymbol{\beta}^\top \mathbf{X_0} \quad \text{and} \quad \mathbf{X} = \mathbf{X}_0 + \frac{\boldsymbol{\beta}_0 T}{Y_0 \|\boldsymbol{\beta}\|_*}.$$

It holds that $\mathbb{E}[|T|^p] \leqslant \varepsilon^p \|\boldsymbol{\beta}\|_*^p$, and thus,

$$\mathbb{E}[\|\mathbf{X} - \mathbf{X}_0\|^p] = \mathbb{E}\left[\left\|\frac{\boldsymbol{\beta}_0 T}{Y_0 \|\boldsymbol{\beta}\|_*}\right\|^p\right] = \frac{\mathbb{E}[|T|^p]}{\|\boldsymbol{\beta}\|_*^p} \leqslant \varepsilon^p.$$

This implies $F_{Y_0 \cdot \boldsymbol{\beta}^\top \mathbf{X}} \in \mathcal{B}$. Noting that $Z = Y_0 \cdot \boldsymbol{\beta}^\top \mathbf{X}$, we have $F \in \mathcal{B}$. Hence, we conclude $\mathcal{B}_p(F_{Y_0 \cdot \boldsymbol{\beta}^\top \mathbf{X_0}}, \varepsilon \|\boldsymbol{\beta}\|_*) \subseteq \mathcal{B}$. This completes the proof of (17). We note that our proof of (17) is similar in spirit to, and includes as a special case, that of Theorem 7 in Mao et al. (2022). □

Proof of Lemma 1. For $\boldsymbol{\beta} \in \mathcal{D}$, note that the light-tailed condition in Assumption 1 implies

$$\mathbb{E}^{F^*_{\boldsymbol{\beta}}}[\exp(U_{\mathcal{D}}^{-a}|Z|^a)] = \mathbb{E}^{F^*}[\exp(U_{\mathcal{D}}^{-a}|Y \cdot \boldsymbol{\beta}^\top \mathbf{X}|^a)] \leqslant \mathbb{E}^{F^*}[\exp(\|\mathbf{X}\|^a)] = A < +\infty,$$

where the first inequality follows from the Hölder inequality. This implies that $F^*_{\boldsymbol{\beta}}$ is a light-tailed distribution on $\mathbb{R}$ which satisfies the condition of Theorem 2 of Fournier and Guillin (2015). Further, noting that $\widehat{F}_{N,\boldsymbol{\beta}}$ is the empirical distribution based on the the independent samples in the population whose true distribution is $F^*_{\boldsymbol{\beta}}$, by Theorem 2 of Fournier and Guillin (2015), we obtain

$$\mathbb{P}\left(W_p\left(\widehat{F}_{N,\boldsymbol{\beta}}, F^*_{\boldsymbol{\beta}}\right) \geqslant \varepsilon\right) \leqslant \begin{cases} c_1 \exp(-c_2 N \varepsilon^2) & \text{if } \varepsilon \leqslant 1, \\ c_1 \exp(-c_2 N \varepsilon^{a/p}) & \text{if } \varepsilon > 1, \end{cases} \tag{31}$$

where c_1, c_2 are positive constants that only depend on $U_{\mathcal{D}}$, a, A, and p. Substituting $\varepsilon = \widehat{\varepsilon}_{p,N}(\eta)$ defined by (12) yields the desired result. □

Proof of Theorem 3. By Assumption 4, we have for any $(Y, \mathbf{X}) \in \{-1, 1\} \times \mathbb{R}^n$,

$$\left|\rho(Y \cdot \boldsymbol{\beta}^\top \mathbf{X}) - \rho(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})\right| \leqslant M\left(\mathbb{E}[|Y \cdot (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^\top \mathbf{X}|^k]\right)^{1/k} \leqslant M\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_* \left(\mathbb{E}[\|Y\mathbf{X}\|^k]\right)^{1/k}, \tag{32}$$

where the last inequality follows from the Hölder inequality. Denote by $\widehat{\mathbf{X}}_N \sim \widehat{F}_{N,\mathbf{X}}$. It holds that

$$\begin{aligned}
&\left| \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon)} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) - \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon)} \rho^F(Y\cdot\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}) \right| \\
&\leqslant \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon)} \left| \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) - \rho^F(Y\cdot\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}) \right| \leqslant M\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_* \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon)} \left(\mathbb{E}^F[\|Y\mathbf{X}\|^k]\right)^{1/k} \\
&= M\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_* \sup_{F\in\mathcal{B}_p(\widehat{F}_{N,\boldsymbol{X}},\varepsilon)} \left(\mathbb{E}^F[\|\mathbf{Z}\|^k]\right)^{1/k} \leqslant M\|\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}\|_* \left(\left(\mathbb{E}^{\widehat{F}_{N,\boldsymbol{X}}}[\|\mathbf{Z}\|^k]\right)^{1/k} + \varepsilon \right), \qquad (33)
\end{aligned}$$

where the first inequality follows from (32), the equality follows from (16) of Theorem 2, and the last inequality follows from $(\mathbb{E}^F[\|\mathbf{X}\|^k])^{1/k} \leqslant (\mathbb{E}^{F_0}[\|\mathbf{X}\|^k])^{1/k} + \varepsilon$ for any $F \in \mathcal{B}_p(F_0,\varepsilon)$. For any $0 < x \leqslant y$ and $t \geqslant 0$, it holds that $y - x > t^{1/k}$ implies $y^k - x^k > t$. Hence, we have

$$\begin{aligned}
&\mathbb{P}\left(\left(\mathbb{E}^{\widehat{F}_N}[\|\mathbf{X}\|^k]\right)^{1/k} - \left(\mathbb{E}^{F^*}[\|\mathbf{X}\|^k]\right)^{1/k} > \left(\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)\right)^{1/2k} \right) \\
&\leqslant \mathbb{P}\left(\mathbb{E}^{\widehat{F}_N}[\|\mathbf{X}\|^k] - \mathbb{E}^{F^*}[\|\mathbf{X}\|^k] > \sqrt{\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)} \right) \leqslant \frac{1}{N}, \qquad (34)
\end{aligned}$$

where the second inequality follows from the Chebyshevs inequality. Let $\tau > 0$ and $\mathcal{D}_\tau$ be an τ-cover of $\mathcal{D}$ with respect to the norm $\|\cdot\|_*$. Denote by $\varepsilon_N(\eta) := \varepsilon_{p,N}(\eta)/L_{\mathcal{D}}$, and we have

$$\begin{aligned}
&\mathbb{P}\left(\exists \boldsymbol{\beta}\in\mathcal{D} \text{ s.t. } \rho^{F^*}(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) > \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N(\eta))} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) \right. \\
&\qquad \left. + \tau M\left(2\left(\mathbb{E}^{F^*}[\|\mathbf{X}\|^k]\right)^{1/k} + \left(\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)\right)^{1/2k} + \varepsilon_N(\eta) \right) \right) \\
&\leqslant \frac{1}{N} + \mathbb{P}\left(\exists \boldsymbol{\beta}\in\mathcal{D} \text{ s.t. } \rho^{F^*}(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) > \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N(\eta))} \rho^F(Y\cdot\boldsymbol{\beta}^\top\mathbf{X}) \right. \\
&\qquad \left. + \tau M\left(\left(\mathbb{E}^{F^*}[\|\mathbf{X}\|^k]\right)^{1/k} + \left(\mathbb{E}^{\widehat{F}_N}[\|\mathbf{X}\|^k]\right)^{1/k} + \varepsilon_N(\eta) \right) \right) \\
&\leqslant \frac{1}{N} + \mathbb{P}\left(\exists \widetilde{\boldsymbol{\beta}}\in\mathcal{D}_\tau \text{ s.t. } \rho^{F^*}(Y\cdot\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}) > \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N(\eta))} \rho^F(Y\cdot\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}) \right) \\
&\leqslant \frac{1}{N} + \sum_{\widetilde{\boldsymbol{\beta}}\in\mathcal{D}_\tau} \mathbb{P}\left(\rho^{F^*}(Y\cdot\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}) > \sup_{F\in\overline{\mathcal{B}}_p(\widehat{F}_N,\varepsilon_N(\eta))} \rho^F(Y\cdot\widetilde{\boldsymbol{\beta}}^\top\mathbf{X}) \right) \\
&\leqslant \frac{1}{N} + \mathcal{N}(\tau;\mathcal{D},\|\cdot\|_*)\eta,
\end{aligned}$$

where the first inequality follows from (34), the second inequality is due to (32) and (33), and we have used Theorem 1 in the last inequality. Letting $\tau = 1/N$ and replacing η by $\eta/\mathcal{N}(\tau;\mathcal{D},\|\cdot\|_*)$ yield the result. □

Proof of Theorem 4. Applying Assumption 5, we have for any $(Y, \mathbf{X}) \in \{-1, 1\} \times \mathbb{R}^n$ with $F_{\mathbf{X}} \in \mathcal{M}_p(\mathbb{R}^n)$,

$$\left|\mathbb{E}[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] - \mathbb{E}[\ell(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})]\right| \leqslant \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_* \mathbb{E}[f(Y \cdot \mathbf{X})] \leqslant \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_* \left(a_1 \mathbb{E}[\|Y \cdot \mathbf{X}\|^k] + a_2\right). \quad (35)$$

Similar to (33), one can show that

$$\begin{aligned} &\left| \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon)} \mathbb{E}^F[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] - \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon)} \mathbb{E}^F[\ell(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})] \right| \\ &\leqslant \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_* \left(a_1 2^{k-1} \left(\mathbb{E}^{\widehat{F}_N}[\|\mathbf{X}\|^k] + \varepsilon^k\right) + a_2\right). \end{aligned} \quad (36)$$

Moreover, by the Chebyshev's inequality, we have

$$\mathbb{P}\left(\mathbb{E}^{\widehat{F}_N}[\|\mathbf{X}\|^k] - \mathbb{E}^{F^*}[\|\mathbf{X}\|^k] > \sqrt{\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)}\right) \leqslant \frac{1}{N}. \quad (37)$$

Let $\tau > 0$ and $\mathcal{D}_\tau$ be an τ-cover of $\mathcal{D}$ with respect to the norm $\|\cdot\|_*$. Denote by $\varepsilon_N(\eta) := \varepsilon_{p,N}(\eta)/L_{\mathcal{D}}$, and we have

$$\begin{aligned} &\mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathcal{D},\ s.t.\ \mathbb{E}^{F^*}[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] > \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N(\eta))} \mathbb{E}^F[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] \right. \\ &\qquad \left. + \tau\left(a_1(2^{k-1}+1)\mathbb{E}^{F^*}[\|\mathbf{X}\|^k] + a_1 2^{k-1}\left(\sqrt{\mathbb{V}\mathrm{ar}^{F^*}(\|\mathbf{X}\|^k)} + \varepsilon_N(\eta)^k\right) + 2a_2\right)\right) \\ &\leqslant \frac{1}{N} + \mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathcal{D},\ s.t.\ \mathbb{E}^{F^*}[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] > \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N(\eta))} \mathbb{E}^F[\ell(Y \cdot \boldsymbol{\beta}^\top \mathbf{X})] \right. \\ &\qquad \left. + \tau\left(a_1 \mathbb{E}^{F^*}[\|\mathbf{X}\|^k] + a_1 2^{k-1}\left(\mathbb{E}^{\widehat{F}_N}[\|\mathbf{X}\|^k] + \varepsilon_N(\eta)^k\right) + 2a_2\right)\right) \\ &\leqslant \frac{1}{N} + \mathbb{P}\left(\exists \widetilde{\boldsymbol{\beta}} \in \mathcal{D}_\tau,\ s.t.\ \mathbb{E}^{F^*}[\ell(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})] > \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N(\eta))} \mathbb{E}^F[\ell(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})]\right) \\ &\leqslant \frac{1}{N} + \sum_{\widetilde{\boldsymbol{\beta}} \in \mathcal{D}_\tau} \mathbb{P}\left(\mathbb{E}^{F^*}[\ell(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})] > \sup_{F \in \overline{\mathcal{B}}_p(\widehat{F}_N, \varepsilon_N(\eta))} \mathbb{E}^F[\ell(Y \cdot \widetilde{\boldsymbol{\beta}}^\top \mathbf{X})]\right) \\ &\leqslant \frac{1}{N} + \mathcal{N}(\tau; \mathcal{D}, \|\cdot\|_*)\eta, \end{aligned}$$

where the first inequality follows from (37), the second inequality is due to (35) and (36), and we have used Theorem 1 in the last inequality. Letting $\tau = 1/N$ and replacing η by $\eta/\mathcal{N}(\tau; \mathcal{D}, \|\cdot\|_*)$ yield the result. □

B Proofs of Section 4

In this section, we will give the proof of the results in Section 4. First, some notation need to be introduced. For a random variable Z, denote $\|Z\|_p$ by its L_p-norm, i.e. $\|Z\|_p = (\mathbb{E}[|Z|^p])^{1/p}$. We use $\mathbb{1}_A$ to represent the indicator function, i.e., $\mathbb{1}_A(\omega) = 1$ if $\omega \in A$, and $\mathbb{1}_A(\omega) = 0$ otherwise. The sign function on $\mathbb{R}$ is defined as

$$\operatorname{sign}(x) = -\mathbb{1}_{(-\infty,0)}(x) + \mathbb{1}_{[0,+\infty)}(x).$$

The main purpose in Section 4 is to study the tractability of the Wasserstein distributionally robust optimization problem for various risk functions, i.e. solving the following problem:

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho^F(\boldsymbol{\beta}^\top \mathbf{Z}). \tag{38}$$

Below we propose a lemma which demonstrates that solving the inner supremum problem of (38) is equivalent to solve a one-dimensional Wasserstein robust optimization problem.

Lemma 2. *Given $\boldsymbol{\beta} \in \mathbb{R}^n$, we have*

$$\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho^F(\boldsymbol{\beta}^\top \mathbf{Z}) = \sup_{F\in\mathcal{B}_p\left(F_{\boldsymbol{\beta}^\top \mathbf{Z}_0},\varepsilon\|\boldsymbol{\beta}\|_*\right)} \rho^F(Z),$$

where $\mathbf{Z}_0 \sim F_0$ and $F_{\boldsymbol{\beta}^\top \mathbf{Z}_0}$ is the distribution of $\boldsymbol{\beta}^\top \mathbf{Z}_0$.

The proof of Lemma 2 follows immediately from Theorem 2. Based on Lemma 2, we give the following lemma which will be used in the proofs of the main results.

Lemma 3. *Let $p \in [1,\infty]$ and $C > 0$. For $\rho : L^p \to \mathbb{R}$, the following two statements are equivalent.*

(i) For any $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$, $\varepsilon \geqslant 0$ and $\mathcal{D} \subseteq \mathbb{R}^n$,

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho^F(\boldsymbol{\beta}^\top \mathbf{Z}) = \inf_{\boldsymbol{\beta}\in\mathcal{D}} \left\{ \rho^{F_0}(\boldsymbol{\beta}^\top \mathbf{Z}) + C\varepsilon\|\boldsymbol{\beta}\|_* \right\}.$$

(ii) For any $Z \in L^p$ and $\varepsilon \geqslant 0$,

$$\sup_{\|V\|_p\leqslant\varepsilon} \rho(Z+V) = \rho(Z) + C\varepsilon. \tag{39}$$

Proof. The implication of (i)⇒(ii) follows from Lemma 2 and setting $\mathcal{D} = \{\boldsymbol{\beta}_0\}$ with $\boldsymbol{\beta}_0 = (1,0,\dots,0)$, and $\mathbf{Z}_0 \sim F_0$ where $\mathbf{Z}_0 = (Z_0,0,\dots,0)$ with $Z_0 \in L^1$. For the implication (ii)⇒(i), let

$\mathbf{Z}_0 \sim F_0$. By Lemma 2, we have

$$\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \rho^F(\boldsymbol{\beta}^\top \mathbf{Z}) = \sup_{F\in\mathcal{B}_p\left(F_{\boldsymbol{\beta}^\top \mathbf{Z}_0},\varepsilon\|\boldsymbol{\beta}\|_*\right)} \rho^F(Z) = \sup_{\{Z:\|Z-\boldsymbol{\beta}^\top \mathbf{Z}_0\|_p\leqslant\varepsilon\|\boldsymbol{\beta}\|_*\}} \rho(Z)$$
$$= \sup_{\|V\|_p\leqslant\varepsilon\|\boldsymbol{\beta}\|_*} \rho(\boldsymbol{\beta}^\top \mathbf{Z}_0 + V) = \rho(\boldsymbol{\beta}^\top \mathbf{Z}_0) + C\varepsilon\|\boldsymbol{\beta}\|_*.$$

This completes the proof. □

B.1 Proof of Theorem 5

To prove Theorem 5, we need the following auxiliary lemma which will also be used in the proof of Theorem 7.

Lemma 4. *Let $p \in (1,\infty)$, $\varepsilon > 0$ and $\eta \in (0,\varepsilon]$. Define $\mathcal{V} = \{V \in L^p : \|V\|_p \leqslant \varepsilon,\ \mathbb{E}[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}] \leqslant \eta\}$. We have $\mathbb{E}[|V|] \leqslant \varepsilon 2^{-p/q} + \eta$ for all $V \in \mathcal{V}$.*

Proof. Using the Chebyshev's inequality, we have $(2\varepsilon)^p\mathbb{P}(|V| > 2\varepsilon) \leqslant \|V\|_p^p$, which implies $\mathbb{P}(|V| > 2\varepsilon) \leqslant 2^{-p}$ for all $\|V\|_p \leqslant \varepsilon$. Hence, for any $V \in \mathcal{V}$, it holds that

$$\mathbb{E}[|V|] = \mathbb{E}\left[|V|\mathbb{1}_{\{|V|>2\varepsilon\}}\right] + \mathbb{E}\left[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}\right] \leqslant \|V\|_p(\mathbb{P}(|V| > 2\varepsilon))^{1/q} + \eta \leqslant \varepsilon 2^{-p/q} + \eta,$$

where the first inequality follows from Hölder inequality. This completes the proof. □

Proof of Theorem 5 By Lemma 3, we know that (i) is equivalent to the following statement.

(i*) For any $Z \in L^p$ and $\varepsilon \geqslant 0$, we have (39) holds.

It suffices to show (i*) $\Leftrightarrow$ (ii). To see (ii)$\Rightarrow$(i*), note that for ℓ_1, denote $C_1 = C$ and $C_2 = -C$, and we have that for $i = 1, 2$,

$$\sup_{\|V\|_p\leqslant\varepsilon} \mathbb{E}[\ell_1(Z_0 + V)] = \sup_{\|V\|_p\leqslant\varepsilon} \{\mathbb{E}[C_i Z_0 + b] + C_i\mathbb{E}[V]\}$$
$$= \mathbb{E}[\ell_1(Z_0)] + C_i \sup_{\|V\|_p\leqslant\varepsilon} \mathbb{E}[V] = \mathbb{E}[\ell_1(Z_0)] + C\varepsilon.$$

For ℓ_2, we have

$$\sup_{\|V\|_p\leqslant\varepsilon} \mathbb{E}[\ell_2(Z_0 + V)] = \sup_{\|V\|_p\leqslant\varepsilon} \mathbb{E}[C|Z_0 + V - m| + b]$$
$$\leqslant \sup_{\|V\|_p\leqslant\varepsilon} \{\mathbb{E}[C|Z_0 - m| + b] + C\mathbb{E}[|V|]\} \leqslant \mathbb{E}[\ell_2(Z_0)] + C\varepsilon.$$

All inequalities above can be equality by letting $V = \varepsilon \text{sign}(Z_0 - m)$. Hence, we complete the proof of (ii)⇒(i*).

To see (i*)⇒(ii), assume without loss of generality that $C = 1$. We first show that $\text{Lip}(\ell) \leqslant 1$. Otherwise, by that a convex function has derivative almost everywhere, there exists x such that $|\ell'(x)| > 1$. It follows that

$$\sup_{\|V\|_p \leqslant \varepsilon} \mathbb{E}[\ell(x+V)] \geqslant \ell(x + \varepsilon \,\text{sign}(\ell'(x))) \geqslant \ell(x) + |\ell'(x)|\varepsilon > \ell(x) + \varepsilon,$$

This yields a contradiction to (39). Next, we aim to verify the following fact.

$$|\ell'(x)| = 1 \text{ for all } x \in \mathbb{R} \text{ such that } \ell \text{ is differentiable at } x. \tag{40}$$

This will complete the proof since the convex function that satisfies (40) must be one of the forms of ℓ_1 and ℓ_2 with $C = 1$. To see (40), assume by contradiction that there exists $x_0 \in \mathbb{R}$ such that $|\ell'(x_0)| < 1$. If $p = \infty$, then we have

$$\sup_{\|V\|_\infty \leqslant \varepsilon} \mathbb{E}[\ell(x_0+V)] = \max\{\ell(x_0-\varepsilon), \ell(x_0+\varepsilon)\} < \ell(x_0) + \varepsilon,$$

where the strict inequality follows from $|\ell'(x_0)| < 1$ and $\text{Lip}(\ell) \leqslant 1$, which yields a contradiction. Suppose now $p \in (1, \infty)$. Define

$$\mathcal{V}_1 = \left\{ V \in L^p : \|V\|_p \leqslant \varepsilon,\ \mathbb{E}\left[|V| \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}\right] \leqslant \frac{(1 - 2^{-p/q})\varepsilon}{2} \right\} \text{ and } \mathcal{V}_2 = \{V : \|V\|_p \leqslant \varepsilon\} \setminus \mathcal{V}_1.$$

By Lemma 4, we have

$$\mathbb{E}[|V|] \leqslant \frac{(1 - 2^{-p/q})\varepsilon}{2} + \varepsilon 2^{-p/q} = \frac{(1 + 2^{-p/q})\varepsilon}{2}, \quad \forall\, V \in \mathcal{V}_1. \tag{41}$$

Below we calculate the value of the left hand side of (39) by letting $Z = x_0$ a.s.

$$\begin{aligned}
\sup_{\|V\|_p \leqslant \varepsilon} \mathbb{E}[\ell(x_0+V)] &= \sup_{\|V\|_p \leqslant \varepsilon} \{\mathbb{E}[\ell(x_0+V)\mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] + \mathbb{E}[\ell(x_0+V)\mathbb{1}_{\{|V| > 2\varepsilon\}}]\} \\
&\leqslant \sup_{\|V\|_p \leqslant \varepsilon} \{\mathbb{E}[(\ell(x_0) + k|V|)\mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] + \mathbb{E}[(\ell(x_0) + |V|)\mathbb{1}_{\{|V| > 2\varepsilon\}}]\} \\
&= \ell(x_0) + \sup_{\|V\|_p \leqslant \varepsilon} \{\mathbb{E}[|V|] - (1-k)\mathbb{E}[|V|\mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}]\} =: \max\{I_1, I_2\},
\end{aligned}$$

where

$$k = \max\left\{\left|\frac{\ell(x_0+2\varepsilon)-\ell(x_0)}{2\varepsilon}\right|, \left|\frac{\ell(x_0)-\ell(x_0-2\varepsilon)}{2\varepsilon}\right|\right\},$$

$$I_i = \ell(x_0) + \sup_{V\in\mathcal{V}_i}\{\mathbb{E}[|V|] - (1-k)\mathbb{E}[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}]\}, \quad i=1,2,$$

and the inequality follows from $\mathrm{Lip}(\ell)\leqslant 1$ and the fact that the convexity of ℓ implies that $|\ell(x_0+v)-\ell(x_0)|\leqslant k|v|$ for all $v\in[-2\varepsilon,2\varepsilon]$. Since $|\ell'(x_0)|<1$ and $\mathrm{Lip}(\ell)\leqslant 1$, we know that $k<1$. Hence, it holds that

$$I_1 \leqslant \ell(x_0) + \sup_{V\in\mathcal{V}_1}\mathbb{E}[|V|] \leqslant \ell(x_0) + \frac{(1+2^{-p/q})\varepsilon}{2} < \ell(x_0)+\varepsilon, \tag{42}$$

where the second inequality follows from (41). For I_2, we have

$$\begin{aligned} I_2 &\leqslant \ell(x_0) + \sup_{V\in\mathcal{V}_2}\left\{\mathbb{E}[|V|] - (1-k)\frac{(1-2^{-p/q})\varepsilon}{2}\right\} \\ &\leqslant \ell(x_0)+\varepsilon-(1-k)\frac{(1-2^{-p/q})\varepsilon}{2} < \ell(x_0)+\varepsilon, \end{aligned} \tag{43}$$

where the first inequality follows from the definition of $\mathcal{V}_2$, and the second one holds because $\|V\|_p\leqslant\varepsilon$ implies $\mathbb{E}[|V|]\leqslant\varepsilon$. Combining with (42) and (43), we conclude that

$$\sup_{\|V\|_p\leqslant\varepsilon}\mathbb{E}[\ell(x_0+V)] \leqslant \max\{I_1,I_2\} < \ell(x_0)+\varepsilon.$$

This yields a contradiction. Hence, (40) is verified, and thus we complete the proof. □

B.2 Proofs of Theorems 6 and 7

Proof of Theorem 6 By Lemma 3, it suffices to verify that for any $Z\in L^p$ and $\varepsilon\geqslant 0$,

$$\sup_{\|V\|_p\leqslant\varepsilon}\|\ell(Z+V)\|_p = \|\ell(Z)\|_p+\varepsilon. \tag{44}$$

For $\ell=\ell_1$, we have

$$\begin{aligned} \sup_{\|V\|_p\leqslant\varepsilon}\|\ell_1(Z+V)\|_p &= \sup_{\|V\|_p\leqslant\varepsilon}\|(Z+V-m)_+\|_p \leqslant \sup_{\|V\|_p\leqslant\varepsilon}\|(Z-m)_+ + |V|\|_p \\ &\leqslant \sup_{\|V\|_p\leqslant\varepsilon}\{\|(Z-m)_+\|_p + \|V\|_p\} \leqslant \|(Z-m)_+\|_p+\varepsilon = \|\ell_1(Z)\|_p+\varepsilon. \end{aligned} \tag{45}$$

If $\mathbb{P}(Z>m)>0$, then one can check that all inequalities are equality, and the maximizer can be

chosen as $V = \lambda(Z-m)_+$ with some $\lambda \geqslant 0$ such that $\|V\|_p = \varepsilon$. If $Z \leqslant m$ a.s., then we take $\{V_n\}_{n\in\mathbb{N}}$ such that V_n has distribution $(1-1/n^p)\delta_0 + (1/n^p)\delta_{n\varepsilon}$, and $\{V_n = n\varepsilon\} \subseteq \{Z \geqslant F_Z^{-1}(1-1/n^p)\}$ where F_Z^{-1} is the left-quantile function of Z for all n. We have $\|V_n\|_p = \varepsilon$ and

$$\sup_{\|V\|_p\leqslant\varepsilon} \|\ell_1(Z+V)\|_p \geqslant \|(Z+V_n-m)_+\|_p \geqslant \frac{1}{n}\left(n\varepsilon + F_Z^{-1}\left(1-\frac{1}{n^p}\right) - m\right)_+ \\ \to \varepsilon = \|\ell_1(Z)\|_p + \varepsilon, \quad \text{as } n\to\infty.$$

Combining with (45), we verify that (44) holds with ℓ_1 for all $Z \leqslant m$ a.s. This completes the proof of the case $\ell = \ell_1$.

For $\ell = \ell_2$, we can follow the similar argument to ℓ_1 to verify the result.

For $\ell = \ell_3$, we have

$$\begin{aligned}\sup_{\|V\|_p\leqslant\varepsilon} \|\ell_3(Z+V)\|_p &= \sup_{\|V\|_p\leqslant\varepsilon} \|(|Z+V-m_1|-m_2)_+\|_p \leqslant \sup_{\|V\|_p\leqslant\varepsilon} \|(|Z-m_1|-m_2+|V|)_+\|_p \\ &\leqslant \sup_{\|V\|_p\leqslant\varepsilon} \|(|Z-m_1|-m_2)_+ + |V|\|_p \leqslant \sup_{\|V\|_p\leqslant\varepsilon} \{\|(|Z-m_1|-m_2)_+\|_p + \|V\|_p\} \\ &\leqslant \|(|Z-m_1|-m_2)_+\|_p + \varepsilon = \|\ell_3(Z)\|_p + \varepsilon. \end{aligned} \tag{46}$$

If $\mathbb{P}(|Z-m_1| > m_2) > 0$, then one can check that all inequalities are equalities, and the maximizer can be chosen as $V = \lambda(|Z-m_1|-m_2)_+\mathrm{sign}(Z-m_1)$ with some $\lambda \geqslant 0$ such that $\|V\|_p = \varepsilon$. If $|Z-m_1| \leqslant m_2$ a.s., then we have $Z \in [m_1-m_2, m_1+m_2]$. Taking a sequence $\{V_n\}_{n\in\mathbb{N}}$ as shown in the case of ℓ_1, i.e., V_n with distribution $(1-1/n^p)\delta_0 + (1/n^p)\delta_{n\varepsilon}$, it holds that for large enough n such that $n\varepsilon \geqslant \max\{m_1, m_2\}$,

$$\begin{aligned}\sup_{\|V\|_p\leqslant\varepsilon} \|\ell_3(Z+V)\|_p &\geqslant \|(|Z+V_n-m_1|-m_2)_+\|_p \\ &\geqslant \left(\mathbb{E}[(|Z-m_1|-m_2)_+^p \mathbb{1}_{\{V_n=0\}}] + \mathbb{E}[(|Z-m_1+n\varepsilon|-m_2)_+^p \mathbb{1}_{\{V_n=n\varepsilon\}}]\right)^{1/p} \\ &\geqslant \left(\mathbb{E}[(n\varepsilon-m_2-m_2)_+^p \mathbb{1}_{\{V_n=n\varepsilon\}}]\right)^{1/p} \\ &= \left(\varepsilon - \frac{2m_2}{n}\right)_+ \to \varepsilon = \|\ell_3(Z)\|_p + \varepsilon, \quad \text{as } n\to\infty. \end{aligned}$$

Combining with (46), we verify that (44) holds with ℓ_3 for all $|Z-m_1| \leqslant m_2$ a.s. This completes the proof of the case ℓ_3.

For $\ell = \ell_4$, we have

$$\begin{aligned}\sup_{\|V\|_p \leqslant \varepsilon} \|\ell_4(Z+V)\|_p &= \sup_{\|V\|_p \leqslant \varepsilon} \||\|Z+V-m|+b\|_p \leqslant \sup_{\|V\|_p \leqslant \varepsilon} \||\|Z-m|+b+|V|\|_p \\ &\leqslant \sup_{\|V\|_p \leqslant \varepsilon} \{\||\|Z-m|+b\|_p + \|V\|_p\} \leqslant \||\|Z-m|+b\|_p + \varepsilon = \|\ell_4(Z)\|_p + \varepsilon,\end{aligned}$$

where all inequalities can be equality, and the maximizer can be chosen as $V = \lambda(|Z-m|+b)\mathrm{sign}(Z-m)$ for some $\lambda \geqslant 0$ such that $\|V\|_p = \varepsilon$. Hence, we conclude that (44) holds with $\ell = \ell_4$. Therefore, we complete the proof for all forms of loss functions. □

The following lemma will be used in the proof of Theorem 7.

Lemma 5. *Let $p \in (1,\infty)$ be an integer, $t, \varepsilon > 0$ and $\eta \in [0,\varepsilon)$. For $V \in L^p$, if $\|V\|_p \leqslant \varepsilon$ and $\mathbb{E}[|V|] \leqslant \varepsilon - \eta$, then $\mathbb{E}[(|V|+t)^p] \leqslant (\varepsilon+t)^p - pt^{p-1}\eta$. In particular, by letting $\eta = 0$, we have $\mathbb{E}[(|V|+t)^p] \leqslant (\varepsilon+t)^p$ for all V such that $\|V\|_p \leqslant \varepsilon$.*

Proof. Note that

$$\begin{aligned}\mathbb{E}[(|V|+t)^p] &= \mathbb{E}\left[\sum_{i=0}^{p} \binom{p}{i} t^{p-i}|V|^i\right] = \sum_{i\neq 1} \binom{p}{i} t^{p-i}\mathbb{E}[|V|^i] + pt^{p-1}\mathbb{E}[|V|] \\ &\leqslant \sum_{i\neq 1} \binom{p}{i} t^{p-i}\varepsilon^i + pt^{p-1}(\varepsilon-\eta) = \sum_{i=0}^{p} \binom{p}{i} t^{p-i}\varepsilon^i - pt^{p-1}\eta = (\varepsilon+t)^p - pt^{p-1}\eta,\end{aligned}$$

where the inequality holds because $\|V\|_p \leqslant \varepsilon$ implies $\|V\|_i \leqslant \varepsilon$ for $i = 1, 2, \ldots, p$. □

Proof of Theorem 7 By Lemma 3, we have for any $Z \in L^p$ and $\varepsilon \geqslant 0$,

$$\sup_{\|V\|_p \leqslant \varepsilon} \|\ell(Z+V)\|_p = \|\ell(Z)\|_p + C\varepsilon. \tag{47}$$

Assume without loss of generality that $C = 1$. We first show that $\mathrm{Lip}(\ell) \leqslant 1$. Otherwise, there exists $x \in \mathbb{R}$ such that $|\ell'(x)| > 1$. It holds that

$$\sup_{\|V\|_p \leqslant \varepsilon} \|\ell(x+V)\|_p \geqslant \ell(x+\varepsilon\mathrm{sign}(\ell'(x))) \geqslant \ell(x) + |\ell'(x)|\varepsilon > \ell(x) + \varepsilon,$$

where the second inequality follows from the convexity of ℓ. This yields a contradiction to (47), and thus, $\mathrm{Lip}(\ell) \leqslant 1$. Next, assume that ℓ is differential at x when we use the notation $\ell'(x)$, and

we show the following facts.

$$\text{If } |\ell'(x)| > 0, \text{ then } |\ell'(x)| = 1. \quad \text{If } \ell'(x) = 0, \text{ then } \ell(x) = 0. \tag{48}$$

This will complete the proof since one can check that (48) implies that ℓ has one of the forms of ℓ_1, ℓ_2, ℓ_3 and ℓ_4 with $C = 1$. To see (48), we assume by contradiction that there exists $x_0 \in \mathbb{R}$ such that $|\ell'(x_0)| < 1$ and $\ell(x_0) > 0$ (note that $|\ell'(x_0)| \in (0, 1)$ implies $\ell(x_0) > 0$). Define $\mathcal{V}_1$ and $\mathcal{V}_2$ as the same as the ones in the proof of Theorem 5, i.e.

$$\mathcal{V}_1 = \left\{ V \in L^p : \|V\|_p \leqslant \varepsilon,\ \mathbb{E}\left[|V| \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}\right] \leqslant \frac{(1 - 2^{-p/q})\varepsilon}{2} \right\}.$$

and $\mathcal{V}_2 = \{V : \|V\|_p \leqslant \varepsilon\} \setminus \mathcal{V}_1$. By Lemma 4, we have

$$\mathbb{E}[|V|] \leqslant \frac{(1 + 2^{-p/q})\varepsilon}{2}, \quad \forall V \in \mathcal{V}_1. \tag{49}$$

Note that

$$\begin{aligned}
\sup_{\|V\|_p \leqslant \varepsilon} \mathbb{E}[\ell^p(x_0 + V)] &= \sup_{\|V\|_p \leqslant \varepsilon} \left\{ \mathbb{E}[\ell^p(x_0 + V) \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] + \mathbb{E}[\ell^p(x_0 + V) \mathbb{1}_{\{|V| > 2\varepsilon\}}] \right\} \\
&\leqslant \sup_{\|V\|_p \leqslant \varepsilon} \left\{ \mathbb{E}[(\ell(x_0) + k|V|)^p \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] + \mathbb{E}[(\ell(x_0) + |V|)^p \mathbb{1}_{\{|V| > 2\varepsilon\}}] \right\} \qquad (50) \\
&= \sup_{\|V\|_p \leqslant \varepsilon} \left\{ \mathbb{E}[(\ell(x_0) + |V|)^p] - \mathbb{E}[((\ell(x_0) + |V|)^p - (\ell(x_0) + k|V|)^p) \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] \right\} \\
&\leqslant \sup_{\|V\|_p \leqslant \varepsilon} \left\{ \mathbb{E}[(\ell(x_0) + |V|)^p] - p\ell^{p-1}(x_0)(1 - k)\mathbb{E}[|V| \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] \right\} \qquad (51) \\
&=: \max\{I_1, I_2\},
\end{aligned}$$

where

$$k = \max\left\{ \left| \frac{\ell(x_0 + 2\varepsilon) - \ell(x_0)}{2\varepsilon} \right|, \left| \frac{\ell(x_0) - \ell(x_0 - 2\varepsilon)}{2\varepsilon} \right| \right\},$$

$$I_i = \sup_{V \in \mathcal{V}_i} \left\{ \mathbb{E}[(\ell(x_0) + |V|)^p] - p\ell^{p-1}(x_0)(1 - k)\mathbb{E}[|V| \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}] \right\}, \quad i = 1, 2,$$

(50) follows from that $0 \leqslant \ell(x_0 + v) \leqslant \ell(x_0) + k|v|$ for all $v \in [-2\varepsilon, 2\varepsilon]$ as ℓ is nonnegative and convex, $0 \leqslant \ell(x_0 + v) \leqslant \ell(x_0) + v$ as $\mathrm{Lip}(\ell) \leqslant 1$, and (51) follows from $|\ell'(x_0)| < 1$ and $\mathrm{Lip}(\ell) \leqslant 1$ which imply $k < 1$ and $(\ell(x_0) + |v|)^p - (\ell(x_0) + k|v|)^p \geqslant p\ell^{p-1}(x_0)(1 - k)|v|$, $\forall v \in \mathbb{R}$. Further, note

that

$$I_1 \leqslant \sup_{V\in\mathcal{V}_1} \mathbb{E}[(\ell(x_0)+|V|)^p] \leqslant (\ell(x_0)+\varepsilon)^p - p\ell^{p-1}(x_0)\varepsilon\left(1-\frac{1+2^{-p/q}}{2}\right) < (\ell(x_0)+\varepsilon)^p, \quad (52)$$

where the second inequalities follows from (49) and Lemma 5. For I_2, we have

$$\begin{aligned} I_2 &\leqslant \sup_{V\in\mathcal{V}_2} \mathbb{E}[(\ell(x_0)+|V|)^p] - \inf_{V\in\mathcal{V}_2} p\ell^{p-1}(x_0)(1-k)\mathbb{E}[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}] \\ &\leqslant (\ell(x_0)+\varepsilon)^p - p\ell^{p-1}(x_0)(1-k)\inf_{V\in\mathcal{V}_2} \mathbb{E}[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}] \\ &\leqslant (\ell(x_0)+\varepsilon)^p - p\ell^{p-1}(x_0)(1-k)\frac{(1-2^{-p/q})\varepsilon}{2} < (\ell(x_0)+\varepsilon)^p, \end{aligned} \quad (53)$$

where the second inequality follows from Lemma 5, and the third inequality is due to the definition of $\mathcal{V}_2$. Combining (52) and (53), we have

$$\sup_{\|V\|_p\leqslant\varepsilon} \|\ell(x_0+V)\|_p \leqslant \max\{I_1^{1/p}, I_2^{1/p}\} < \ell(x_0)+\varepsilon,$$

which yields a contradiction to (47). Hence, (48) holds which completes the proof. □

B.3 Proofs of Corollaries 1 and 2

The proofs of Corollaries 1 and 2 rely on the following lemma and the regularization results in Theorem 6. Recall that

$$\pi_{1,\ell}(F,t) = t + \left(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t)]\right)^{1/p}, \quad \pi_{2,\ell}(F,t) = \mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t)].$$

Lemma 6. *For any $p\in[1,\infty)$, $F_0\in\mathcal{M}_p(\mathbb{R}^n)$ and $\varepsilon\geqslant 0$, the following two statements hold.*

(i) Suppose that $\ell(z,t)$ is nonnegative on $\mathbb{R}^2$, and convex in t with $\lim_{t\to-\infty}\partial\ell(z,t)/\partial t<-1$ for all $z\in\mathbb{R}$, and Lipschitz continuous in z for all $t\in\mathbb{R}$ with a uniform Lipschitz constant, i.e. there exists $M>0$ such that

$$|\ell(z_1,t)-\ell(z_2,t)| \leqslant M|z_1-z_2|, \quad \forall t,z_1,z_2\in\mathbb{R}.$$

Then we have

$$\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \pi_{1,\ell}(F,t) = \inf_{t\in\mathbb{R}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \pi_{1,\ell}(F,t).$$

(ii) Suppose that $\ell(z,t)$ is convex in t with $\lim_{t\to-\infty}\partial\ell(z,t)/\partial t<0$ and $\lim_{t\to+\infty}\partial\ell(z,t)/\partial t>0$ for all $z\in\mathbb{R}$, and Lipschitz continuous in z for all $t\in\mathbb{R}$ with a uniform Lipschitz constant. Then we have

$$\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)}\inf_{t\in\mathbb{R}}\pi_{2,\ell}(F,t)=\inf_{t\in\mathbb{R}}\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)}\pi_{2,\ell}(F,t).$$

Proof. (i) First, we show three facts below. (a) $\pi_{1,\ell}(F,t)$ is concave in F for all $t\in\mathbb{R}$; (b) $\pi_{1,\ell}(F,t)$ is convex in t for all $F\in\mathcal{M}_p(\mathbb{R}^n)$; (c) $\lim_{t\to\pm\infty}\pi_{1,\ell}(F,t)=\infty$ for all $F\in\mathcal{M}_p(\mathbb{R}^n)$. The fact (a) is trivial. For $F\in\mathcal{M}_p(\mathbb{R}^n)$, $\lambda\in[0,1]$ and $t_1,t_2\in\mathbb{R}$, it holds that

$$\begin{aligned}\left(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},\lambda t_1+(1-\lambda)t_2)]\right)^{1/p}&\leqslant\left(\mathbb{E}^F\left[\left(\lambda\ell(\boldsymbol{\beta}^\top\mathbf{Z},t_1)+(1-\lambda)\ell(\boldsymbol{\beta}^\top\mathbf{Z},t_2)\right)^p\right]\right)^{1/p}\\&\leqslant\lambda\left(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t_1)]\right)^{1/p}+(1-\lambda)\left(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t_2)]\right)^{1/p},\end{aligned}$$

where the first step holds because ℓ is nonnegative, and the second follows from the triangle inequality. This implies (b). To see (c), it is obvious that $\lim_{t\to\infty}\pi_{1,\ell}(F,t)=\infty$. Note that $(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t)])^{1/p}\geqslant\mathbb{E}^F[\ell(\boldsymbol{\beta}^\top\mathbf{Z},t)]$. Combining with $\lim_{t\to-\infty}\partial\ell(z,t)/\partial t<-1$ and the convexity of $\ell(z,t)$ in t, we have $\lim_{t\to-\infty}\pi_{1,\ell}(F,t)=\infty$. Hence, we conclude the proof of (c). Using (b) and (c), the set of all minimizers of the problem $\inf_{t\in\mathbb{R}}\pi_{1,\ell}(F,t)$ is a closed interval. Denote by $t(F):=\inf\arg\min_t\pi_{1,\ell}(F,t)$. We will show that $\{t(F):F\in\mathcal{B}_p(F_0,\varepsilon)\}$ is a subset of a compact set. For any $F\in B_p(F_0,\varepsilon)$ and $t\in\mathbb{R}$, let $\mathbf{Z}\sim F$ and $\mathbf{Z_0}\sim F_0$, and we have

$$\begin{aligned}|\pi_{1,\ell}(F,t)-\pi_{1,\ell}(F_0,t)|&=\left|\left(\mathbb{E}[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z},t)]\right)^{1/p}-\left(\mathbb{E}^F[\ell^p(\boldsymbol{\beta}^\top\mathbf{Z_0},t)]\right)^{1/p}\right|\\&\leqslant\left(\mathbb{E}[|\ell(\boldsymbol{\beta}^\top\mathbf{Z},t)-\ell(\boldsymbol{\beta}^\top\mathbf{Z}_0,t)|^p]\right)^{1/p}\leqslant\left(\mathbb{E}[M^p|\boldsymbol{\beta}^\top(\mathbf{Z}-\mathbf{Z}_0)|^p]\right)^{1/p}\\&\leqslant M\|\boldsymbol{\beta}\|_*\left(\mathbb{E}[\|\mathbf{Z}-\mathbf{Z}_0\|^p]\right)^{1/p}\leqslant M\|\boldsymbol{\beta}\|_*\varepsilon,\end{aligned}\tag{54}$$

where the first and the third inequalities follow from the triangle inequality and Hölder inequality, respectively, and we have used the definition of the Wasserstein ball $\mathcal{B}_p(F_0,\varepsilon)$ in the last step. Hence, it holds that

$$\pi_{1,\ell}(F,t(F_0))\leqslant\pi_{1,\ell}(F_0,t(F_0))+M\|\boldsymbol{\beta}\|_*\varepsilon.\tag{55}$$

Note that $\pi_{1,\ell}(F_0,t)\to+\infty$ as $t\to\pm\infty$. There exists $\Delta>0$ such that $\pi_{1,\ell}(F_0,t)>\pi_{1,\ell}(F_0,t(F_0))+$

$2M\|\boldsymbol{\beta}\|_*\varepsilon$ for all $t \notin [t(F_0)-\Delta, t(F_0)+\Delta]$. This, combined with (54), imply that

$$\pi_{1,\ell}(F,t) \geqslant \pi_{1,\ell}(F_0,t) - M\|\boldsymbol{\beta}\|_*\varepsilon > \pi_{1,\ell}(F_0,t(F_0)) + M\|\boldsymbol{\beta}\|_*\varepsilon, \quad \forall t \notin [t(F_0)-\Delta, t(F_0)+\Delta]. \quad (56)$$

Applying (55) and (56), we have $\{t(F) : F \in \mathcal{B}_p(F_0,\varepsilon)\} \subseteq [t(F_0)-\Delta, t(F_0)+\Delta]$. Using a minimax theorem (see e.g., Sion (1958)), it holds that

$$\begin{aligned}\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \pi_{1,\ell}(F,t) &= \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \inf_{t\in[t(F_0)-\Delta,t(F_0)+\Delta]} \pi_{1,\ell}(F,t) \\ &= \inf_{t\in[t(F_0)-\Delta,t(F_0)+\Delta]} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \pi_{1,\ell}(F,t) \geqslant \inf_{t\in\mathbb{R}} \sup_{F\in\mathcal{B}_p(F_0,\varepsilon)} \pi_{1,\ell}(F,t).\end{aligned}$$

The converse direction is trivial. Hence, we complete the proof.

(ii) The proof is similar to (i) by considering the minimax problem of $(\pi_{2,\ell}(F,t))^{1/p}$. □

Proof of Corollary 1 One can check that all forms of ℓ in this corollary satisfy the conditions in Lemma 6 (i). Hence, the result follows immediately from Lemma 6 and Theorem 6. □

Proof of Corollary 2 One can check that all forms of ℓ in this corollary satisfy the conditions in Lemma 6 (ii). Hence, the result follows immediately from Lemma 6 and Theorem 6. □

B.4 Proofs of the cases of exponential functions and distortion functionals

Proof of Proposition 1 Suppose that $\rho : L^\infty \to \mathbb{R}$ is a monotonic function. It holds that

$$\sup_{F\in\mathcal{B}_\infty(F_0,\varepsilon)} \rho^F(Z) = \sup_{\|Z-Z_0\|_\infty\leqslant\varepsilon} \rho(Z) = \sup_{\|V\|_\infty\leqslant\varepsilon} \rho(Z_0+V) = \rho(Z_0+\varepsilon),$$

where $Z_0 \sim F_0$, and the last equality follows from the monotonicity of ρ. Then the desired result follows by applying Lemma 3. □

To prove the result in the case of distortion functionals, we need some preliminaries for distortion functionals.

Lemma 7 (Wang et al. (2020)). *Let $h : [0,1] \to \mathbb{R}$ be a distortion function. The following statements hold.*

(i) ρ_h is monotone, i.e. $\rho(Z_1) \leqslant \rho(Z_2)$ for $Z_1 \leqslant Z_2$ a.s., if and only if h is increasing.

(ii) ρ_h is subadditive, i.e. $\rho_h(Z_1+Z_2) \leqslant \rho_h(Z_1)+\rho_h(Z_2)$ for any Z_1 and Z_2, if and only if h is convex. The equality holds when Z_1 and Z_2 are comonotonic[2].

[2]Two random variables Z_1 and Z_2 are said to be comonotonic if (Z_1, Z_2) is distributionally equivalent to

Proof of Proposition 2 (i) For a distortion function h, we have $\rho_h^F(Z-\mathbb{E}^F[Z])=\rho_{\widetilde{h}}^F(Z)$, where $\widetilde{h}(s)=h(s)+(h(0)-h(1))s$ for $s\in[0,1]$. Note that $\widetilde{h}$ is convex whenever h is convex. Hence, it suffices to prove that for any convex distortion function h

$$\inf_{\boldsymbol{\beta}\in\mathcal{D}}\sup_{F\in\mathcal{B}_p(F_0,\varepsilon)}\rho_h^F(\boldsymbol{\beta}^\top\mathbf{X})=\inf_{\boldsymbol{\beta}\in\mathcal{D}}\left\{\rho_h^{F_0}(\boldsymbol{\beta}^\top\mathbf{X})+\|h'_-\|_q\varepsilon||\boldsymbol{\beta}||_*\right\},$$

which, by applying Lemma 3, is equivalent to prove that

$$\sup_{\|V\|_p\leqslant\varepsilon}\rho_h(Z+V)=\rho_h(Z)+\varepsilon\|h'_-\|_q,\quad\forall Z\in L^p. \tag{57}$$

To see this, by Lemma 7, we have for any $\|V\|_p\leqslant\varepsilon$,

$$\rho_h(Z+V)\leqslant\rho_h(Z)+\rho_h(V)\leqslant\rho_h(Z)+\|V\|_p\|h'_-\|_q\leqslant\rho_h(Z)+\varepsilon\|h'_-\|_q, \tag{58}$$

where the second inequality follows from Hölder inequality. Suppose that $Z=F_Z^{-1}(U)$ a.s. where U is a uniform random variable on $[0,1]$ (see Lemma A.28 of Föllmer and Schied (2016) for the existence of U). In the following, we consider the cases of $p=1$ and $p\in(1,+\infty]$ separately.

(a) If $p=1$, then $q=+\infty$ and $\|h'_-\|_\infty=\max\{|h'_-(0+)|,h'_-(1-)\}$. In this case, we first assume that $h'_-(1-)\geqslant|h'_-(0+)|$. Define a sequence $\{V_n\}_{n\in\mathbb{N}}$ such that $V_n=n\varepsilon\mathbb{1}_{\{U>1-1/n\}}$. For all $n\in\mathbb{N}$, it holds that $\|V_n\|_1=\varepsilon$, and V_n and Z are comonotonic. Hence, we have

$$\begin{aligned}\sup_{\|V\|_p\leqslant\varepsilon}\rho_h(Z+V)\geqslant\rho_h(Z+V_n)&=\rho_h(Z)+n\varepsilon\int_{1-1/n}^1h'_-(s)\mathrm{d}s\\&\to\rho_h(Z)+\varepsilon h'_-(1-)=\rho_h(Z)+\varepsilon\|h'_-\|_\infty,\quad\text{as } n\to\infty.\end{aligned} \tag{59}$$

Combining (58) and (59), we have verified (57). Assume now $h'_-(1-)<|h'_-(0+)|$. We can construct a sequence $\{V_n\}_{n\in\mathbb{N}}$ such that $V_n=-n\varepsilon\mathbb{1}_{\{U<1/n\}}$, and then follow the same analysis as the previous argument to verify the result.

(b) If $p\in(1,\infty]$, then we define $\widetilde{V}=\mathrm{sgn}(h'_-(U))\varepsilon|h'_-(U)|^{q/p}/\|h'_-\|_q^{q/p}$. One can verify that $\|\widetilde{V}\|_p=\varepsilon$ and $\rho_h(Z+\widetilde{V})=\rho_h(Z)+\varepsilon\|h'_-\|_q$. This, together with (58), implies that (57) holds.

Hence, we complete the proof of (i).

$(F_{Z_1}^{-1}(U),F_{Z_2}^{-1}(U))$, where U is a random variable uniformly distributed on the interval $[0,1]$ (see e.g., Dhaene et al. (2002) for a discussion of comonotonic random variables).

(ii) By Lemma 3, it is sufficient to prove that

$$\sup_{\|V\|_p \leqslant \varepsilon} \rho_h(|Z+V|) = \rho_h(|Z|) + \varepsilon \|h'_-\|_q, \quad \forall Z \in L^p. \tag{60}$$

To see this, note that h is an increasing and convex distortion function. By Lemma 7, we have for any $\|V\|_p \leqslant \varepsilon$,

$$\rho_h(|Z+V|) \leqslant \rho_h(|Z|) + \rho_h(|V|) \leqslant \rho_h(|Z|) + \|V\|_p \|h'_-\|_q \leqslant \rho_h(|Z|) + \varepsilon \|h'_-\|_q, \tag{61}$$

where the second inequality follows from Hölder inequality. Suppose that $|Z| = F^{-1}_{|Z|}(U)$ a.s. where U is a uniform random variable on $[0,1]$. Define $\widetilde{V} = \mathrm{sign}(Z)\varepsilon (h'_-(U))^{q/p} / \|h'_-\|_q^{q/p}$. One can verify that $\|\widetilde{V}\|_p = \varepsilon$ and $\rho_h(|Z+\widetilde{V}|) = \rho_h(|Z|) + \varepsilon \|h'_-\|_q$. Therefore, we conclude that (60) holds. Hence, we complete the proof. □