When Wasserstein DRO Reduces Exactly: Complete Characterization, Projection Equivalence, and Regularization

Chenling Huang[1], Jonathan Yu-Meng Li[2] and Tiantian Mao[1]

[1] *Department of Statistics and Finance, School of Management*
University of Science and Technology of China, Hefei, Anhui, China.
[2] *Telfer School of Management, University of Ottawa*
Ottawa, Ontario K1N 6N5, Canada
Emails: hcl0210@mail.ustc.edu.cn; jonathan.li@telfer.uottawa.ca; tmao@ustc.edu.cn

September 6, 2025

Abstract

Wasserstein distributionally robust optimization (DRO), a leading paradigm in data-driven decision-making, entails the evaluation of worst-case risk over a high-dimensional Wasserstein ball–a major computational burden. In this paper, we study when the worst-case risk problem admits an exact reduction to the evaluation of risk over a one-dimensional projected Wasserstein ball—a property we refer to as projection equivalence. This reduction depends on the class of risk functions used to evaluate risk: starting from the most general law-invariant risk functions and progressing through monotone risk functions, coherent risk measures, and further specialized classes, we provide a complete characterization–namely, necessary and sufficient conditions on the loss function under which projection equivalence holds. This not only simplifies the evaluation of worst-case risk but also enables a further characterization of cases in which the worst-case problem admits an exact regularization reformulation, significantly extending beyond previously known results. Applications to distributionally robust chance-constrained programs and classification problems are presented.

1 Introduction

Wasserstein distributionally robust optimization (DRO) has emerged as a dominant paradigm for optimization under uncertainty, gaining prominence across operations research, statistics, finance, and machine learning. Its strength lies in safeguarding decisions against distributional ambiguity while maintaining strong out-of-sample guarantees. In its most general form, Wasserstein DRO can be expressed as

$$\min_{f\in\mathcal{F}} \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \rho^F\big(f(\boldsymbol{\xi})\big),$$

where $\mathcal{F}$ denotes the admissible decision class, f represents a decision-dependent loss function, ρ is a risk functional, and $\mathbb{B}_p(F_0, \varepsilon)$ is the p-Wasserstein ball of radius $\varepsilon > 0$ centered at a nominal distribution F_0. This formulation accommodates a wide array of performance criteria through the choice of ρ: expectation in the classical Wasserstein DRO setting, risk measures in finance, statistical functionals in inference, and loss- or utility-based objectives in machine learning.

The primary challenge lies in the inner problem, which we refer to as the worst-case risk problem:

$$\sup_{F \in \mathbb{B}_p(F_0, \varepsilon)} \rho^F\big(f(\boldsymbol{\xi})\big). \tag{1}$$

When the random vector $\boldsymbol{\xi}$ is high-dimensional, evaluating (1) becomes the dominant bottleneck, and addressing this difficulty is critical for both the theoretical analysis and the practical implementation of Wasserstein DRO. In certain cases, however, the high-dimensional worst-case problem (1) admits an *exact* reduction to a one-dimensional problem over a Wasserstein ball centered at the distribution of $f(\boldsymbol{\xi})$ under F_0—a property we term *projection equivalence*:

$$\sup_{F \in \mathbb{B}_p(F_0, \varepsilon)} \rho^F(f(\boldsymbol{\xi})) = \sup_{G \in \mathcal{C}_p(f \mid F_0, \varepsilon)} \rho^G(X), \tag{2}$$

where $\mathcal{C}_p(f \mid F_0, \varepsilon)$ denotes the one-dimensional p-Wasserstein ball centered at G_0, the distribution of $f(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \sim F_0$. Thus the high-dimensional worst-case risk evaluation reduces exactly to its one-dimensional counterpart, yielding substantial benefits for both computation and analysis, and in many cases leading to closed-form or efficiently computable solutions.

Projection equivalence has thus far been observed only in limited settings, most notably when the loss function f is linear (Mao et al. (2022)). In this case, Wu et al. (2022) and Aolaritei et al. (2023) obtain projection equivalence by establishing a *set-level equivalence*: the projection of the high-dimensional Wasserstein ball through f coincides exactly with the one-dimensional Wasserstein ball $\mathcal{C}_p(f \mid F_0, \varepsilon)$. However, as noted in Wu et al. (2022) and Aolaritei et al. (2023), such set-level equivalence is generally impossible beyond the affine case. In more general settings, only a set-inclusion relationship can be established: the projection of the high-dimensional Wasserstein ball through f is contained in the one-dimensional Wasserstein ball $\mathcal{C}_p(f \mid F_0, \varepsilon)$ (see, e.g., Santambrogio (2015)). This in turn implies that the one-dimensional worst-case problem, i.e., the right-hand side of (2), provides at best an upper bound on the full-dimensional worst-case problem, i.e., the left-hand side of (2).

Set-level equivalence is, however, a stronger property than projection equivalence, and hence may be unnecessarily restrictive for reducing the worst-case risk problem. For the purpose of problem

reduction, what ultimately matters is whether the worst-case evaluations coincide, not whether the underlying ambiguity sets are identical. The linear case therefore represents only a narrow special instance, leaving open the fundamental question of when exact reduction is possible across broader classes of loss functions and risk functionals, even in the absence of set-level equivalence.

To the best of our knowledge, no prior work has provided a complete characterization of when projection equivalence holds beyond the linear setting. More generally, whether such an equivalence holds depends on the class of risk functionals used to evaluate risk. In this paper, we close this gap by developing a hierarchy of results: starting from the most general law-invariant risk functionals and then specializing to monotone functionals, coherent risk measures, and further subclasses, we derive necessary and sufficient conditions on the loss functions f under which projection equivalence (2) holds, thus providing a complete characterization. On the one hand, our results reveal that solving the high-dimensional worst-case risk problem via its one-dimensional counterpart is possible for classes of loss functions that extend far beyond the linear case. On the other hand, and perhaps even more theoretically intriguing, our results also constitute impossibility results: such a reduction is provably not possible for any loss function outside the identified classes. This establishes a sharp boundary for projection equivalence in Wasserstein DRO, delineating precisely when exact reduction is feasible and when it is not. As an application, we show how our reduction results enable an exact reformulation of Wasserstein chance-constrained programs, extending in a nontrivial way the previous results of Xie (2021) and Chen et al. (2024) from the type-1 Wasserstein setting to general type-p.

As another key benefit of reduction, our results show that projection equivalence enables the identification of broader conditions under which Wasserstein DRO problems admit an exact reformulation as regularized optimization problems (Pflug et al., 2012; Blanchet et al., 2019; Shafieezadeh-Abadeh et al., 2019; Gao et al., 2024; Wu et al., 2022). Such reformulations are of great interest in both optimization and machine learning, as they reveal when Wasserstein DRO can be interpreted and solved through regularization schemes commonly applied in practice. Previously, exact regularization reformulations were known only in restricted cases: in particular, when the risk functional is the expectation (Shafieezadeh-Abadeh et al., 2019; Gao et al., 2024), or more generally, for other risk functionals but limited to linear loss functions (Wu et al., 2022). Our results extend these findings substantially by pinpointing precisely when such reformulations exist across broader classes of risk functionals and loss functions.

We further extend the reduction result (2) to the classification setting, showing that exact reduction remains possible but only for a more restricted class of loss functions, whereas the existing

literature has primarily focused on linear classifiers (e.g., Kuhn et al. (2019), Ho-Nguyen and Wright (2023)).

Finally, to provide a high-level perspective on our key results—including set equivalence, projection equivalence, regularization, and the exact characterization of loss functions—we summarize their relationships in Figure 1.

Our contributions. This paper makes the following advances:

1. **Complete characterization.** We provide necessary and sufficient conditions under which the high-dimensional worst-case risk problem reduces exactly to its one-dimensional counterpart, offering the first full characterization of projection equivalence in Wasserstein DRO.

2. **Beyond set preservation.** We show that projection equivalence can hold even without set-level equivalence, revealing a broader class of (f, ρ) pairs than previously recognized and establishing sharp impossibility boundaries beyond them.

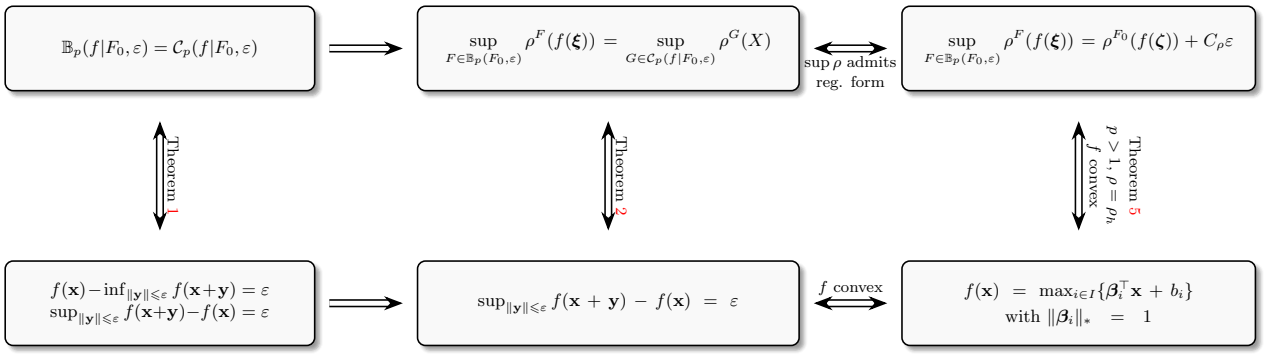
3. **Functional characterization.** For convex losses, we identify the exact family of functions admitting projection equivalence via convex isometric subgradients, subsuming affine and piecewise-linear forms as special cases.

4. **Regularization reformulations.** Leveraging projection equivalence, we derive precise conditions under which Wasserstein DRO admits exact regularization representations, substantially extending prior results beyond expectation and linear losses.

2 Preliminaries

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an atomless probability space. A random vector $\boldsymbol{\xi}$ is a measurable mapping from Ω to $\mathbb{R}^n$, $n \in \mathbb{N}$. Denote by $F_{\boldsymbol{\xi}}$ the distribution of $\boldsymbol{\xi}$ under $\mathbb{P}$. Denote by $\mathcal{M}(\mathbb{R}^n)$ the the set of all distributions on $\mathbb{R}^n$. For $p \geqslant 1$, let $L^p := L^p(\Omega, \mathcal{A}, \mathbb{P})$ be the set of all random variables with finite pth moment and $\mathcal{M}_p(\mathbb{R}^n)$ be the set of all distributions on $\mathbb{R}^n$ with finite pth moment in each component. For any norm $\|\cdot\|$ on $\mathbb{R}^n$, its dual norm $\|\cdot\|_*$ is defined as $\|\mathbf{y}\|_* = \sup_{\|\mathbf{x}\|\leqslant 1} \mathbf{x}^\top \mathbf{y}$. Let q denote the Hölder conjugate of p, i.e., $1/p + 1/q = 1$. For a real number $x \in \mathbb{R}$, we use $x_+ = \max\{x, 0\}$ and $x_- = \max\{-x, 0\}$; and for $m \in \mathbb{N}$, denote by $[m] = \{1, \ldots, m\}$. Let $\mathbf{e}_i \in \mathbb{R}^n$ be the vector whose ith element is 1 and all other elements are 0 for $i \in [n]$. Let $\mathbf{x}_{-i}$ denote the vector obtained by removing the i-th component from $\mathbf{x} \in \mathbb{R}^n$. Similarly, $\mathbf{x}_{-(i,j)}$ denote the vector obtained by removing the i-th and j-th components. Denote by $\delta_{\mathbf{z}}$ the Dirac distribution at $\mathbf{z} \in \mathbb{R}^n$.

Figure 1: Illustration of the relationships in Theorems 1 and 2



Notes. The three conditions in the first layer are assumed to hold for all $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and all $\varepsilon > 0$, while the first two conditions in the second layer are assumed to hold for all $\mathbf{x} \in \mathbb{R}^n$ and all $\varepsilon > 0$. We assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz function. For simplicity, we set $\mathrm{Lip}(f) = c_f = 1$. The equivalence stated in Theorem 2 holds under the assumption that $\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \rho^F(f(\boldsymbol{\xi})) = \sup_{G\in\mathcal{C}_p(f|F_0,\varepsilon)} \rho^G(X)$ holds for any monotone risk measure ρ. We say that "$\sup\rho$ admits reg. form" if there exists a constant $C_\rho \in \mathbb{R}$ such that $\sup_{G\in\mathcal{C}_p(f|F_0,\varepsilon)} \rho^G(X) = \rho^{F_0}(f(\boldsymbol{\zeta})) + C_\rho\varepsilon$. Theorem 5 shows that, when $p > 1$, $\rho = \rho_h$ is a convex distortion risk measure, and f is convex, the two conclusions on the right are equivalent, with $C_\rho = \|h'\|_q$, where q is the Hölder conjugate of p.

We denote by $\mathbf{x} \circ \mathbf{y}$ the Hadamard (element-wise) product of vectors $\mathbf{x}$ and $\mathbf{y}$, i.e., the vector whose i-th component is given by $(\mathbf{x} \circ \mathbf{y})_i = x_i y_i$.

For any two n-dimensional distributions F_1 and F_2, the type-p Wasserstein metric is defined as

$$\mathrm{W}_p\left(F_1, F_2\right) := \inf_{\pi\in\Pi(F_1,F_2)} \left(\mathbb{E}^\pi[\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|^p]\right)^{1/p}, \tag{3}$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^n$, and $\Pi(F_1, F_2)$ denotes the set of all distributions on $\mathbb{R}^n \times \mathbb{R}^n$ with marginals F_1 and F_2.

We define the ball of distributions $\mathbb{B}_p(F_0, \varepsilon)$ on $\mathbb{R}^n$ as

$$\mathbb{B}_p(F_0, \varepsilon) = \{F \in \mathcal{M}_p(\mathbb{R}^n) : \mathrm{W}_p(F, F_0) \leqslant \varepsilon\}, \tag{4}$$

and refer to it as the type-p Wasserstein ball throughout this paper.

We begin by formalizing the notion of a *projection-induced ambiguity set*, along with the classes of risk functionals under which our main results are developed. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a loss function, $F_0 \in \mathcal{M}(\mathbb{R}^n)$ be a nominal distribution, and $\varepsilon > 0$. The *projection-induced p-Wasserstein ball* is defined as

$$\mathbb{B}_p(f \,|\, F_0, \varepsilon) := \left\{ F_{f(\boldsymbol{\xi})} : F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon) \right\},$$

where $F_{f(\boldsymbol{\xi})}$ denotes the distribution of the scalar random variable $f(\boldsymbol{\xi})$ for $\boldsymbol{\xi} \sim F_{\boldsymbol{\xi}}$. We also write

$\mathcal{C}_p(f \,|\, F_0, \varepsilon) := \mathbb{B}_p(G_0, \varepsilon) \subseteq \mathcal{M}(\mathbb{R})$ for the *one-dimensional* type-p Wasserstein ball centered at G_0, the distribution of $f(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \sim F_0$. If f is Lipschitz continuous, we define its Lipschitz constant with respect to the norm used in the Wasserstein metric as

$$\mathrm{Lip}(f) := \sup_{\mathbf{x} \neq \mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|}.$$

Let $\mathcal{X}$ denote a space of real-valued random variables. We refer to any functional $\rho : \mathcal{X} \to \mathbb{R}$ as a *risk functional*, and consider the following nested classes:

(i) Law-invariant finite-valued risk functionals. A risk functional ρ is *law-invariant* if $\rho(X) = \rho(Y)$ for all $X, Y \in \mathcal{X}$ such that $X \overset{d}{=} Y$. We assume that $\rho(X) < \infty$ for all $X \in \mathcal{X}$.

(ii) Law-invariant monotone risk functionals. A risk functional ρ is *monotone* if $\rho(X) \leqslant \rho(Y)$ whenever $X \leqslant Y$ almost surely.

(iii) Law-invariant coherent risk measures. A risk functional ρ is *coherent* if it is monotone and satisfies the following properties:

- **Translation invariance:** $\rho(X + m) = \rho(X) + m$ for all $m \in \mathbb{R}$,
- **Positive homogeneity:** $\rho(\lambda X) = \lambda \rho(X)$ for all $\lambda \geqslant 0$,
- **Subadditivity:** $\rho(X + Y) \leqslant \rho(X) + \rho(Y)$.

It is well known that any law-invariant, coherent, and lower semicontinuous risk measure $\rho : L^p \to \mathbb{R}$ admits a Kusuoka representation (Kusuoka, 2001; Filipovic and Svindland, 2007; Shapiro, 2013) of the form

$$\rho^F(X) = \sup_{\mu \in \mathscr{M}_\rho} \int_0^1 \mathrm{CVaR}_\alpha^F(X) \, \mathrm{d}\mu(\alpha), \tag{5}$$

where $\mathscr{M}_\rho$ is a set of probability measures on $[0, 1]$, and $\mathrm{CVaR}_\alpha^F(X)$ is the Conditional Value-at-Risk (CVaR, also called Expected Shortfall, ES) at level $\alpha \in [0, 1]$ defined as

$$\mathrm{CVaR}_\alpha^F(X) = \frac{1}{1-\alpha} \int_\alpha^1 F^{-1}(s) \mathrm{d}s, \ \alpha \in [0, 1), \quad \text{and} \quad \mathrm{CVaR}_1^F(X) = F^{-1}(1).$$

In this paper, we call a coherent risk measure *regular* if it admits a Kusuoka representation of the form (5) with

$$C_\rho := \sup_{\mu \in \mathscr{M}_\rho} \int_0^1 \frac{1}{1-\alpha} \, \mathrm{d}\mu(\alpha) < \infty.$$

We denote by $\mathcal{R}_{\text{law}}$, $\mathcal{R}_{\text{mon}}$, and $\mathcal{R}_{\text{coh}}$ the respective classes of law-invariant finite-valued risk functionals, monotone risk functionals, and regular coherent risk measures, where

$$\mathcal{R}_{\text{coh}} \subseteq \mathcal{R}_{\text{mon}} \subseteq \mathcal{R}_{\text{law}}.$$

3 Projection Equivalence in Wasserstein DRO

We now present our main results: a complete characterization of the loss functions f for which projection equivalence (2) holds. Our analysis proceeds hierarchically, beginning with the most general class of risk functionals, $\mathcal{R}_{\text{law}}$, and moving to the more structured classes of monotone ($\mathcal{R}_{\text{mon}}$) and coherent ($\mathcal{R}_{\text{coh}}$) risk measures. This progression reveals progressively broader classes of loss functions that admit projection equivalence.

3.1 Case $\mathcal{R}_{\text{law}}$

When projection equivalence is required to hold across the most general class of risk functionals, it coincides with set equivalence. Theorem 1 gives the necessary and sufficient conditions for this equivalence and a complete characterization of the admissible loss functions.

Theorem 1. *For $p \geqslant 1$, let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. Then the following statements are equivalent.*

(i) There exists $c_f \geqslant 0$ such that

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \rho^F\left(f(\boldsymbol{\xi})\right) = \sup_{G \in \mathcal{C}_p(f|F_0,c_f\varepsilon)} \rho^G(X) \tag{6}$$

holds for any $\rho \in \mathcal{R}_{\text{law}}$, $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$.

(ii) There exists $c_f \geqslant 0$ such that for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$, it holds that

$$\mathbb{B}_p(f|F_0,\varepsilon) = \mathcal{C}_p(f|F_0,c_f\varepsilon). \tag{7}$$

(iii) The function f is Lipschitz continuous, and

$$f(\mathbf{x}) - \inf_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x}+\mathbf{y}) = \sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = \text{Lip}(f)\varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, \varepsilon > 0. \tag{8}$$

While the set-equivalence condition (ii) is sufficient—since $(ii) \Rightarrow (i)$ follows directly from the definition—establishing its necessity, i.e., $(i) \Rightarrow (ii)$, is considerably more delicate. As shown later,

this necessity no longer holds once we move to more refined classes of risk functionals, where projection equivalence may arise even without set-level coincidence. Furthermore, whereas set-level equivalence (7) was previously known only for linear loss functions f, characterization (*iii*) reveals that linearity is not required.

To shed light on when linearity is essential and when it is not, we provide a refined characterization of functions with $\mathrm{Lip}(f) > 0$ that satisfy (8) under common norms, including $\|\cdot\|_1$ and $\|\cdot\|_a$ for $a \in (1, \infty)$. In particular, we first show that (8) can be sharpened under a strictly convex norm (Clarkson, 1936), i.e., a norm such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ and $\|\mathbf{x} - \mathbf{y}\| \neq 0$, one has $\|\mathbf{x} + \mathbf{y}\| < 2$. It is well known that $\|\cdot\|_a$ with $a \in (1, \infty)$ is strictly convex. Under these norms, linearity of the loss function f becomes necessary.

Proposition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with $\mathrm{Lip}(f) > 0$. If $\|\cdot\|$ is a strictly convex norm, then f satisfies (8) is equivalent to there exists $\boldsymbol{v}$ with $\|\boldsymbol{v}\| = 1$ such that*

$$f(\mathbf{x} + t\boldsymbol{v}) - f(\mathbf{x}) = \mathrm{Lip}(f)t, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, t \in \mathbb{R}. \tag{9}$$

In particular, if $\|\cdot\| = \|\cdot\|_a$, $a \in (1, \infty)$, then f satisfies (9) if and only if $f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + b$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

We next show that (8) can also be sharpened under $\|\cdot\|_1$, where a significantly richer class of admissible loss functions emerges.

Proposition 2. *If $\|\cdot\| = \|\cdot\|_1$ is the ℓ_1-norm, then for any $\mathbf{x} \in \mathbb{R}^n$, there exist $i, j \in [n]$ such that*

$$f(\mathbf{x} + t\tilde{\mathbf{e}}_i) - f(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x} + t\tilde{\mathbf{e}}_j) = \mathrm{Lip}(f)t, \quad \forall\, t > 0, \tag{10}$$

where $\tilde{\mathbf{e}}_i \in \{\pm \mathbf{e}_i\}, i \in [n]$.

Corollary 1. *Given $\|\cdot\| = \|\cdot\|_1$, if f is given by the following two cases, then f satisfies (10).*

(a) $f(\mathbf{x}) = c\left(\boldsymbol{\beta}^\top \mathbf{x} + g(\boldsymbol{\eta} \circ \mathbf{x})\right)$ where $c := \mathrm{Lip}(f)$, $\|\boldsymbol{\beta}\|_\infty = 1$, $\beta_i, \eta_i \in \{1, -1, 0\}$, $i \in [n]$, $\boldsymbol{\beta} \circ \boldsymbol{\eta} = \mathbf{0}$, and g is a Lipschitz function with $\mathrm{Lip}(g) \leqslant 1$.

(b) $f(\mathbf{x}) = c\left(|\boldsymbol{\beta}^\top \mathbf{x}| - |\boldsymbol{\nu}^\top \mathbf{x}| + g(\boldsymbol{\eta} \circ \mathbf{x})\right)$, where $c := \mathrm{Lip}(f)$, $\|\boldsymbol{\beta}\|_\infty = \|\boldsymbol{\nu}\|_\infty = 1$, $\beta_i, \nu_i, \eta_i \in \{1, -1, 0\}$, $i \in [n]$, $\boldsymbol{\beta} \circ \boldsymbol{\eta} = \boldsymbol{\nu} \circ \boldsymbol{\eta} = \boldsymbol{\beta} \circ \boldsymbol{\nu} = \mathbf{0}$, and g is a Lipschitz function with $\mathrm{Lip}(g) \leqslant 1$.

In other words, such functions may consist of a linear term, or the difference of two absolute-value terms depending on disjoint subsets of the coordinates of $\mathbf{x}$, together with a nontrivial Lipschitz component acting on the complementary coordinates.

3.2 Case $\mathcal{R}_{\mathbf{mon}}$

We next show that, when projection equivalence is required to hold across all monotone risk measures, the stringent set-level equivalence (7) is no longer necessary. Instead, a weaker condition on the loss function f suffices—one that admits a substantially richer class of functions than in the previous setting.

Theorem 2. *For $p \geqslant 1$, let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. The following statements are equivalent.*

(i) There exists $c_f \geqslant 0$ such that

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \rho^F\left(f(\boldsymbol{\xi})\right) = \sup_{G \in \mathcal{C}_p(f|F_0, c_f\varepsilon)} \rho^G(X) \tag{11}$$

holds for any $\rho \in \mathcal{R}_{\mathrm{mon}}$, $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$.

(ii) There exists $c_f \geqslant 0$ such that

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = c_f\varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, \varepsilon > 0. \tag{12}$$

When f is convex, condition (12) can be further sharpened and, in fact, admits a closed-form characterization, yielding a complete representation of all functions that satisfy it.

Proposition 3. *If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $p \geqslant 1$, then the function in (12) must admit the form of $c_f \geqslant 0$, $\boldsymbol{\beta}_i \in \mathbb{R}^n$, $i \in I$, with $\|\boldsymbol{\beta}_i\|_* = 1$ and $b_i \in \mathbb{R}$ such that*

$$f(\mathbf{x}) = \max_{i \in I}\{c_f \boldsymbol{\beta}_i^\top \mathbf{x} + b_i\}. \tag{13}$$

We refer to any function f of the form (13) with $c_f = 1$ as a *convex isometric subgradient function*. Such a function satisfies

$$f(\mathbf{x}) = \sup_{\|\mathbf{y}\|_* = 1} \{\mathbf{x}^\top \mathbf{y} - f^*(\mathbf{y})\},$$ [1]

[1] One can verify that a function f is a convex isometric subgradient function if and only if f satisfies $f(\mathbf{x}) = \sup_{\|\mathbf{y}\|_*=1}\{\mathbf{x}^\top\mathbf{y} - f^*(\mathbf{y})\}$. Suppose $f(\mathbf{x}) = \sup_{\|\mathbf{y}\|_*=1}\{\mathbf{x}^\top\mathbf{y} - f^*(\mathbf{y})\}$. Let $I := \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\|_* = 1\}$, and define $\boldsymbol{\beta}_\mathbf{y} := \mathbf{y}$ and $b_\mathbf{y} := -f^*(\mathbf{y})$ for $\mathbf{y} \in I$. Then f can be written as $f(\mathbf{x}) = \sup_{\mathbf{y}\in I}\{\mathbf{x}^\top\boldsymbol{\beta}_\mathbf{y} + b_\mathbf{y}\}$, which is a desired form. Conversely, suppose f take the form of (13). We claim that for any $\mathbf{x} \in \mathbb{R}^n$ and $\varepsilon > 0$, the following holds: $\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = \varepsilon$. Indeed, for any $\|\mathbf{y}\| \leqslant \varepsilon$, we have $f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) \leqslant \sup_i \mathbf{y}^\top\boldsymbol{\beta}_i \leqslant \varepsilon$ by Hölder's inequality. On the other hand, for each $\mathbf{x} \in \mathbb{R}^n$, there exists $i_0 \in I$ such that $f(\mathbf{x}) = \boldsymbol{\beta}_{i_0}^\top\mathbf{x} + b_{i_0}$ and since $\|\boldsymbol{\beta}_{i_0}\|_* = 1$, there exists $\mathbf{z}_0$ with $\|\mathbf{z}_0\| = 1$ such that $\boldsymbol{\beta}_{i_0}^\top\mathbf{z}_0 = 1$. Setting $\mathbf{y}_0 := \varepsilon\mathbf{z}_0$ gives $f(\mathbf{x}+\mathbf{y}_0) - f(\mathbf{x}) \geqslant \varepsilon$. Thus, the claim is established. It follows that for any $\mathbf{x} \in \mathbb{R}^n$, the subgradient of f at $\mathbf{x}$, denoted by $\nabla f(\mathbf{x})$, satisfies $\|\nabla f\|_* = 1$. Therefore, since all subgradients lie on $\|\mathbf{y}\|_* = 1$, the Fenchel Moreau theorem gives $f(\mathbf{x}) = \max_{\mathbf{y}\in\mathrm{dom} f^*}\{\mathbf{x}^\top\mathbf{y} - f^*(\mathbf{y})\} = \max_{\|\mathbf{y}\|_*=1}\{\mathbf{x}^\top\mathbf{y} - f^*(\mathbf{y})\}$.

where f^* denotes the conjugate of f.

Example 1. The following are concrete instances of convex isometric subgradient functions:

(i) The norm function: $f(\mathbf{x}) = \|\mathbf{x}\| = \sup_{\|\boldsymbol{\beta}\|_*=1} \boldsymbol{\beta}^\top \mathbf{x}$. We have $c_f = 1$ and by Theorem 2, for any monotone risk measure ρ, $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$ it holds that

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \rho^F\left(\|\boldsymbol{\xi}\|\right) = \sup_{G \in \mathbb{B}_p(F_{\|\boldsymbol{\zeta}\|},\varepsilon)} \rho^G(X).$$

(ii) The absolute value linear function: $f(\mathbf{x}) = |\boldsymbol{\beta}^\top \mathbf{x} + b|$, $\boldsymbol{\beta} \in \mathbb{R}^n$, $b \in \mathbb{R}$. We have $c_f = \|\boldsymbol{\beta}\|_*$ and by Theorem 2, for any monotone risk measure ρ, $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$ it holds that

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \rho^F\left(|\boldsymbol{\beta}^\top \boldsymbol{\xi} + b|\right) = \sup_{G \in \mathbb{B}_p(F_{|\boldsymbol{\beta}^\top \boldsymbol{\zeta}+b|},\|\boldsymbol{\beta}\|_* \varepsilon)} \rho^G(X).$$

(iii) If $\|\cdot\| = \|\cdot\|_1$ is the ℓ_1-norm, then $f(\mathbf{x}) = \max_i(\beta_{i1}x_1 + \cdots + \beta_{in}x_n + b_i)$ with $\max_{j\in[n]} |\beta_{ij}| = c_f$ for $i \in I$, is a convex isometric subgradient function. In particular, $f(\mathbf{x}) = \max_i(\beta_{i1}x_1 + \cdots + \beta_{in-1}x_{n-1} + b_i) + x_n$ with $\max_{i\in I} \max_{j\in[n-1]} |\beta_{ij}| \leqslant 1$ is a special case.

Remark 1. It is worth noting that Proposition 3 holds even if we do not assume that f is convex when $n = 1$, while the convexity is necessary when $n \geqslant 2$. We give a counter example here. Consider $\mathbb{R}^2$ and the norm is given by $\|(x_1, x_2)\| = \sqrt{x_1^2 + x_2^2}$, $(x_1, x_2) \in \mathbb{R}^2$. Define

$$f(x_1, x_2) = \begin{cases} \max\left\{\frac{-x_1+2x_2}{\sqrt{5}}, \frac{-x_1-2x_2}{\sqrt{5}}\right\}, & x_1 \geqslant 0, \\ \max\left\{\frac{x_1+2x_2}{\sqrt{5}}, \frac{x_1-2x_2}{\sqrt{5}}\right\}, & x_1 < 0. \end{cases}$$

One can check the function f satisfies (12) but f is not convex on the line $x_2 = 0$ and thus could not be written as (13).

One may wonder whether the characterization (12) can be further relaxed when projection equivalence is required to hold only across a more restrictive class of risk measures. We show that, perhaps surprisingly, (12) remains necessary even in as specific a case as the widely used Value-at-Risk. For a random variable X with distribution F, the Value-at-Risk (VaR) at level $\alpha \in [0, 1]$ is defined as

$$\mathrm{VaR}_0^F(X) = \inf\{x : F(x) > 0\} \quad \text{and} \quad \mathrm{VaR}_\alpha^F(X) = \inf\{x : F(x) \geqslant \alpha\}, \ \alpha \in (0, 1].$$

Proposition 4. *Fix $\alpha \in [0,1)$ and $p \geqslant 1$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. There exists $c_f \geqslant 0$ such that*

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \mathrm{VaR}_\alpha^F\left(f(\boldsymbol{\xi})\right) = \sup_{G \in \mathcal{C}_p(f|F_0,c_f\varepsilon)} \mathrm{VaR}_\alpha^G(X) \tag{14}$$

holds for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$ if and only if f satisfies (12).

Remark 2. It is worth noting that Proposition 4 does not hold for $\alpha = 1$. Note that for any f and $p \in [1,\infty)$, $\sup_{G \in \mathcal{C}_p(f|F_0,c_f\varepsilon)} \mathrm{VaR}_1^G(X) = \infty$, and if, moreover, f satisfies $\sup_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = \infty$, then $\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \mathrm{VaR}_1^F\left(f(\boldsymbol{\xi})\right) = \infty$. This implies that if f is unbounded from above, then (14) always holds for $\alpha = 1$. Therefore, we have Proposition 4 does not hold for $\alpha = 1$.

Before proceeding further, we highlight how Proposition 4 facilitates the solution of important OR/MS problems. Observe that the function $f(\boldsymbol{\xi}) := \max_{i\in I} \frac{\boldsymbol{\beta}_i^\top \boldsymbol{\xi}}{\|\boldsymbol{\beta}_i\|_*}$ is a special case of (13). By Proposition 4, this yields

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \mathrm{VaR}_\alpha^F\left(\max_{i\in I} \frac{\boldsymbol{\beta}_i^\top \boldsymbol{\xi}}{\|\boldsymbol{\beta}_i\|_*}\right) = \sup_{G \in \mathbb{B}_p(G_0,\varepsilon)} \mathrm{VaR}_\alpha^G(X), \tag{15}$$

where G_0 is the distribution of $f(\boldsymbol{\zeta})$ and $\boldsymbol{\zeta} \sim F_0$.

We present two important applications of (15). In each case, the reduction to the one-dimensional worst-case Value-at-Risk problem (the right-hand side of (15)) allows us to invoke the CVaR reformulation in Lemma 1 to solve the original high-dimensional problem (the left-hand side of (15)).

Lemma 1. *The worst-case value-at-risk (15) is the unique $x \in \mathbb{R}$ satisfying*

$$\frac{\varepsilon^p}{1-\alpha} + \mathrm{CVaR}_\alpha^{F_0}\left(-\left(x - \max_{i\in I} \frac{\boldsymbol{\beta}_i^\top \boldsymbol{\zeta}}{\|\boldsymbol{\beta}_i\|_*}\right)_+^p\right) = 0. \tag{16}$$

Moreover, the value of (15) is strictly increasing in $\alpha \in [0,1]$.

Example 2 (Worst-case risk over multiple portfolios)**.** Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \in \mathbb{R}^n$ denote m portfolio weight vectors with no short positions, and assume that all of them share the same dual norm, i.e., $\|\boldsymbol{\beta}_j\|_* = \|\boldsymbol{\beta}_1\|_*$ for all $j = 1, \ldots, m$. The problem of evaluating the worst-case value-at-risk of the poorest-performing portfolio is

$$\sup_{F \in \mathbb{B}_p(F_0,\varepsilon)} \mathrm{VaR}_\alpha^F\left(\max_{i\in[m]} \boldsymbol{\beta}_i^\top \boldsymbol{\xi}\right).$$

By applying (15) together with Lemma 1, the problem can be reformulated as the following opti-

mization problem:

$$\min\ x \quad \text{s.t.} \quad \frac{\varepsilon^p}{1-\alpha}\|\boldsymbol{\beta}_1\|_*^p + \mathrm{CVaR}_\alpha^{F_0}\left(-\left[x - \max_{i\in[m]} \boldsymbol{\beta}_i^\top \boldsymbol{\zeta}\right]_+^p\right) \leqslant 0. \tag{17}$$

This formulation generalizes the result of Chen and Xie (2021), which is restricted to the case $p = 1$, to all $p \geqslant 1$.

Example 3 (Distributionally robust chance-constrainted program)**.** One important application of Wasserstein DRO is to distributionally robust chance-constrained programs (Wasserstein DRCCPs), which aim to ensure that constraints hold with high probability under distributional uncertainty. A typical Wasserstein DRCCP takes the form

$$\min_{\mathbf{x}\in S} \quad \mathbf{c}^\top \mathbf{x}, \tag{18}$$

$$\text{s.t.} \quad \inf_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathbb{P}^F\left(a_i(\mathbf{x})^\top \boldsymbol{\xi} \leqslant b_i(\mathbf{x}),\ \forall\ i \in [m]\right) \geqslant 1-\eta, \tag{19}$$

where $S \subseteq \mathbb{R}^k$ is a feasible set of the decision vector $\mathbf{x}$, the vector $\mathbf{c} \in \mathbb{R}^k$ denotes the objective function coefficients, and $S(\mathbf{x}) = \{\boldsymbol{\xi} : a_i(\mathbf{x})^\top \boldsymbol{\xi} \leqslant b_i(\mathbf{x}),\ \forall\ i \in [m]\} \subseteq \mathbb{R}^n$ is referred to as a safety set for each $\mathbf{x} \in S$. For $p = 1$, the feasible set has been reformulated in terms of CVaR by Xie (2021) and Chen et al. (2024). Here, we provide a new perspective and extend the result to general $p \geqslant 1$. First, the constraint (19) is equivalent to [2]

$$\sup_{F\in\mathbb{B}_p(F_0,\delta)} \mathrm{VaR}_{1-\eta}^F\left(\max_{i\in[m]}\left\{\frac{a_i(\mathbf{x})^\top \boldsymbol{\xi} - b_i(\mathbf{x})}{\|a_i(\mathbf{x})\|_*}\right\}\right) \leqslant 0.$$

By applying (15) together with Lemma 1, the problem (18) is equivalent to

$$\min_{\mathbf{x}\in S} \mathbf{c}^\top \mathbf{x}, \quad \text{s.t.} \quad a_i(\mathbf{x}) \neq 0,\ -\mathrm{CVaR}_{1-\eta}^{F_0}\left[-f(\mathbf{x},\boldsymbol{\zeta})^p\right] \geqslant \frac{\varepsilon^p}{\eta}, \quad i \in I(\mathbf{x}), \tag{20}$$
$$\text{or} \quad a_i(\mathbf{x}) = 0,\ b_i(\mathbf{x}) \geqslant 0, \quad i \notin I(\mathbf{x}),$$

where $f(\mathbf{x},\boldsymbol{\zeta}) = \min_{i\in I(\mathbf{x})} \frac{(b_i(\mathbf{x}) - a_i(\mathbf{x})^\top \boldsymbol{\zeta})_+}{\|a_i(\mathbf{x})\|_*}$ and $I(\mathbf{x}) = \{i \in [m] : a_i(\mathbf{x}) \neq 0\}$.

Moreover, observing that

$$\min_{i\in I(\mathbf{x})} -\frac{1}{\|a_i(\mathbf{x})\|_*^p}\mathrm{CVaR}_{1-\eta}^{F_0}\left[-(b_i(\mathbf{x}) - a_i(\mathbf{x})^\top \boldsymbol{\xi})_+^p\right] \geqslant -\mathrm{CVaR}_{1-\eta}^{F_0}\left[-f(\mathbf{x},\boldsymbol{\zeta})^p\right].$$

[2] For a random variable X with distribution F and $\eta \in (0,1)$, it holds that $F(0) \geqslant 1-\eta$ if and only if $\mathrm{VaR}_{1-\eta}^F(X) \leqslant 0$. To see this, first assume that $F(0) \geqslant 1-\eta$. By definition of VaR, we have $\mathrm{VaR}_{1-\eta}(X) = \inf\{x : F(x) \geqslant 1-\eta\} \leqslant 0$ as $0 \in \{x : F(x) \geqslant 1-\eta\}$. Next assume that $\mathrm{VaR}_{1-\eta}(X) \leqslant 0$, that is, $\inf\{x : F(x) \geqslant 1-\eta\} \leqslant 0$. There exists $x \downarrow 0$ such that $F(x) \geqslant 1-\eta$. By right-continuity of F, we have $F(0) \geqslant 1-\eta$.

we obtain the following tractable optimization problem, which provides a conservative approximation of (18):

$$\min_{\mathbf{x}\in S} \mathbf{c}^\top \mathbf{x}, \quad \text{s.t.} \quad a_i(\mathbf{x}) \neq 0, \ \min_{i\in I(\mathbf{x})} -\frac{1}{\|a_i(\mathbf{x})\|_*^p} \mathrm{CVaR}_{1-\eta}^{F_0}\left[-(b_i(\mathbf{x}) - a_i(\mathbf{x})^\top \xi_i)_+^p\right] \geqslant \frac{\varepsilon^p}{\eta} \quad i \in I(\mathbf{x}),$$
$$\text{or} \quad a_i(\mathbf{x}) = 0, \ b_i(\mathbf{x}) \geqslant 0, \quad i \notin I(\mathbf{x}),$$

This upper-bound formulation generalizes the result of Xie (2021), which is restricted to $p = 1$, to arbitrary $p \geqslant 1$.

3.3 Case $\mathcal{R}_{\mathbf{coh}}$ under Wasserstein ($p = 1$)

As our final general characterization of projection equivalence, we show that projection equivalence can hold for a significantly broader class of loss functions f when restricted to the Wasserstein ball with $p = 1$ and to coherent risk measures admitting a Kusuoka representation.

Theorem 3. *Let $\rho \in \mathcal{R}_{\mathrm{coh}}$ and $f : \mathbb{R}^n \to \mathbb{R}$ be a function. We have the following statements hold.*

(i) If f is a Lipschitz continuous function and there exist $\mathbf{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^n$ with $\|\boldsymbol{v}_k\| = 1, k \in \mathbb{N}$ such that

$$\lim_{k\to\infty} \limsup_{m\to\infty} \frac{1}{m}(f(\mathbf{x}_0 + m\boldsymbol{v}_k) - f(\mathbf{x}_0)) = \mathrm{Lip}(f), \tag{21}$$

then we have

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \rho^F\left(f\left(\boldsymbol{\xi}\right)\right) = \sup_{G\in\mathcal{C}_1(f|F_0, c_f\varepsilon)} \rho^G(X) \tag{22}$$

holds for any $F_0 \in \mathcal{M}\left(\mathbb{R}^n\right)$ and $\varepsilon > 0$.

(ii) If there exists a compact set $\mathcal{D} \subseteq \mathbb{R}^n$ such that f coincides with some convex function on $\mathbb{R}^n \backslash \mathcal{D}$, then we have (21) is also a necessary for (22) to hold for any $F_0 \in \mathcal{M}\left(\mathbb{R}^n\right)$ and $\varepsilon > 0$.

(iii) If, moreover, f is a convex function, then there exists $c_f \geqslant 0$ such that (22) holds for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$ if and only if f is Lipschitz continuous.

4 Exact Regularization via Projection Equivalence

Projection equivalence greatly simplifies worst-case risk problems; in particular, its one dimensional reduction lets us pinpoint exactly when these problems admit a regularization reformulation—a connection of both theoretical and practical significance. Building on the previous section's

results and focusing on several prominent families of risk functionals, we characterize the loss functions f for which worst-case risk problems in Wasserstein DRO admit exact reformulations as regularized empirical risk problems. As a byproduct, we show that for higher orders $p > 1$ the condition (12) is often necessary for both projection equivalence and the existence of such regularization, even within specific families of risk functionals. By contrast, for $p = 1$ (Wasserstein-1), exact regularization holds for a substantially larger class of loss functions. We begin with the higher-order case $p > 1$.

4.1 Higher-Order Case ($p > 1$)

Higher-order risk functionals

As a first higher-order risk functional, we study the worst-case L_p risk

$$\sup_{F \in \mathbb{B}_p(F_0, \varepsilon)} \left(\mathbb{E}^F \left[f^p\left(\boldsymbol{\xi}\right)\right]\right)^{1/p},$$

where $f : \mathbb{R}^n \to \mathbb{R}_+$. This formulation generalizes the classical worst-case expectation ($p = 1$) and underpins many important higher-order risk functionals, making it a natural basis for our analysis. Below, we present an exact characterization of the loss functions f that are necessary and sufficient for the existence of an exact regularization reformulation.

Theorem 4. *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be a Lipschitz continuous and convex function. For any $p \in (1, \infty)$, there exists $c_f \geqslant 0$ such that*

$$\sup_{F \in \mathbb{B}_p(F_0, \varepsilon)} \left(\mathbb{E}^F \left[f^p\left(\boldsymbol{\xi}\right)\right]\right)^{1/p} = \sup_{G \in \mathcal{C}_p(f|F_0, c_f\varepsilon)} \left(\mathbb{E}^G \left[|X|^p\right]\right)^{1/p} = \left(\mathbb{E}^{F_0} \left[f^p\left(\boldsymbol{\zeta}\right)\right]\right)^{1/p} + c_f\varepsilon \tag{23}$$

holds for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$ if and only if there exist $c_f \geqslant 0$, $\boldsymbol{\beta}_i \in \mathbb{R}^n$, $i \in I$, with $\|\boldsymbol{\beta}_i\|_* = 1$ and $b_i \in \mathbb{R}$ such that

$$f(\mathbf{x}) = \left(\max_{i \in I} \{c_f \boldsymbol{\beta}_i^\top \mathbf{x} + b_i\}\right)_+. \tag{24}$$

The functional form (24) is a slight generalization of (12) from the projection equivalence result (Theorem 2), underscoring the close connection between projection equivalence and regularization reformulation.

We next consider risk functionals of the form

$$\mathcal{D}_p^F(X) = \inf_{t \in \mathbb{R}} \left(\mathbb{E}^F \left[\ell^p(X, t)\right]\right)^{1/p} \quad \text{and} \quad \mathcal{H}_p^F(X) = \inf_{t \in \mathbb{R}} \left\{t + \left(\mathbb{E}^F \left[\ell^p(X, t)\right]\right)^{1/p}\right\} \tag{25}$$

for some loss function ℓ. These two forms cover many widely used risk measures, including variance and other higher-moment measures. By Theorem 4, each admits an exact regularization counterpart.

Corollary 2. *For any $p \in (1,\infty)$, $c > 0$, and $f(\boldsymbol{\xi}) = \max_{i\in I_1}\{\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\}$ with $\boldsymbol{\beta}_i \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}_i\|_* = 1$ and $b_i \in \mathbb{R}$, $i \in I_1$, let $\mathcal{D}_p$ and $\mathcal{H}_p$ be defined by (25) with $\ell(z,t) = c(z + \max_{j\in I_2}\{d_j t\})_+$ for some $d_j \in \mathbb{R}$, $j \in I_2$. Then we have the following statements hold.*

(i) If $\min_{j\in I_2} d_j \leqslant 0 \leqslant \max_{j\in I_2} d_j$, then

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathcal{D}_p^F\left(f(\boldsymbol{\xi})\right) = \mathcal{D}_p^{F_0}\left(f(\boldsymbol{\zeta})\right) + c\varepsilon. \tag{26}$$

If $\min_{j\in I_2} d_j > 0$ or $\max_{j\in I_2} d_j < 0$, then we have $\mathcal{D}_p^F\left(f(\boldsymbol{\xi})\right) = 0$ for any $F \in \mathcal{M}_p\left(\mathbb{R}^n\right)$.

(ii) If $c\min_{j\in I_2} d_j \leqslant -1$, then

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathcal{H}_p^F\left(f(\boldsymbol{\xi})\right) = \mathcal{H}_p^{F_0}\left(f(\boldsymbol{\zeta})\right) + c\varepsilon. \tag{27}$$

If $c\min_{j\in I_2} d_j > -1$, then $\mathcal{H}_p^F\left(\boldsymbol{\xi}\right) = -\infty$ for any $F \in \mathcal{M}_p\left(\mathbb{R}^n\right)$.

Corollary 3. *For any $p \in (1,\infty)$, $c > 0$ and $f(\boldsymbol{\xi}) := \boldsymbol{\beta}^\top \boldsymbol{\xi} + b$ with $\boldsymbol{\beta} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, let $\mathcal{D}_p$ and $\mathcal{H}_p$ be defined by (25) with $\ell(z,t) = c(\max\{z - d_1 t, -z - d_2 t\} + d_3)_+$, $j \in [3]$. We have the following statements hold.*

(i) If $\min\{d_1,d_2\} \leqslant 0 \leqslant \max\{d_1,d_2\}$, then (26) holds. If $\min\{d_1,d_2\} > 0$ or $\max\{d_1,d_2\} < 0$, we have $\mathcal{D}_p^F\left(f(\boldsymbol{\xi})\right) = 0$ for any $F \in \mathcal{M}_p\left(\mathbb{R}^n\right)$.

(ii) If $c\max\{d_1,d_2\} \geqslant 1$, then (27) holds. If $c\max\{d_1,d_2\} < 1$, then $\mathcal{H}_p^F\left(\boldsymbol{\xi}\right) = -\infty$ for any $F \in \mathcal{M}_p\left(\mathbb{R}^n\right)$.

Remark 3. For $c > 1$, the following special cases of Corollaries 2 and 3 recover some standard loss functions:

(i) Taking $f(\boldsymbol{\xi}) = \boldsymbol{\beta}^\top \boldsymbol{\xi}$ and $\ell(z,t) = c(z-t)_+$ which responds to $d_j = -1$ for all $j \in I_2$ in Corollary 2, we get $\ell(f(\boldsymbol{\xi}),t) = c(\boldsymbol{\beta}^\top \boldsymbol{\xi} - t)_+$.

(ii) Taking $f(\boldsymbol{\xi}) = -\boldsymbol{\beta}^\top \boldsymbol{\xi}$ and $\ell(z,t) = c(z+t)_+$ which responds to $d_j = 1$ for all $j \in I_2$ in Corollary 2, we get $\ell(f(\boldsymbol{\xi}),t) = c(\boldsymbol{\beta}^\top \boldsymbol{\xi} - t)_-$.

(iii) Taking $f(\boldsymbol{\xi}) = |\boldsymbol{\beta}^\top \boldsymbol{\xi}| = \max\{\boldsymbol{\beta}^\top \boldsymbol{\xi}, -\boldsymbol{\beta}^\top \boldsymbol{\xi}\}$ and $\ell(z,t) = c(z-t)_+$ which responds to $d_j = -1$ for all $j \in I_2$ in Corollary 2, we have $\ell(f(\boldsymbol{\xi}),t) = c(|\boldsymbol{\beta}^\top \boldsymbol{\xi}| - t)_+$.

(iv) Taking $f(\boldsymbol{\xi}) = \boldsymbol{\beta}^\top \boldsymbol{\xi} - b_1$ and $\ell(z,t) = c\left(|z-t| + b_2\right)_+$ which corresponds to $d_1 = 1$, $d_2 = -1$, and $d_3 = b_2 \in \mathbb{R}$ in Corollary 3, we have $\ell(f(\boldsymbol{\xi}),t) = c\left(|\boldsymbol{\beta}^\top \boldsymbol{\xi} - t - b_1| + b_2\right)_+$. In particular, if $b_2 \geqslant 0$, the loss reduces to $\ell(f(\boldsymbol{\xi}),t) = c\left(|\boldsymbol{\beta}^\top \boldsymbol{\xi} - t - b_1| + b_2\right)$.

These cases demonstrate that Corollaries 2 and 3 broaden the scope of Corollary 1 in Wu et al. (2022), encompassing it as a special case.

Example 4 (Higher-order risk measure)**.** For any $p \geqslant 1$ and $c > 1$, let $f(\boldsymbol{\xi}) = \max_{i\in I}\{\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\}$ for some $\boldsymbol{\beta}_i \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}_i\|_* = 1$ and $b_i \in \mathbb{R}$ for all $i \in I$. Consider the loss function $\ell(z,t) := c(z-t)_+$. Then, by Corollary 2 (ii), we have

$$\sup_{F\in\mathbb{B}_p(F_0,\delta)} \mathcal{H}_p^F\left(\max_{i\in I}\{\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\}\right) = \mathcal{H}_p^{F_0}\left(\max_{i\in I}\{\boldsymbol{\beta}_i^\top \boldsymbol{\zeta} + b_i\}\right) + c\varepsilon.$$

Distortion Risk Functionals

We now turn to distortion risk functionals, another important generalization of the expectation. A risk measure ρ_h is called a distortion risk measure if

$$\rho_h^F(Z) = \int_0^1 \mathrm{VaR}_u^F(Z)\,\mathrm{d}h(u),$$

where $h : [0,1] \to [0,1]$ is increasing with $h(0) = 0$ and $h(1) = 1$, referred to as the *distortion function*. We focus on the case where h is convex, in which case ρ_h is coherent (i.e., $\rho_h \in \mathcal{R}_{\mathrm{coh}}$), and define

$$\|h'\|_q := \left(\int_0^1 (h'(u))^q\,\mathrm{d}u\right)^{1/q},$$

where h' denotes the left-hand derivative of h.

We now establish a precise characterization of the loss functions f for which projection equivalence holds—conditions that are both necessary and sufficient for an exact regularization reformulation.

Theorem 5. *Let h be an increasing and convex distortion function satisfying $\|h'\|_q \in \mathbb{R}$, and $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. For $p > 1$, there exists $c_f \geqslant 0$ such that*

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \rho_h^F(f(\boldsymbol{\xi})) = \sup_{G\in\mathcal{C}_p(f|F_0,c_f\varepsilon)} \rho_h^G(X) = \rho_h^{F_0}(f(\boldsymbol{\zeta})) + c_f\varepsilon\|h'\|_q \tag{28}$$

holds for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ *and* $\varepsilon > 0$ *if and only if* f *is given by* (13).

Strikingly, the same structural form identified in Proposition 3 re-emerges here as the necessary and sufficient condition.

4.2 First-Order Case ($p = 1$)

Following the observation in Section 3.3 that projection equivalence holds for a substantially broader class of loss functions f when $p = 1$, we show that the same phenomenon extends to the risk measures considered in the previous section: restricting to $p = 1$ yields exact regularization for a strikingly larger class of loss functions.

We first examine the two functionals in (25) specialized to $p = 1$.

Proposition 5. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *with* $\mathrm{Lip}(f) \in \mathbb{R}_+$ *satisfy that there exist* $\mathbf{x}_0 \in \mathbb{R}^n$ *and* $\boldsymbol{v}_k \in \mathbb{R}^n$ *with* $\|\boldsymbol{v}_k\| = 1$, $k \in \mathbb{N}$ *such that* (21) *holds, and* $\ell : \mathbb{R}^2 \to \mathbb{R}$ *be a function satisfying* $\mathrm{Lip}(\ell(\cdot, t)) = b \in \mathbb{R}$ *for all* $t \in \mathbb{R}$ *and for each* $t \in \mathbb{R}$ *there exists* $z_0(t)$ *such that*

$$\lim_{m\to\infty} \frac{\ell(z_0(t) + m, t) - \ell(z_0(t), t)}{m} = b. \tag{29}$$

We have the following statements hold:

(i) If $\ell(z, t)$ *is convex in* t *with* $\lim_{t\to-\infty} \ell'_t(z, t) < 0 < \lim_{t\to\infty} \ell'_t(z, t)$ *for all* $z \in \mathbb{R}$, *where* $\ell'_t(z, t)$ *denotes the left derivative of* ℓ *with respect to* t, *then for any* $F_0 \in \mathcal{M}(\mathbb{R}^n)$ *and* $\varepsilon > 0$,

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \mathcal{D}_1^F(f(\boldsymbol{\xi})) = \mathcal{D}_1^{F_0}(f(\boldsymbol{\zeta})) + b\,\mathrm{Lip}(f)\varepsilon. \tag{30}$$

(ii) If $\ell(z, t)$ *is convex in* t *with* $\lim_{t\to-\infty} \ell'_t(z, t) < -1 < \lim_{t\to\infty} \ell'_t(z, t)$ *for all* $z \in \mathbb{R}$, *then for any* $F_0 \in \mathcal{M}(\mathbb{R}^n)$ *and* $\varepsilon > 0$,

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \mathcal{H}_1^F(f(\boldsymbol{\xi})) = \mathcal{H}_1^{F_0}(f(\boldsymbol{\zeta})) + b\,\mathrm{Lip}(f)\varepsilon. \tag{31}$$

Remark 4. Note that the functions ℓ in Corollaries 2 and 3 both satisfy assumption (29) with $b = c$. In Corollary 2, if $\min_{j\in I_2} d_j \leqslant 0 \leqslant \max_{j\in I_2} d_j$, then (30) holds, and if $c \min_{j\in I_2} d_j \leqslant -1$, then (31) holds. Likewise, in Corollary 3, if $\min\{d_1, d_2\} \leqslant 0 \leqslant \max\{d_1, d_2\}$, then (30) holds, and if $c \max\{d_1, d_2\} \geqslant 1$, then (31) holds.

Next, we turn to distortion risk functionals under the Wasserstein-1 ball.

Proposition 6. *Let h be an increasing and convex distortion function satisfying $\|h'\|_q \in (0,\infty)$. and $f:\mathbb{R}^n\to\mathbb{R}$ satisfy the conditions in Proposition 5. We have*

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \rho_h^F\left(f\left(\boldsymbol{\xi}\right)\right) = \sup_{G\in\mathcal{C}_1(f|F_0,c_f\varepsilon)} \rho_h^G(X) = \rho_h^{F_0}\left(f\left(\boldsymbol{\zeta}\right)\right) + \mathrm{Lip}(f)\varepsilon\left\|h'\right\|_\infty. \tag{32}$$

holds for any $F_0\in\mathcal{M}(\mathbb{R}^n)$ and $\varepsilon>0$.

Beyond the risk functionals discussed above, further "regularization-like" reformulations can be derived for other coherent risk measures through projection equivalence. A notable example is the expectile. For a risk X with distribution F, the expectile at level α, $\mathrm{ex}_\alpha^F(X)$, is defined as the unique solution to

$$\alpha\mathbb{E}^F[(X-x)_+] = (1-\alpha)\mathbb{E}^F[(X-x)_-], \tag{33}$$

and it is coherent when $\alpha\geqslant 1/2$ (Bellini et al., 2014). By Theorem 3, if f satisfies the conditions in Proposition 6, then

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \mathrm{ex}_\alpha^F\left(f\left(\boldsymbol{\xi}\right)\right) = \sup_{G\in\mathcal{C}_1(f|F_0,\mathrm{Lip}(f)\varepsilon)} \mathrm{ex}_\alpha^G(X).$$

The next proposition shows that the worst-case expectile coincides with the solution to the regularized version of (33).

Proposition 7. *For $\alpha\in[1/2,1]$, $F_0\in\mathcal{M}(\mathbb{R}^n)$, $\varepsilon>0$, and a function f satisfying the conditions in Proposition 5, we have $\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)}\mathrm{ex}_\alpha^F\left(f\left(\boldsymbol{\xi}\right)\right)$ is the unique solution to*

$$\mathbb{E}^{F_0}[\alpha(f(\boldsymbol{\zeta})-x)_+ - (1-\alpha)(f(\boldsymbol{\zeta})-x)_-] + \alpha\mathrm{Lip}(f)\varepsilon = 0.$$

5 Classification

As a natural extension of the main Wasserstein DRO problem (1), we now turn to a setup motivated by classification in machine learning. Here, the random vector consists of a class label and a feature vector, $\boldsymbol{\xi}=(Y,\mathbf{X})\in\Xi$, where $\Xi:=\{-1,1\}\times\mathbb{R}^n\subseteq\mathbb{R}^{n+1}$, with Y denoting a binary label and $\mathbf{X}$ the associated features. The classification task is to select a decision function $f:\mathbb{R}^n\to\mathbb{R}$, interpreted as a classifier, from a class $\mathcal{A}$ to predict the sign of Y given $\mathbf{X}$.

To capture distributional robustness, we equip Ξ with the type-p Wasserstein metric

$$\overline{\mathrm{W}}_p\left(F_1,F_2\right) := \inf_{\pi\in\overline{\Pi}(F_1,F_2)} \left(\mathbb{E}^\pi[d(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)^p]\right)^{1/p},$$

where $\overline{\Pi}(F_1, F_2)$ is the set of all distributions on Ξ with marginals F_1 and F_2 supported on Ξ, the distance between $\boldsymbol{\xi}_1 = (Y_1, \mathbf{X}_1)$ and $\boldsymbol{\xi}_2 = (Y_2, \mathbf{X}_2)$ is defined via the additively separable form

$$d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \|\mathbf{X}_1 - \mathbf{X}_2\| + \Theta\left(Y_1 - Y_2\right), \tag{34}$$

with $\|\cdot\|$ a norm on $\mathbb{R}^n$. The penalty function $\Theta : \mathbb{R} \to \{0, \infty\}$ is specified by $\Theta(0) = 0$ and $\Theta(s) = \infty$ for $s \neq 0$. Hence, the metric prohibits perturbations in the label Y while allowing adversarial shifts in the feature space $\mathbf{X}$.

For a nominal distribution $F_0 \in \mathcal{M}_p(\Xi)$ and robustness radius $\varepsilon > 0$, the distributionally robust classification problem is given by

$$\inf_{f \in \mathcal{A}} \sup_{F \in \overline{\mathbb{B}}_p(F_0, \varepsilon)} \rho^F\big(Y \cdot f(\mathbf{X})\big), \tag{35}$$

where $\overline{\mathbb{B}}_p(F_0, \varepsilon) := \{F \in \mathcal{M}(\Xi) : \overline{\mathrm{W}}_p\left(F, F_0\right) \leqslant \varepsilon\}$, and ρ is a risk functional applied to the *margin* $Z := Y \cdot f(\mathbf{X})$.

To study projection equivalence in this setting, let $(Y_0, \mathbf{X}_0) \sim F_0$ and denote by G_0 the distribution of the baseline margin $Y_0 \cdot f(\mathbf{X}_0)$. We introduce the one-dimensional Wasserstein ball

$$\overline{\mathcal{C}}_p(f \mid F_0, \varepsilon) := \mathbb{B}_p(G_0, \varepsilon) \subseteq \mathcal{M}(\mathbb{R}).$$

A classifier f admits *classification projection equivalence* if

$$\sup_{F \in \overline{\mathbb{B}}_p(F_0, \varepsilon)} \rho^F(Y \cdot f(\mathbf{X})) = \sup_{G \in \overline{\mathcal{C}}_p(f | F_0, \varepsilon)} \rho^G(Z). \tag{36}$$

As shown below, in contrast to projection equivalence in Section 3, classification projection equivalence holds only for classifiers f that strictly satisfy set-level equivalence. This requirement remains even when ρ is restricted to the monotone class $\mathcal{R}_{\mathrm{mon}}$, underscoring that exact reduction in classification demands stronger conditions than in the general Wasserstein DRO framework (1).

Proposition 8. *For $p \geqslant 1$ and $\alpha \in [0, 1)$, let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. The following statements are equivalent.*

(i) There exists $c_f \geqslant 0$ such that

$$\sup_{F \in \overline{\mathbb{B}}_p(F_0, \varepsilon)} \rho^F\left(Y \cdot f(\mathbf{X})\right) = \sup_{G \in \overline{\mathcal{C}}_p(f | F_0, c_f \varepsilon)} \rho^G(X) \tag{37}$$

holds for any $\rho \in \mathcal{R}_{\mathrm{mon}}$, $F_0 \in \mathcal{M}(\Xi)$, *and* $\varepsilon > 0$.

(ii) There exists $c_f \geqslant 0$ *such that for any* $F_0 \in \mathcal{M}(\Xi)$ *and* $\varepsilon > 0$, *it holds that*

$$\{F_{Yf(\mathbf{X})} : \ F_{(Y,\mathbf{X})} \in \overline{\mathbb{B}}_p(F_0, \varepsilon)\} = \overline{\mathcal{C}}_p(f|F_0, c_f\varepsilon). \tag{38}$$

(iii) The function f *is Lipschitz continuous, and satisfies* (8).

Moreover, this necessity remains even under specialized risk measures; in particular, the characterization (8) is still required when the risk functional is Value-at-Risk.

Proposition 9. *Fix* $\alpha \in [0, 1)$ *and* $p \geqslant 1$ *and let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a function. There exists* $c_f \geqslant 0$ *such that*

$$\sup_{F \in \overline{\mathbb{B}}_p(F_0,\varepsilon)} \mathrm{VaR}_\alpha^F(Y \cdot f(\mathbf{X})) = \sup_{G \in \overline{\mathcal{C}}_p(f|F_0,c_f\varepsilon)} \mathrm{VaR}_\alpha^G(X) \tag{39}$$

holds for any $F_0 \in \mathcal{M}(\Xi)$ *and* $\varepsilon > 0$ *if and only if* f *satisfies* (8).

Recall that when the norm $\|\cdot\| = \|\cdot\|_a$ is the ℓ_a-norm for some $a \in [1, \infty)$, Propositions 1 and 2 identify explicit classifier forms for which classification projection equivalence holds. In combination with this result, Proposition 8 yields the following regularization reformulation, linking robust classification directly to a familiar paradigm in machine learning.

Corollary 4. *For* $p \geqslant 1$, $F_0 \in \mathcal{M}(\Xi)$ *with* $(Y_0, \mathbf{X}_0) \sim F_0$, *let* h *be an increasing and convex distortion function satisfying* $\|h'\|_q \in (0, \infty)$. *We have the following statements hold.*

(i) If $f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + b$ *for some* $\boldsymbol{\beta} \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$, *we have*

$$\sup_{F \in \overline{\mathbb{B}}_p(F_0,\varepsilon)} \rho_h^F(Y \cdot f(\mathbf{X})) = \rho_h^{F_0}(Y_0 \cdot f(\mathbf{X_0})) + \|\boldsymbol{\beta}\|_* \varepsilon \|h'\|_q.$$

(ii) If $\|\cdot\| = \|\cdot\|_1$ *is the* ℓ_1*-norm and* f *is given by Corollary 1, then we have*

$$\sup_{F \in \overline{\mathbb{B}}_p(F_0,\varepsilon)} \rho_h^F(Y \cdot f(\mathbf{X})) = \rho_h^{F_0}(Y_0 \cdot f(\mathbf{X_0})) + c\varepsilon \|h'\|_q.$$

We note that while case (i) can also be obtained through a direct analysis of linear classifiers, as shown in Wu et al. (2022), case (ii) emerges only within the more general framework of Proposition 8.

6 Conclusion

In this work, we provide the first complete characterization of *projection equivalence* in Wasserstein distributionally robust optimization. Our central finding reveals that this powerful high-dimensional reduction is not confined to the restrictive case of set-level equivalence but extends to a much broader class of loss functions. By systematically navigating a hierarchy of risk functionals, we establish a sharp boundary delineating precisely when a high-dimensional worst-case risk evaluation simplifies to its one-dimensional counterpart. This foundational result, in turn, enables us to derive necessary and sufficient conditions for Wasserstein DRO problems to admit exact regularization reformulations, unifying two central paradigms in optimization and machine learning. Ultimately, our analysis delivers new classes of tractable models and establishes the fundamental limits of such reductions, clarifying the structural properties that govern computational feasibility in distributionally robust optimization.

A Appendix: Proofs of the Main results

subsectionProofs for Section 3 To show Theorem 1, we need the following lemma.

Lemma A1. *For any* $p \geqslant 1$, $\varepsilon > 0$, $\alpha \in [0,1)$ *and* $f : \mathbb{R}^n \to \mathbb{R}$, *it holds that* $\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p]\leqslant\varepsilon^p} \mathrm{VaR}_\alpha(f(\boldsymbol{\xi})) = \sup_{\|\mathbf{z}-\mathbf{x}\|\leqslant\varepsilon/(1-\alpha)^{1/p}} f(\mathbf{z})$.

Proof. First note that for any $\mathbf{z} \in \mathbb{R}^n$ with $\|\mathbf{z}-\mathbf{x}\| \leqslant \varepsilon/(1-\alpha)^{1/p}$ with $f(\mathbf{z}) \geqslant f(\mathbf{x})$, take $\boldsymbol{\xi} \sim \alpha\delta_{\mathbf{x}} + (1-\alpha)\delta_{\mathbf{z}}$. We have $\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p] \leqslant \varepsilon^p$ and $\mathrm{VaR}_\alpha(f(\boldsymbol{\xi})) = f(\mathbf{z})$. Thus, $\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p]\leqslant\varepsilon^p} f(\boldsymbol{\xi}) \geqslant \sup_{\|\mathbf{z}-\mathbf{x}\|\leqslant\varepsilon/(1-\alpha)^{1/p}} f(\mathbf{z})$.

On the other hand, for any random vector $\boldsymbol{\xi}$ with $\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p] \leqslant \varepsilon^p$, define

$$\boldsymbol{\xi}^* \sim \alpha\delta_{\mathbf{x}} + (1-\alpha)\delta_{\mathbf{x_2}}$$

with $\mathbf{x}_2 = \arg\min_{\{\mathbf{z}:f(\mathbf{z})\geqslant\mathrm{VaR}_\alpha(f(\boldsymbol{\xi}))\}} \|\mathbf{z}-\mathbf{x}\|$. Then we have $\mathrm{VaR}_\alpha(f(\boldsymbol{\xi}^*)) \geqslant f(\mathbf{x}_2) \geqslant \mathrm{VaR}_\alpha(f(\boldsymbol{\xi}))$ and $\mathbb{E}[\|\boldsymbol{\xi}^*-\mathbf{x}\|^p] \leqslant \varepsilon^p$. This implies that $\|\mathbf{x}_2-\mathbf{x}\| \leqslant \varepsilon/(1-\alpha)^{1/p}$ and thus, $\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p]\leqslant\varepsilon^p} \mathrm{VaR}_\alpha(f(\boldsymbol{\xi})) \leqslant \sup_{\|\mathbf{z}-\mathbf{x}\|\leqslant\varepsilon/(1-\alpha)^{1/p}} f(\mathbf{z})$. This completes the proof. □

Proof of Theorem 1. Note that the implication (ii) ⇒ (i) is trivial. We only give the proof of (i) ⇒ (iii) and (iii) ⇒ (ii).

For (i) ⇒ (iii), we first consider the case where $c_f = 0$. We choose $\rho = \mathrm{VaR}_\alpha$ for some $\alpha \in [0,1)$

and take $F_0 = \delta_{\mathbf{x}}$ for $\mathbf{x} \in \mathbb{R}^n$. In this case, equation (6) reduces to

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p] \leqslant \varepsilon^p} \mathrm{VaR}_\alpha\left(f(\boldsymbol{\xi})\right) = \sup_{\mathbb{E}[|X-f(\mathbf{x})|^p] \leqslant 0} \mathrm{VaR}_\alpha^G\left(X\right). \tag{A1}$$

By Lemma A1, we have

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p] \leqslant \varepsilon^p} \mathrm{VaR}_\alpha(f(\boldsymbol{\xi})) = \sup_{\|\mathbf{z}-\mathbf{x}\| \leqslant \frac{\varepsilon}{(1-\alpha)^{1/p}}} f(\mathbf{z}).$$

For $\alpha \in [0,1)$, since $\varepsilon > 0$ is arbitrary, it follows that for any $\mathbf{x} \in \mathbb{R}^n$, (A1) is equivalent to

$$\sup_{\|\mathbf{z}-\mathbf{x}\| \leqslant \varepsilon} f(\mathbf{z}) = f(\mathbf{x}), \quad \forall\, \varepsilon > 0.$$

Then, we have $f(\mathbf{z}) \leqslant f(\mathbf{x})$ and $f(\mathbf{z}) \geqslant f(\mathbf{x})$ for any $\mathbf{z}, \mathbf{x} \in \mathbb{R}^n$, which implies f is constant. Thus, there exists $b \in \mathbb{R}$ such that $f(\mathbf{x}) \equiv b$ for $\mathbf{x} \in \mathbb{R}^n$. We next consider the case where $c_f > 0$ and, without loss of generality, set $c_f = 1$. For $\rho = \mathrm{VaR}_\alpha$ with $\alpha \in [0,1)$, we first show that (6) implies that

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, \varepsilon > 0. \tag{A2}$$

To that end, take $F_0 = \delta_{\mathbf{x}}$ for $\mathbf{x} \in \mathbb{R}^n$. In this case, (6) reduces to

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p] \leqslant \varepsilon^p} \mathrm{VaR}_\alpha(f(\boldsymbol{\xi})) = \sup_{\mathbb{E}[|Z-f(\mathbf{x})|^p] \leqslant \varepsilon^p} \mathrm{VaR}_\alpha(Z). \tag{A3}$$

By Lemma A1, we have

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{x}\|^p] \leqslant \varepsilon^p} \mathrm{VaR}_\alpha(f(\boldsymbol{\xi})) = \sup_{\|\mathbf{z}-\mathbf{x}\| \leqslant \frac{\varepsilon}{(1-\alpha)^{1/p}}} f(\mathbf{z})$$

and

$$\sup_{\mathbb{E}[|Z-f(\mathbf{x})|^p] \leqslant \varepsilon^p} \mathrm{VaR}_\alpha(Z) = \sup_{|z-f(\mathbf{x})| \leqslant \frac{\varepsilon}{(1-\alpha)^{1/p}}} z = f(\mathbf{x}) + \frac{\varepsilon}{(1-\alpha)^{1/p}}.$$

Combining these two equations with (A3) yields

$$\sup_{\|\mathbf{z}-\mathbf{x}\| \leqslant \frac{\varepsilon}{(1-\alpha)^{1/p}}} f(\mathbf{z}) = f(\mathbf{x}) + \frac{\varepsilon}{(1-\alpha)^{1/p}}.$$

For $\alpha \in [0,1)$, since $\varepsilon > 0$ is arbitrary, it follows that for any $\varepsilon > 0$ and $\mathbf{x} \in \mathbb{R}^n$,

$$\sup_{\|\mathbf{z}-\mathbf{x}\| \leqslant \varepsilon} f(\mathbf{z}) = f(\mathbf{x}) + \varepsilon,$$

that is, (A2) holds. This identity immediately implies that

$$|f(\mathbf{x}) - f(\mathbf{y})| \leqslant \|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

and thus f is Lipschitz continuous with Lipschitz constant $\mathrm{Lip}(f) = 1$. Moreover, since (6) holds for any risk measure, by taking $\rho = -\mathrm{VaR}_\alpha$ and repeating a similar argument, we obtain that for all $\varepsilon > 0$ and $\mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) - \inf_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) = \varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, \varepsilon > 0,$$

and thus (8) follows. This completes the proof of this direction.

For (iii) $\Rightarrow$ (ii), if $\mathrm{Lip}(f) = 0$, then f is constant, i.e., there exists $b \in \mathbb{R}$ such that $f(\mathbf{x}) \equiv b$ for all $\mathbf{x} \in \mathbb{R}^n$. In this case, one can take $c_f = 0$, and

$$\mathbb{B}_p(f|F_0, \varepsilon) = \{\delta_b\} = \mathcal{C}_p(f|F_0, 0),$$

for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$. If $\mathrm{Lip}(f) > 0$, we assume without loss generality that $\mathrm{Lip}(f) = 1$. Then, since f is 1-Lipschitz continuous, it follows that for any $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^n$,

$$|f(\boldsymbol{\xi}_1) - f(\boldsymbol{\xi}_2)| \leqslant \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|.$$

Hence, if $\mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] \leqslant \varepsilon^p$, then

$$\mathbb{E}\left[|f(\boldsymbol{\xi}) - f(\boldsymbol{\zeta})|^p\right] \leqslant \mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] \leqslant \varepsilon^p,$$

which implies $\mathbb{B}_p(f|F_0, \varepsilon) \subseteq \mathcal{C}_p(f|F_0, \varepsilon)$. To prove the reverse inclusion, it suffices to show that for any $G \in \mathcal{C}_p(f|F_0, \varepsilon)$ and $Z \sim G$ with $\mathbb{E}[|Z - f(\boldsymbol{\zeta})|^p] \leqslant \varepsilon^p$, there exists a random vector $\boldsymbol{\xi}$ with distribution $F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon)$ such that $f(\boldsymbol{\xi}) = Z$ almost surely. To this end, denote by $T := Z - f(\boldsymbol{\zeta})$. Then, we have $\mathbb{E}\left[|T|^p\right] \leqslant \varepsilon^p$. We first assert that there exist measurable mappings $\mathbf{V}_1$ and $\mathbf{V}_2$ such that

$$\mathbf{V}_1(\omega) \in \operatorname*{arg\,max}_{\|\mathbf{y}\| \leqslant |T(\omega)|} f\left(\boldsymbol{\zeta}(\omega) + \mathbf{y}\right) \quad \text{and} \quad \mathbf{V}_2(\omega) \in \operatorname*{arg\,min}_{\|\mathbf{y}\| \leqslant |T(\omega)|} f\left(\boldsymbol{\zeta}(\omega) + \mathbf{y}\right), \quad \omega \in \Omega.$$

To justify this, we invoke a measurable selection theorem (Theorem 3.5 in Rieder (1978)). For $\mathbf{V}_1$,

define

$$v(\omega) = \sup_{\mathbf{y} \in D(\omega)} u(\omega, \mathbf{y}), \quad \omega \in \Omega,$$

where $D(\omega) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\| \leqslant |T(\omega)|\}$ and $u(\omega, \mathbf{y}) = f(\boldsymbol{\zeta}(\omega) + \mathbf{y})$. Further set

$$E := \{(\omega, \mathbf{y}) \in \Omega \times \mathbb{R}^n : \|\mathbf{y}\| \leqslant (|T(\omega)|\}.$$

Let $\mathrm{P}_1(E)$ be the projection of E onto its first coordinate. The first three conditions of Theorem 3.5 in Rieder (1978) are immediate in our setting. For the last condition, take $c \in \mathbb{R}$ and $\omega \in \mathrm{P}_1(E)$. Then, we have

$$\{\mathbf{y} \in D(\omega) : u(\omega, \mathbf{y}) \geqslant c\} = D(\omega) \cap \{\mathbf{y} \in \mathbb{R}^n : f(\boldsymbol{\zeta}(\omega) + \mathbf{y}) \geqslant c\}.$$

Note that $D(\omega)$is compact and, since f is continuous, the set $\{\mathbf{y} \in \mathbb{R}^n : f(\boldsymbol{\zeta}(\omega) + \mathbf{y}) \geqslant c\}$ is closed. Consequently, $\{\mathbf{y} \in D(\omega) : u(\omega, \mathbf{y}) \geqslant c\}$ is compact, which verifies the last condition in Theorem 3.5 of Rieder (1978). Hence a measurable maximizer $\mathbf{V}_1$ exists and the existence of $\mathbf{V}_2$ follows analogously. Denote by $A_+ := \{\omega : T(\omega) \geqslant 0\}$ and $A_- := \{\omega : T(\omega) < 0\}$. We define $\boldsymbol{\xi}$ as

$$\boldsymbol{\xi}(\omega) = (\boldsymbol{\zeta}(\omega) + \mathbf{V}_1(\omega))\mathbb{1}_{A_+}(\omega) + (\boldsymbol{\zeta}(\omega) + \mathbf{V}_2(\omega))\mathbb{1}_{A_-}(\omega).$$

Since $\mathbf{V}_1$ and $\mathbf{V}_2$ are measurable, it follows that $\boldsymbol{\xi}$ is measurable. As $\Omega = A_+ \cup A_-$, we have

$$\|\boldsymbol{\xi} - \boldsymbol{\zeta}\| \leqslant \max\{\|\mathbf{V}_1(\omega)\|, \|\mathbf{V}_2(\omega)\|\} \leqslant |T(\omega)|, \quad \omega \in \Omega.$$

Then, we have $\mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] \leqslant \mathbb{E}[|T|^p] \leqslant \varepsilon^p$. This implies $F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon)$. Moreover, by the definitions of $\mathbf{V}_1, \mathbf{V}_2$ and (8), we have

$$f(\boldsymbol{\xi}(\omega)) = f(\boldsymbol{\zeta}(\omega)) + T(\omega) = Z(\omega), \quad \omega \in A_+,$$

and

$$f(\boldsymbol{\xi}(\omega)) = f(\boldsymbol{\zeta}(\omega)) - (-T(\omega)) = Z(\omega), \quad \omega \in A_-.$$

Therefore, $\mathcal{C}_p(f|F_0, \varepsilon) \subseteq \mathbb{B}_p(f|F_0, \varepsilon)$ and thus, (7) holds. This completes the proof. □

Proof of Proposition 1. Without loss of generality, assume $\mathrm{Lip}(f) = 1$. It suffices to show that if f satisfies (8), then (9) holds. We first show that (8) implies that for each $\mathbf{x} \in \mathbb{R}^n$ there exist $\boldsymbol{\beta}_{\mathbf{x}}$ and

$\boldsymbol{\eta}_{\mathbf{x}}$ with $\|\boldsymbol{\beta}_{\mathbf{x}}\| = \|\boldsymbol{\eta}_{\mathbf{x}}\| = 1$ such that

$$f(\mathbf{x}) - f(\mathbf{x} + \varepsilon\boldsymbol{\eta}_{\mathbf{x}}) = f(\mathbf{x} + \varepsilon\boldsymbol{\beta}_{\mathbf{x}}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \varepsilon > 0. \tag{A4}$$

Indeed, since

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \varepsilon > 0, \mathbf{x} \in \mathbb{R}^n,$$

it follows that for each for each $\varepsilon > 0$ and $\mathbf{x}$, there exists $\boldsymbol{\beta}_{\mathbf{x},\varepsilon}$ with $\|\boldsymbol{\beta}_{\mathbf{x},\varepsilon}\| = \varepsilon$ such that

$$f(\mathbf{x} + \boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x}) = \|\boldsymbol{\beta}_{\mathbf{x},\varepsilon}\| = \varepsilon. \tag{A5}$$

We next assert that $\boldsymbol{\beta}_{\mathbf{x},\varepsilon}$ can be chosen as the same for different ε. To show this, suppose that (A5) holds for some $\varepsilon > 0$. We aim to show that for any $c < 1$,

$$f(\mathbf{x} + c\boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x}) = c\|\boldsymbol{\beta}_{\mathbf{x},\varepsilon}\| = c\varepsilon. \tag{A6}$$

Suppose otherwise, $f(\mathbf{x} + c\boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x}) < c\varepsilon$ for some $c < 1$. By (A5), we have $f(\mathbf{x} + \boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x} + c\boldsymbol{\beta}_{\mathbf{x},\varepsilon}) \leqslant (1 - c)\varepsilon$, and thus,

$$f(\mathbf{x} + \boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x}) = f(\mathbf{x} + \boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x} + c\boldsymbol{\beta}_{\mathbf{x},\varepsilon}) + f(\mathbf{x} + c\boldsymbol{\beta}_{\mathbf{x},\varepsilon}) - f(\mathbf{x}) < \varepsilon,$$

which yields a contradiction to (A5). Therefore, for each $\mathbf{x} \in \mathbb{R}^n$, there exists $\boldsymbol{\beta}_{\mathbf{x}} \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}_{\mathbf{x}}\| = 1$ such that

$$f(\mathbf{x} + \varepsilon\boldsymbol{\beta}_{\mathbf{x}}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \varepsilon > 0.$$

By symmetry, applying the analogous argument to

$$f(\mathbf{x}) - \inf_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) = \varepsilon, \quad \forall\, \varepsilon > 0, \mathbf{x} \in \mathbb{R}^n,$$

we obtain (A4). We now prove that (9) holds by contradiction. If there exist $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$ such that $f(\mathbf{x}) \leqslant f(\mathbf{y})$ and $\|\boldsymbol{\eta}_{\mathbf{x}} - \boldsymbol{\beta}_{\mathbf{y}}\| < \|\boldsymbol{\eta}_{\mathbf{x}}\| + \|\boldsymbol{\beta}_{\mathbf{y}}\| = 2$, then define $\mathbf{z} = \mathbf{x} + c\boldsymbol{\eta}_{\mathbf{x}}$ and $\mathbf{z}_1 = \mathbf{y} + c\boldsymbol{\beta}_{\mathbf{y}}$. There exists c large enough such that $\|\mathbf{x} - \mathbf{y}\| < c(2 - \|\boldsymbol{\beta}_{\mathbf{y}} - \boldsymbol{\eta}_{\mathbf{x}}\|)$, and thus,

$$c_0 := \|\mathbf{z}_1 - \mathbf{z}\| \leqslant \|\mathbf{x} - \mathbf{y}\| + c\|\boldsymbol{\beta}_{\mathbf{y}} - \boldsymbol{\eta}_{\mathbf{x}}\| < 2c.$$

This implies that

$$\sup_{\|\mathbf{w}\|\leqslant c_0} f(\mathbf{z}+\mathbf{w}) - f(\mathbf{z}) \geqslant f(\mathbf{z}_1) - f(\mathbf{z}) \geqslant f(\mathbf{z}_1) - f(\mathbf{y}) + f(\mathbf{x}) - f(\mathbf{z}) = 2c,$$

where the second inequality follows from $f(\mathbf{y}) \geqslant f(\mathbf{x})$, and the equality follows from (A4). This yields a contradiction to (8). Thus, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, if $f(\mathbf{x}) \leqslant f(\mathbf{y})$ (hence including $\mathbf{x} = \mathbf{y}$), then $\|\boldsymbol{\eta}_\mathbf{x} - \boldsymbol{\beta}_\mathbf{y}\| = \|\boldsymbol{\eta}_\mathbf{x}\| + \|\boldsymbol{\beta}_\mathbf{y}\|$. Since $\|\boldsymbol{\eta}_\mathbf{x}\| = \|\boldsymbol{\beta}_\mathbf{y}\| = 1$ and the norm is strictly convex, it follows that $\boldsymbol{\beta}_\mathbf{x} = -\boldsymbol{\eta}_\mathbf{x} = \boldsymbol{\beta}_\mathbf{y}$ for all $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^n$ with $f(\mathbf{x}) \leqslant f(\mathbf{y})$. Otherwise we have $\|\boldsymbol{\eta}_\mathbf{x} - \boldsymbol{\beta}_\mathbf{y}\| < 2$, contradicting the requirement $\|\boldsymbol{\eta}_\mathbf{x} - \boldsymbol{\beta}_\mathbf{y}\| = \|\boldsymbol{\eta}_\mathbf{x}\| + \|\boldsymbol{\beta}_\mathbf{y}\| = 2$. Therefore, we have $\boldsymbol{\beta}_\mathbf{x}$ can be chosen as the same $\boldsymbol{\beta}$ for all $\mathbf{x} \in \mathbb{R}^n$. It then follows that there exists $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\| = 1$ such that

$$f(\mathbf{x}) - f(\mathbf{x} - \varepsilon\boldsymbol{v}) = f(\mathbf{x} + \varepsilon\boldsymbol{v}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\ \varepsilon > 0.$$

Since the case $\varepsilon = 0$ is trivially satisfied, it implies that (9) holds for any $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$.

For $\|\cdot\| = \|\cdot\|_a$, $a \in (1, \infty)$, it suffices to show for any $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$,

$$f(\mathbf{y}) - f(\mathbf{x}) = \boldsymbol{\beta}^\top(\mathbf{y} - \mathbf{x}), \tag{A7}$$

where $\boldsymbol{\beta} \in \mathbb{R}^n$ is the unique vector satisfying $\boldsymbol{\beta}^\top\boldsymbol{v} = \|\boldsymbol{\beta}\|_*\|\boldsymbol{v}\| = 1$, that is, $\operatorname{sgn}(\boldsymbol{\beta}_i) = \operatorname{sgn}(\boldsymbol{v}_i)$ for $i \in [n]$ and $|\boldsymbol{\beta}|^b = k|\boldsymbol{v}|^a$ for some $k > 0$ and b is the conjugate constant of a. By (9), we know (A7) holds for the case $\mathbf{y} - \mathbf{x} = t\boldsymbol{v}$, $t \in \mathbb{R}$. By $\boldsymbol{\beta}^\top\boldsymbol{v} = 1$, we have any vector in $\mathbb{R}^n$ can be written as a linear combination of $\boldsymbol{v}$ and a vector $\boldsymbol{u}$ with $\boldsymbol{\beta}^\top\boldsymbol{u} = 0$ and thus, it suffices to show (A7) for any $\mathbf{y} - \mathbf{x} = t\boldsymbol{u}$ with $\boldsymbol{\beta}^\top\boldsymbol{u} = 0$. We show it by contradiction. Suppose that there exist $\mathbf{y}$ and $\mathbf{x}$ such that $\mathbf{y} - \mathbf{x} = \boldsymbol{u}$, $\boldsymbol{\beta}^\top\boldsymbol{u} = 0$, and $f(\mathbf{x}) < f(\mathbf{y})$. Denote by $\varepsilon = f(\mathbf{y}) - f(\mathbf{x}) > 0$. Then we have

$$f(\mathbf{y} + t\boldsymbol{v}) - f(\mathbf{x}) = f(\mathbf{y} + t\boldsymbol{v}) - f(\mathbf{y}) + f(\mathbf{y}) - f(\mathbf{x}) = t + \varepsilon,$$

and

$$\|\mathbf{y} + t\boldsymbol{v} - \mathbf{x}\|^a = \|\boldsymbol{u} + t\boldsymbol{v}\|^a = \sum_{i=1}^n |u_i + tv_i|^a.$$

It follows that

$$\frac{f(\mathbf{y} + t\boldsymbol{v}) - f(\mathbf{x})}{\|\mathbf{y} + t\boldsymbol{v} - \mathbf{x}\|} = \frac{t + \varepsilon}{(\sum_{i=1}^n |u_i + tv_i|^a)^{1/a}} = \frac{1 + s\varepsilon}{(\sum_{i=1}^n |su_i + v_i|^a)^{1/a}}, \tag{A8}$$

where $s = 1/t$. Denote by $I_1 = \{i : v_i > 0\}$, $I_2 = \{i : v_i < 0\}$, $I_3 = \{i : v_i = 0, u_i \neq 0\}$, and

$g(s) = (1+s\varepsilon)^a - \sum_{i=1}^n |su_i + v_i|^a$. We have for $s > 0$ small enough,

$$\begin{aligned} g'(s) &= a(1+s\varepsilon)^{a-1}\varepsilon - \sum_{i\in I_1} a(su_i+v_i)^{a-1}u_i + \sum_{i\in I_2} a(-su_i - v_i)^{a-1}u_i - \sum_{i\in I_3} as^{a-1}|u_i|^a \\ &\overset{\text{sgn}}{=} (1+s\varepsilon)^{a-1}\varepsilon - \sum_{i\in I_1}(su_i+v_i)^{a-1}u_i + \sum_{i\in I_2}(-su_i - v_i)^{a-1}u_i - \sum_{i\in I_3} s^{a-1}|u_i|^a \\ &\to \varepsilon - \sum_{i\in I_1} v_i^{a-1}u_i + \sum_{i\in I_2}(-v_i)^{a-1}u_i = \varepsilon - \sum_{i=1}^n \operatorname{sgn}(v_i)|v_i|^{a-1}u_i = \varepsilon > 0 \quad \text{as } s \to 0, \end{aligned}$$

where the last equality follows from that $\boldsymbol{\beta}^\top \boldsymbol{u} = 0$, $|\boldsymbol{\beta}| = k_1|\boldsymbol{v}|^{a-1}$ and $\operatorname{sgn}(v_i) = \operatorname{sgn}(\beta_i)$. This, together with $g(0) = 0$, implies $g(s) > 0$ for some $s > 0$. Substituting this into (A8) yields that $\frac{f(\mathbf{y}+t\boldsymbol{v})-f(\mathbf{x})}{\|\mathbf{y}+t\boldsymbol{v}-\mathbf{x}\|} > 1$ for some $t > 0$. This yields a contradiction to (9). Therefore, we have (A7) holds and thus we complete the proof. □

Proof of Proposition 2. If $\|\cdot\| = \|\cdot\|_1$, then (A4) holds as well. For any $\mathbf{x}$, there exists $\boldsymbol{\beta}_\mathbf{x}$ with $\|\boldsymbol{\beta}_\mathbf{x}\|_1 = 1$ such that

$$f(\mathbf{x} + \varepsilon\boldsymbol{\beta}_\mathbf{x}) - f(\mathbf{x}) = \varepsilon = \varepsilon \sum_{i=1}^n |\beta_{\mathbf{x}i}| \quad \forall\, \varepsilon > 0.$$

Denote by $\mathbf{x}_0 := \mathbf{x}$ and $\mathbf{x}_k := \mathbf{x}_{k-1} + \varepsilon\beta_{\mathbf{x}k}\mathbf{e}_k$, $k \in [n]$, where $\beta_{\mathbf{x}k}$ denotes the k-th component of $\boldsymbol{\beta}_\mathbf{x}$. By Lipschitz continuity,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k-1}) \leqslant \varepsilon|\beta_{\mathbf{x}k}|, \quad k \in [n]. \tag{A9}$$

Summing over k yields

$$f(x + \varepsilon\boldsymbol{\beta}_\mathbf{x}) - f(x) \leqslant \varepsilon \sum_{k=1}^m |\beta_{\mathbf{x}k}| = \varepsilon.$$

Since the inequality holds with equality, each inequality in (A9) must also hold with equality. Moreover, the construction of the sequence $\mathbf{x}_k$ is independent of the order of coordinates. Hence, for any $\mathbf{x} \in \mathbb{R}^n$, there exists an index $i \in [n]$ such that

$$f(\mathbf{x} + t\tilde{\mathbf{e}}_i) - f(\mathbf{x}) = t, \quad \forall\, t > 0,$$

where $\tilde{\mathbf{e}}_i \in \{\pm\mathbf{e}_i\}, i \in [n]$. By (A4), there also exists $\boldsymbol{\eta}_\mathbf{x}$ with $\|\boldsymbol{\eta}_\mathbf{x}\|_1 = 1$ such that

$$f(\mathbf{x}) - f(\mathbf{x} + \varepsilon\boldsymbol{\eta}_\mathbf{x}) = \varepsilon, \quad \forall\, \varepsilon > 0.$$

Applying the same reasoning as above, we conclude that (10) holds. This completes the proof. □

Proof of Corollary 1. (a) It suffices to verify that f satisfies (10). Denote by $I := \{i \in [n], \beta_i \neq 0\}$.

For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\mathbf{x} \neq \mathbf{y}$,

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{y}) &= c(\boldsymbol{\beta}^\top(\mathbf{x}-\mathbf{y}) + g(\boldsymbol{\eta}\circ\mathbf{x}) - g(\boldsymbol{\eta}\circ\mathbf{y})) \\ &\leqslant c(\sum_{i\in I}|x_i - y_i| + \mathrm{Lip}(g)\sum_{i\in[n]\setminus I}|x_i - y_i|) \leqslant c\|\mathbf{x}-\mathbf{y}\|_1, \end{aligned}$$

where the first inequality follows from Hölder's inequality, and the second from the fact that $\mathrm{Lip}(g) \leqslant 1$. Thus f is Lipschitz continuous with Lipschitz constant c. Next, for any $\mathbf{x} \in \mathbb{R}^n$ and $\varepsilon > 0$, let $\boldsymbol{v}_1 := \mathrm{sgn}(\beta_i)\mathbf{e}_i$, $i \in I$. Then,

$$\begin{aligned} f(\mathbf{x}+\varepsilon\boldsymbol{v}_1) - f(\mathbf{x}) &= c\left(\boldsymbol{\beta}^\top(\mathbf{x}+\varepsilon\boldsymbol{v}_1) + g(\boldsymbol{\eta}\circ(\mathbf{x}+\varepsilon\boldsymbol{v}_1))\right) - c\left(\boldsymbol{\beta}^\top\mathbf{x} + g(\boldsymbol{\eta}\circ\mathbf{x})\right) \\ &= c\left(\varepsilon\boldsymbol{\beta}^\top\boldsymbol{v}_1 + g(\boldsymbol{\eta}\circ\mathbf{x} + \varepsilon\boldsymbol{\eta}\circ\boldsymbol{v}_1) - g(\boldsymbol{\eta}\circ\mathbf{x})\right) = c\varepsilon, \end{aligned}$$

where the last equality follows from $\boldsymbol{\beta}^\top\boldsymbol{v}_1 = \mathrm{sgn}(\beta_i)\boldsymbol{\beta}^\top\mathbf{e}_i = |\beta_i| = 1$ and $\boldsymbol{\eta}\circ\mathbf{e}_i = \mathbf{0}$ for $i \in I$, by the relation $\boldsymbol{\beta}\circ\boldsymbol{\eta} = \mathbf{0}$. Similarly, for any $\mathbf{x} \in \mathbb{R}^n$ and $\varepsilon > 0$, let $\boldsymbol{v}_2 := -\mathrm{sgn}(\beta_i)\mathbf{e}_i$, $i \in I$. Then,

$$f(\mathbf{x}) - f(\mathbf{x}+\varepsilon\boldsymbol{v}_2) = c\left(\boldsymbol{\beta}^\top\mathbf{x} + g(\boldsymbol{\eta}\circ\mathbf{x})\right) - c\left(\boldsymbol{\beta}^\top(\mathbf{x}+\varepsilon\boldsymbol{v}_2) + g(\boldsymbol{\eta}\circ(\mathbf{x}+\varepsilon\boldsymbol{v}_2))\right) = c\varepsilon.$$

Hence, f satisfies (10).

(b) We next verify that f satisfies (10). Denote by $I_1 := \{i \in [n], \beta_i \neq 0\}$ and $I_2 := \{i \in [n], \nu_i \neq 0\}$. It is straightforward to verify that f is Lipschitz continuous with Lipschitz constant c. For any $\mathbf{x} \in \mathbb{R}^n$ with $\boldsymbol{\beta}^\top\mathbf{x} \geqslant 0$ and $\varepsilon > 0$, let $\boldsymbol{v}_1 := \mathrm{sgn}(\beta_i)\mathbf{e}_i$, $i \in I_1$. Then,

$$\begin{aligned} f(\mathbf{x}+\varepsilon\boldsymbol{v}_1) - f(\mathbf{x}) &= c\left(|\boldsymbol{\beta}^\top(\mathbf{x}+\varepsilon\boldsymbol{v}_1)| - |\boldsymbol{\nu}^\top(\mathbf{x}+\varepsilon\boldsymbol{v}_1)| + g(\boldsymbol{\eta}\circ(\mathbf{x}+\varepsilon\boldsymbol{v}_1))\right) - f(\mathbf{x}) \\ &= c\left(\varepsilon\boldsymbol{\beta}^\top\boldsymbol{v}_1 + g(\boldsymbol{\eta}\circ\mathbf{x}+\varepsilon\boldsymbol{\eta}\circ\boldsymbol{v}_1) - g(\boldsymbol{\eta}\circ\mathbf{x}) + |\boldsymbol{\nu}^\top\mathbf{x}| - |\boldsymbol{\nu}^\top(\mathbf{x}+\varepsilon\boldsymbol{v}_1)|\right) = c\varepsilon, \end{aligned}$$

where the second equality follows from that $|\boldsymbol{\beta}^\top\mathbf{x}| = \boldsymbol{\beta}^\top\mathbf{x} \geqslant 0$ and $\boldsymbol{\beta}^\top\boldsymbol{v}_1 = \mathrm{sgn}(\beta_i)\boldsymbol{\beta}^\top\mathbf{e}_i = |\beta_i| = 1$, and the last equality follows from that $\boldsymbol{\nu}^\top\boldsymbol{v}_1 = 0$ and $\boldsymbol{\eta}\circ\boldsymbol{v}_1 = \mathbf{0}$ for $i \in I_1$, by the relations $\boldsymbol{\beta}\circ\boldsymbol{\eta} = \boldsymbol{\nu}\circ\boldsymbol{\eta} = \boldsymbol{\beta}\circ\boldsymbol{\nu} = \mathbf{0}$. Similarly, for any $\mathbf{x} \in \mathbb{R}^n$ with $\boldsymbol{\beta}^\top\mathbf{x} < 0$ and $\varepsilon > 0$, let $\boldsymbol{v}_2 := -\mathrm{sgn}(\beta_i)\mathbf{e}_i$, $i \in I_1$. Then,

$$f(\mathbf{x}+\varepsilon\boldsymbol{v}_2) - f(\mathbf{x}) = c\left(-\varepsilon\boldsymbol{\beta}^\top\boldsymbol{v}_2 + g(\boldsymbol{\eta}\circ\mathbf{x}+\varepsilon\boldsymbol{\eta}\circ\boldsymbol{v}_2) - g(\boldsymbol{\eta}\circ\mathbf{x}) + |\boldsymbol{\nu}^\top\mathbf{x}| - |\boldsymbol{\nu}^\top(\mathbf{x}+\varepsilon\boldsymbol{v}_2)|\right) = c\varepsilon,$$

since $|\boldsymbol{\beta}^\top\mathbf{x}| = -\boldsymbol{\beta}^\top\mathbf{x}$, $\boldsymbol{\beta}^\top\boldsymbol{v}_2 = -\mathrm{sgn}(\beta_i)\boldsymbol{\beta}^\top\mathbf{e}_i = -1$, $\boldsymbol{\nu}^\top\boldsymbol{v}_2 = 0$ and $\boldsymbol{\eta}\circ\boldsymbol{v}_2 = \mathbf{0}$ for $i \in I_1$. Moreover, for any $\mathbf{x} \in \mathbb{R}^n$ with $\boldsymbol{\nu}^\top\mathbf{x} \geqslant 0$ and $\varepsilon > 0$, let $\boldsymbol{v}_3 := \mathrm{sgn}(\nu_i)\mathbf{e}_i$, $i \in I_2$. Here $\boldsymbol{\nu}^\top\boldsymbol{v}_3 = 1$, $\boldsymbol{\beta}^\top\boldsymbol{v}_3 = 0$ and

$\boldsymbol{\eta} \circ \boldsymbol{v}_3 = \mathbf{0}$. Analogous arguments yield

$$f(\mathbf{x}) - f(\mathbf{x} + \varepsilon \boldsymbol{v}_3) = c\varepsilon.$$

For any $\mathbf{x} \in \mathbb{R}^n$ with $\boldsymbol{\nu}^\top \mathbf{x} < 0$ and $\varepsilon > 0$, let $\boldsymbol{v}_4 := -\mathrm{sgn}(\nu_i)\mathbf{e}_i$, $i \in I_2$. Again $\boldsymbol{\nu}^\top \boldsymbol{v}_4 = -1$, $\boldsymbol{\beta}^\top \boldsymbol{v}_4 = 0$ and $\boldsymbol{\eta} \circ \boldsymbol{v}_4 = \mathbf{0}$. Hence $f(\mathbf{x}) - f(\mathbf{x} + \varepsilon \boldsymbol{v}_4) = c\varepsilon$. Thus, we have established that f satisfies (10). This completes the proof. □

Proof of Theorem 2. For (i) ⇒ (ii), choose $\rho = \mathrm{VaR}_\alpha$ for some $\alpha \in [0, 1)$, and invoke Lemma A1. Then, by arguments similar to those in the proof of Theorem 1, direction (i) ⇒ (iii), the desired result follows.

For (ii) ⇒ (i), if $c_f = 0$, then (12) reduces to

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = 0, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, \varepsilon > 0,$$

which implies that f is constant. In this case, (11) holds trivially. We now consider the case where $c_f > 0$. Without loss of generality, assume $c_f = 1$. Suppose that f satisfies (12). Then, for any random vector $\boldsymbol{\xi}$ taking values in $\mathbb{R}^n$, the definition of (12) implies that

$$|f(\boldsymbol{\xi}) - f(\boldsymbol{\zeta})| \leqslant \|\boldsymbol{\xi} - \boldsymbol{\zeta}\|.$$

Hence, if $\mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] \leqslant \varepsilon^p$, then

$$\mathbb{E}\left[|f(\boldsymbol{\xi}) - f(\boldsymbol{\zeta})|^p\right] \leqslant \mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] \leqslant \varepsilon^p.$$

This shows that $\{F_{f(\boldsymbol{\xi})} : F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon)\} \subseteq \mathcal{C}_p(f|F_0, \varepsilon)$. It follows that

$$\sup_{F \in \mathbb{B}_p(F_0, \varepsilon)} \rho^F\left(f(\boldsymbol{\xi})\right) \leqslant \sup_{G \in \mathcal{C}_p(f|F_0, \varepsilon)} \rho^G(X).$$

We now show the reverse direction of (11). It suffices to demonstrate that for any $G \in \mathbb{B}_p(F_{f(\boldsymbol{\zeta})}, \varepsilon)$ and $Z \sim G$ with $\mathbb{E}[|Z - f(\boldsymbol{\zeta})|^p] \leqslant \varepsilon^p$, there exists $F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon)$ and $\boldsymbol{\xi} \sim F_{\boldsymbol{\xi}}$ such that

$$\rho(f(\boldsymbol{\xi})) \geqslant \rho^G(Z).$$

To this end, take $G \in \mathbb{B}_p(F_{f(\boldsymbol{\zeta})}, \varepsilon)$ and $Z \sim G$. Define $Z^* = \max\{f(\boldsymbol{\zeta}), Z\}$ and denote by G^* the distribution of Z^*. We have $\rho^G(Z) \leqslant \rho^{G^*}(Z^*)$ and $\mathbb{E}[|Z^* - f(\boldsymbol{\zeta})|^p] \leqslant \mathbb{E}[|Z - f(\boldsymbol{\zeta})|^p] \leqslant \varepsilon^p$, and thus,

$G^* \in \mathbb{B}_p(F_{f(\boldsymbol{\zeta})}, \varepsilon)$. So, without loss of generality, assume that $Z \geqslant f(\boldsymbol{\zeta})$ almost surely. Denote by $T := Z - f(\boldsymbol{\zeta})$. We have $T \geqslant 0$ almost surely and $\mathbb{E}[T^p] \leqslant \varepsilon^p$. By Rieder (1978) and following similar arguments as in the proof of Theorem 1 (iii) $\Rightarrow$ (ii), there exists a measurable mapping $\mathbf{V}$ such that

$$\mathbf{V}(\omega) \in \underset{\|\mathbf{y}\| \leqslant T(\omega)}{\arg\max} f\left(\boldsymbol{\zeta}(\omega) + \mathbf{y}\right), \quad \omega \in \Omega.$$

Define $\boldsymbol{\xi}(\omega) = \boldsymbol{\zeta}(\omega) + \mathbf{V}(\omega)$, $\omega \in \Omega$, so that $\boldsymbol{\xi}$ is measurable. By (12), one can verify that

$$f(\boldsymbol{\xi}(\omega)) = f(\boldsymbol{\zeta}(\omega) + \mathbf{V}(\omega)) = f(\boldsymbol{\zeta}(\omega)) + T(\omega) = Z(\omega), \quad \omega \in \Omega.$$

Moreover, we have $\mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|^p] = \mathbb{E}[\|\mathbf{V}\|^p] \leqslant \mathbb{E}[T^p] \leqslant \varepsilon^p$, which implies that $F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon)$. Hence

$$\sup_{F \in \{F_{f(\boldsymbol{\xi})}:\ F_{\boldsymbol{\xi}} \in \mathbb{B}_p(F_0, \varepsilon)\}} \rho^F(X) \geqslant \sup_{G \in \mathbb{B}_p(F_{f(\boldsymbol{\zeta})}, \varepsilon)} \rho^G(X),$$

and thus (11) holds, completing the proof. □

Proof of Proposition 3. Since the case $c_f = 0$ is trivial, we consider the case where $c_f > 0$. Without loss of generality, we assume $c_f = 1$. We first show that (12) implies that for any $\mathbf{x} \in \mathbb{R}^n$, the subgradient $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$ satisfies $\|\nabla f\|_* \leqslant 1$, where $\partial f(\mathbf{x})$ denotes the subdifferential of f at $\mathbf{x}$. Since otherwise, suppose that $\|\nabla f(\mathbf{x}_0)\|_* > 1$ for some $\mathbf{x}_0$. Then there exists $\boldsymbol{v}_{\mathbf{x}_0}$ with $\|\boldsymbol{v}_{\mathbf{x}_0}\| = 1$ such that $\boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\|_*$, and there exists $\varepsilon > 0$ small enough such that

$$f(\mathbf{x}_0 + \varepsilon \boldsymbol{v}_{\mathbf{x}_0}) \geqslant f(\mathbf{x}_0) + \varepsilon \boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = f(\mathbf{x}_0) + \varepsilon \|\nabla f(\mathbf{x}_0)\|_* > f(\mathbf{x}_0) + \varepsilon,$$

where the first inequality follows from that f is convex and the strict inequlity follows from $\|\nabla f(\mathbf{x}_0)\|_* > 1$. This yields a contradiction with (12). It follows that all subgradients of f satisfy $\|\nabla f\|_* \leqslant 1$, which in turn implies that f is Lipschitz continuous with $\mathrm{Lip}(f) \leqslant 1$. Next, we assert that for each $\mathbf{x} \in \mathbb{R}^n$, there exists a subgradient $\nabla f(\mathbf{x})$ such that $\|\nabla f(\mathbf{x})\|_* = 1$. As in the proof of Proposition 1, we have from (12) that for every $\mathbf{x} \in \mathbb{R}^n$, there exists $\boldsymbol{\beta}_{\mathbf{x}} \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}_{\mathbf{x}}\| = 1$ such that

$$f(\mathbf{x} + \varepsilon \boldsymbol{\beta}_{\mathbf{x}}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \varepsilon > 0.$$

This implies that the directional derivative of f at $\mathbf{x}$ at direction $\boldsymbol{\beta}_{\mathbf{x}}$ equals to 1. By Theorem 23.4 of Rockafellar (1970), there exist a subgradient $\boldsymbol{\eta}_{\mathbf{x}} \in \partial f(\mathbf{x})$ such that $\boldsymbol{\eta}_{\mathbf{x}}^\top \boldsymbol{\beta}_{\mathbf{x}} = 1$. Since $\|\boldsymbol{\eta}_{\mathbf{x}}\|_* \leqslant 1$

and

$$\boldsymbol{\eta}_{\mathbf{x}}^{\top}\boldsymbol{\beta}_{\mathbf{x}} = 1 \leqslant \|\boldsymbol{\eta}_{\mathbf{x}}\|_*\|\boldsymbol{\beta}_{\mathbf{x}}\| = \|\boldsymbol{\eta}_{\mathbf{x}}\|_*,$$

we conclude that $\|\boldsymbol{\eta}_{\mathbf{x}}\|_* = 1$. Since a convex function f can be written as

$$f(\mathbf{x}) = \max_{\mathbf{z}\in\mathbb{R}^n}\{f(\mathbf{z}) + \boldsymbol{\eta}_{\mathbf{z}}^{\top}(\mathbf{x}-\mathbf{z})\},$$

we have (13) holds and thus, we complete the proof. □

Proof of Proposition 4. For the "if" part, by Theorem 2, we can take $\rho = \mathrm{VaR}_\alpha$ for some $\alpha \in [0,1)$, which implies that (14) holds.

For the "only if" part, applying Lemma A1 and following arguments similar to those in the proof of Theorem 1, direction (i) ⇒ (iii), the desired result follows. This completes the proof. □

Proof of Lemma 1. Let $G_0 \in \mathcal{M}(\mathbb{R})$. Note that $f(\boldsymbol{\xi}) = \max_{i\in I}\{\boldsymbol{\beta}_i^{\top}\boldsymbol{\xi}/\|\boldsymbol{\beta}_i\|_*\}$ is a convex isometric function. By Propositions 3 and 4, equation (15) holds. Hence, it suffices to the worst case VaR over the one-dimensional Wasserstein ball $\sup_{F\in\mathbb{B}_p(G_0,\varepsilon)}\mathrm{VaR}_\alpha^F(X)$. By Proposition 4 of Liu et al. (2022), we have that $\sup_{F\in\mathbb{B}_p(G_0,\varepsilon)}\mathrm{VaR}_\alpha^F(X)$ equals to the unique solution to

$$\int_\alpha^1 (x - \mathrm{VaR}_u^{G_0}(X))_+^p \mathrm{d}u = \varepsilon^p. \tag{A10}$$

Note that $\mathrm{VaR}_u(h(X)) = h(\mathrm{VaR}_u(X))$ for any increasing and continuous function h. By taking $h(t) = -(x-t)_+^p$ which is an increasing and continuous function in $t \in \mathbb{R}$, the above equation is equivalent to

$$\int_\alpha^1 -\mathrm{VaR}_u^{G_0}\left(-(x-X)_+^p\right)\mathrm{d}u = \varepsilon^p.$$

It follows that (16) holds. To show the strict monotonicity, for $\varepsilon > 0$ and $\alpha_1 < \alpha_2$, denote by $x_i^* = \sup_{F\in\mathbb{B}_p(G_0,\varepsilon)}\mathrm{VaR}_{\alpha_i}^F(X)$, i.e., the solution to (A10) with $\alpha = \alpha_i$, $i = 1,2$. We have $x_2^* > \mathrm{VaR}_{\alpha_2}^{G_0}(X) \geqslant \mathrm{VaR}_u^{G_0}(X)$ for $u \in (\alpha_1,\alpha_2)$, and thus,

$$\int_{\alpha_1}^1 (x_2^* - \mathrm{VaR}_u^{G_0}(X))_+^p \mathrm{d}u < \int_{\alpha_2}^1 (x_2^* - \mathrm{VaR}_u^{G_0}(X))_+^p \mathrm{d}u = \varepsilon^p.$$

This implies $x_1^* < x_2^*$, that is, $\sup_{F\in\mathbb{B}_p(G_0,\varepsilon)}\mathrm{VaR}_\alpha^F(X)$ is strictly increasing in $\alpha \in [0,1]$. This completes the proof. □

To show Theorem 3, we need the following lemma.

Lemma A2. *For $\alpha \in [0,1)$ and $\varepsilon > 0$, if f is a Lipschitz continuous function and there exist $\mathbf{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{v}_k \in \mathbb{R}^n$ with $\|\boldsymbol{v}_k\| = 1$, $k \in \mathbb{N}$ such that (21) holds, then*

$$\sup_{F \in \mathbb{B}_1(F_0,\varepsilon)} \mathrm{CVaR}_\alpha^F\left(f\left(\boldsymbol{\xi}\right)\right) = \sup_{G \in \mathcal{C}_1(f|F_0,\, c_f\varepsilon)} \mathrm{CVaR}_\alpha^G(X) = \mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon c_f}{1-\alpha} \tag{A11}$$

holds for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ with $c_f = \mathrm{Lip}(f)$.

Proof. Note that the second equality in (A11) follows from Proposition 2 in Pflug et al. (2012). To establish the first equality, it suffices to show that

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|] \leqslant \varepsilon} \mathrm{CVaR}_\alpha\left(f\left(\boldsymbol{\xi}\right)\right) = \mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon c_f}{1-\alpha} \tag{A12}$$

holds for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$. To see it, we assume without loss generality that $\mathrm{Lip}(f) = 1$. First note that

$$\begin{aligned}
\sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\| \leqslant \varepsilon} \mathrm{CVaR}_\alpha\left(f\left(\boldsymbol{\xi}\right)\right) &\leqslant \sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\| \leqslant \varepsilon} \left\{ \mathrm{CVaR}_\alpha\left(f\left(\boldsymbol{\zeta}\right)\right) + \frac{\mathbb{E}[|f(\boldsymbol{\xi}) - f(\boldsymbol{\zeta})|]}{1-\alpha} \right\} \\
&\leqslant \sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\| \leqslant \varepsilon} \left\{ \mathrm{CVaR}_\alpha\left(f\left(\boldsymbol{\zeta}\right)\right) + \frac{\mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|]}{1-\alpha} \right\} \\
&\leqslant \mathrm{CVaR}_\alpha\left(f\left(\boldsymbol{\zeta}\right)\right) + \frac{\varepsilon}{1-\alpha},
\end{aligned} \tag{A13}$$

where the first inequality follows from the Lipschitz property of CVaR, and the second from the 1-Lipschitz continuity of f. Let us now verify the other direction. By assumption, for $\mathbf{x}_0 \in \mathbb{R}^n$ define $s(\boldsymbol{v}) := \limsup_{m\to\infty}(f(\mathbf{x}_0 + m\boldsymbol{v}) - f(\mathbf{x}_0))/m < \infty$. There exist $\boldsymbol{v}_k$ with $\|\boldsymbol{v}_k\| = 1$, $k \in \mathbb{N}$, and $\mathbf{x}_0 \in \mathbb{R}^n$ such that $\lim_{k\to\infty} s(\boldsymbol{v}_k) = \mathrm{Lip}(f) = 1$. Hence, for any $\eta > 0$ there is $K > 0$ with $s(\boldsymbol{v}_k) > 1 - \eta/2$ for all $k > K$. For each $k > K$ and any $M > 0$, one can then find $m > M$ such that $(f(\mathbf{x}_0 + m\boldsymbol{v}_k) - f(\mathbf{x}_0))/m \geqslant 1 - \eta$. Now set $\eta_j := 1/j$ for $j \in \mathbb{N}$. Choosing $k_j > K_j$ and $m_j > \max\{m_{j-1}, j, \varepsilon/(1-\alpha)\}$ sufficiently large with $m_j \to \infty$, we obtain $(f(\mathbf{x}_0 + m_j\boldsymbol{v}_{k_j}) - f(\mathbf{x}_0))/m_j \geqslant 1 - \eta_j$. Then, for any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\begin{aligned}
f(\mathbf{x} + m_j\boldsymbol{v}_{k_j}) - f(\mathbf{x}) &= f(\mathbf{x} + m_j\boldsymbol{v}_{k_j}) - f(\mathbf{x}_0 + m_j\boldsymbol{v}_{k_j}) + f(\mathbf{x}_0 + m_j\boldsymbol{v}_{k_j}) - f(\mathbf{x}_0) + f(\mathbf{x}_0) - f(\mathbf{x}) \\
&\geqslant m_j(1-\eta_j) - 2\|\mathbf{x} - \mathbf{x}_0\|,
\end{aligned}$$

where the inequality uses the fact that f is 1-Lipschitz continuous. Denote by U a uniform random variable on $[0,1]$ such that U and $f(\boldsymbol{\zeta})$ are comonotonic. For chosen η_j, k_j, and m_j, define $\boldsymbol{\xi}_j = m_j\boldsymbol{v}_{k_j}\mathbb{1}_{A_j} + \boldsymbol{\zeta}$, where $A_j := \{\omega : 1 - \varepsilon/m_j < U(\omega) \leqslant 1\}$. One can verify that $\mathbb{E}\left[\|\boldsymbol{\xi}_j - \boldsymbol{\zeta}\|\right] = \varepsilon$. We

next show that $\mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}_j)) \to \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta})) + \varepsilon/(1-\alpha)$ as $j \to \infty$. Note that for m_j sufficiently large, we have $f(\boldsymbol{\zeta} + m_j \boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}) \geqslant m_j(1-\eta_j) - 2\|\boldsymbol{\zeta} - \mathbf{x}_0\|$, and

$$\begin{aligned}
\mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}_j)) - \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta})) &= \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta} + m_j \boldsymbol{v}_{k_j})\mathbb{1}_{A_j} + f(\boldsymbol{\zeta})\mathbb{1}_{A_j^c}) - \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta})) \\
&= \mathrm{CVaR}_\alpha((f(\boldsymbol{\zeta} + m_j \boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}))\mathbb{1}_{A_j} + f(\boldsymbol{\zeta})) - \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta})) \\
&\geqslant \mathbb{E}\left[(f(\boldsymbol{\zeta} + m_j \boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}))\mathbb{1}_{A_j}(U)\frac{\mathbb{1}_{[\alpha,1]}(U)}{1-\alpha}\right] \\
&\geqslant (1-\eta_j)\mathbb{E}\left[m_j \mathbb{1}_{A_{m_j}}(U)\frac{\mathbb{1}_{[\alpha,1]}(U)}{1-\alpha}\right] - 2\mathbb{E}\left[\|\boldsymbol{\zeta} - \mathbf{x}_0\|\mathbb{1}_{A_j}(U)\frac{\mathbb{1}_{[\alpha,1]}(U)}{1-\alpha}\right] \\
&= (1-\eta_j)\frac{\int_{1-\varepsilon/m_j}^1 \frac{\mathbb{1}_{[\alpha,1]}(s)}{1-\alpha}\mathrm{d}s}{\varepsilon/m_j}\varepsilon - 2\mathbb{E}\left[\|\boldsymbol{\zeta} - \mathbf{x}_0\|\mathbb{1}_{A_j}(U)\frac{\mathbb{1}_{[\alpha,1]}(U)}{1-\alpha}\right] \\
&\to \frac{\varepsilon}{1-\alpha} \quad \text{as } j \to \infty,
\end{aligned}$$

where the first inequality uses the definition and dual representation of CVaR_α (see, e.g., Theorem 4.79 in Föllmer and Schied (2016)), and the last limit follows from $\mathbb{E}[\|\boldsymbol{\zeta} - \mathbf{x}_0\|\mathbb{1}_{A_j}(U)\mathbb{1}_{[\alpha,1]}(U)/(1-\alpha)] \to 0$ as $j \to \infty$ by the dominated convergence theorem. It follows that

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|\leqslant\varepsilon} \mathrm{CVaR}_\alpha\left(f\left(\boldsymbol{\xi}\right)\right) \geqslant \liminf_{j\to\infty} \mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}_j)) \geqslant \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta})) + \varepsilon/(1-\alpha).$$

Combining this with (A13) yields (A12). This completes the proof. □

Proof of Theorem 3. (i) Without loss generality assume $\mathrm{Lip}(f) = 1$. First note that

$$\begin{aligned}
\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \rho^F\left(f\left(\boldsymbol{\xi}\right)\right) &= \sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|]\leqslant\varepsilon} \sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \mathrm{CVaR}_\alpha(f\left(\boldsymbol{\xi}\right))\mathrm{d}\mu(\alpha) \\
&= \sup_{\mu\in\mathscr{M}_\rho} \sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|\leqslant\varepsilon} \int_0^1 \mathrm{CVaR}_\alpha(f\left(\boldsymbol{\xi}\right))\mathrm{d}\mu(\alpha) \\
&\leqslant \sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|]\leqslant\varepsilon} \mathrm{CVaR}_\alpha(f\left(\boldsymbol{\xi}\right))\mathrm{d}\mu(\alpha) \\
&= \sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \left(\mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon}{1-\alpha}\right)\mathrm{d}\mu(\alpha), \qquad \text{(A14)}
\end{aligned}$$

where the third equality follows from Lemma A2. We now show that the inequality in (A14) is an equality. By assumption, for each $\eta_j := 1/j$, $j \in \mathbb{N}$, we can choose $k_j > K_j$ for some $K_j > 0$ and $m_j > \max\{m_{j-1}, j, \varepsilon/(1-\alpha)\}$, with $m_j \to \infty$, such that $(f(\mathbf{x}_0 + m_j \boldsymbol{v}_{k_j}) - f(\mathbf{x}_0))/m_j \geqslant 1 - \eta_j$. Denote by U a uniform random variable on $[0,1]$ such that U and $f(\boldsymbol{\zeta})$ are comonotonic. For chosen η_j, k_j, and m_j, define $\boldsymbol{\xi}_j = m_j \boldsymbol{v}_{k_j}\mathbb{1}_{A_j} + \boldsymbol{\zeta}$, where $A_j := \{\omega : 1 - \varepsilon/m_j < U(\omega) \leqslant 1\}$. With similar

arguments as in the proof of Lemma A2, we obtain

$$\liminf_{j\to\infty} \mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}_j)) \geqslant \mathrm{CVaR}_\alpha(f(\boldsymbol{\zeta})) + \varepsilon/(1-\alpha).$$

Therefore, for any $\mu \in \mathscr{M}_\rho$,

$$\begin{aligned}\sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|\leqslant\varepsilon} \int_0^1 \mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}))\mathrm{d}\mu(\alpha) &\geqslant \sup_{j\in\mathbb{N}} \int_0^1 \mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}_j))\mathrm{d}\mu(\alpha)\\ &\geqslant \int_0^1 \liminf_{j\to\infty} \mathrm{CVaR}_\alpha^F(f(\boldsymbol{\xi}_j))\mathrm{d}\mu(\alpha)\\ &\geqslant \int_0^1 \left(\mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon}{1-\alpha}\right)\mathrm{d}\mu(\alpha),\end{aligned}$$

where the first inequality follows from $\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\xi}_j\|] \leqslant \varepsilon$ and the second inequality follows from Fatou's Lemma. It then follows that

$$\sup_{\mu\in\mathscr{M}_\rho}\sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|\leqslant\varepsilon} \int_0^1 \mathrm{CVaR}_\alpha(f(\boldsymbol{\xi}))\mathrm{d}\mu(\alpha) \geqslant \sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \left(\mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon}{1-\alpha}\right)\mathrm{d}\mu(\alpha).$$

This means that the inequality of (A14) is actually an equality. That is,

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \rho^F(f(\boldsymbol{\xi})) = \sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \left(\mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon}{1-\alpha}\right)\mathrm{d}\mu(\alpha). \tag{A15}$$

Similarly, one can prove that

$$\begin{aligned}\sup_{G\in\mathcal{C}_1(f|F_0,\varepsilon)} \rho^G(X) &= \sup_{\mathbb{E}\|X-f(\boldsymbol{\zeta})\|\leqslant\varepsilon}\sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \mathrm{CVaR}_\alpha(X)\mathrm{d}\mu(\alpha)\\ &= \sup_{\mu\in\mathscr{M}_\rho} \int_0^1 \left(\mathrm{CVaR}_\alpha^{F_0}(f(\boldsymbol{\zeta})) + \frac{\varepsilon}{1-\alpha}\right)\mathrm{d}\mu(\alpha).\end{aligned} \tag{A16}$$

Combining (A15) and (A16), we have (22) holds. This completes the proof of (i).

(ii) To prove necessity, we first show that f is Lipschitz continuous. Suppose that f is not Lipschitz continuous. Then there exist $\mathbf{x}_n$, $\mathbf{y}_n$, $n \in \mathbb{N}$, such that $|f(\mathbf{x}_n) - f(\mathbf{y}_n)| > n\|\mathbf{x}_n - \mathbf{y}_n\|$, $n \in \mathbb{N}$. Without loss of generality assume $f(\mathbf{x}_n) - f(\mathbf{y}_n) > n\|\mathbf{x}_n - \mathbf{y}_n\|$, $n \in \mathbb{N}$. Take $F_n = \delta_{\mathbf{y}_n}$ and $\varepsilon_n = \|\mathbf{x}_n - \mathbf{y}_n\|$. It follows that

$$\begin{aligned}\sup_{F\in\mathbb{B}_1(F_n,\varepsilon_n)} \rho^F(f(\boldsymbol{\xi})) &= \sup_{\mathbb{E}[\|\boldsymbol{\xi}-\mathbf{y}_n\|]\leqslant\varepsilon_n} \rho(f(\boldsymbol{\xi}))\\ &\geqslant \sup_{\|\mathbf{x}-\mathbf{y}_n\|\leqslant\varepsilon_n} f(\mathbf{x}) \geqslant f(\mathbf{x}_n) > f(\mathbf{y}_n) + n\varepsilon_n.\end{aligned} \tag{A17}$$

Note that

$$\sup_{G\in\mathcal{C}_1(f|F_n,c_f\varepsilon_n)}\rho^G(X)=\sup_{\mathbb{E}[|X-f(\mathbf{y}_n)|]\leqslant c_f\varepsilon_n}\rho(X)\leqslant\rho(f(\mathbf{y}_n))+C_\rho c_f\varepsilon_n,$$

where $C_\rho:=\sup_{\mu\in\mathscr{M}_\rho}\int_0^1\frac{1}{1-\alpha}\mathrm{d}\mu(\alpha)$. This, together with (A17), yields a contradiction with (22) by noting n can be arbitrarily large. Therefore, f is a Lipschitz continuous function.

We next show that (21) holds. Without loss of generality, assume $\mathrm{Lip}(f)=1$. Since $\mathcal{D}$ is compact, for each $\mathbf{x}_0\in\mathbb{R}^n$ and any $\boldsymbol{v}\in\mathbb{R}^n$ with $\|\boldsymbol{v}\|=1$, we have $\mathbf{x}_0+t\boldsymbol{v}\notin\mathcal{D}$ for all $t>\sup_{\{\mathbf{z}\in\mathcal{D}\}}\|\mathbf{z}\|+\|\mathbf{x}_0\|+1$. Hence, for a fixed $\mathbf{x}_0\in\mathbb{R}^n$, one can choose $T>0$ such that, for all $t>T$ sufficiently large,

$$\frac{f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0\right)}{t}=\frac{f\left(\mathbf{x}_0+T\boldsymbol{v}\right)-f\left(\mathbf{x}_0\right)}{t}+\frac{t-T}{t}\frac{f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0+T\boldsymbol{v}\right)}{t-T}.\tag{A18}$$

Define $\phi(\boldsymbol{v}):=\lim_{t\to\infty}(f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0\right))/t$. Since $(f(\mathbf{x}_0+T\boldsymbol{v})-f(\mathbf{x}_0))/t\to 0$ as $t\to\infty$ and f coincides with a convex function on $\mathbb{R}^n\setminus\mathcal{D}$, it follows that $\phi(\boldsymbol{v})$ is well defined, and $\lim_{t\to\infty}\frac{f(\mathbf{x}_0+t\boldsymbol{v})-f(\mathbf{x}_0+T\boldsymbol{v})}{t-T}=\phi(\boldsymbol{v})$. Moreover, by (A18), for any $t>T$ and $\boldsymbol{v}$ with $\|\boldsymbol{v}\|=1$,

$$\frac{f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0\right)}{t}\leqslant\frac{T}{t}+\frac{t-T}{t}\lim_{t\to\infty}\frac{f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0+T\boldsymbol{v}\right)}{t-T}\leqslant\frac{T}{t}+\phi(\boldsymbol{v}),$$

where the first inequality uses that f is 1-Lipschitz and coincides with a convex function outside $\mathcal{D}$, implying that $(f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0+T\boldsymbol{v}\right))/(t-T)$ is increasing in $t>T$. Thus, for all $t>T$,

$$\sup_{\|\boldsymbol{v}\|=1}\frac{f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0\right)}{t}\leqslant\frac{T}{t}+\sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v}).\tag{A19}$$

We next show that $\sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v})=\mathrm{Lip}(f)=1$. Suppose, on the contrary, that $\sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v})\leqslant 1-2\delta$ for some $\delta>0$. Then, by (A19), there exists $T_1:=\max\{T,T/\delta\}$ such that $f(\mathbf{x}_0+t\boldsymbol{v})-f(\mathbf{x}_0)\leqslant(1-\delta)t$ for all $t>T_1$ and $\|\boldsymbol{v}\|=1$. Let

$$B:=\max\left\{0,\sup_{\{\|\boldsymbol{v}\|=1,\,0<t\leqslant T_1\}}\{f\left(\mathbf{x}_0+t\boldsymbol{v}\right)-f\left(\mathbf{x}_0\right)-(1-\delta)t\}\right\}<\infty.$$

For any $t>0$ and $\boldsymbol{v}\in\mathbb{R}^n$ with $\|\boldsymbol{v}\|=1$, we have

$$f(\mathbf{x}_0+t\boldsymbol{v})\leqslant f(\mathbf{x}_0)+B+(1-\delta)t.$$

Take $F_0 = \delta_{\mathbf{x}_0}$ and $\rho = \mathrm{CVaR}_\alpha$ with $\alpha \in (0,1)$. For $\varepsilon > (1-\alpha)B/\delta$ sufficiently large,

$$\begin{aligned}\sup_{\mathbb{E}\|\boldsymbol{\xi}-\mathbf{x}_0\|\leqslant\varepsilon} \mathrm{CVaR}_\alpha(f(\boldsymbol{\xi})) &\leqslant \sup_{\mathbb{E}\|\boldsymbol{\xi}-\mathbf{x}_0\|\leqslant\varepsilon} \mathrm{CVaR}_\alpha(f(\mathbf{x}_0)+B+(1-\delta)\|\boldsymbol{\xi}-\mathbf{x}_0\|)\\ &= f(\mathbf{x}_0)+B+\sup_{\mathbb{E}\|\boldsymbol{\xi}-\mathbf{x}_0\|\leqslant\varepsilon}(1-\delta)\,\mathrm{CVaR}_\alpha(\|\boldsymbol{\xi}-\mathbf{x}_0\|)\\ &\leqslant f(\mathbf{x}_0)+B+\frac{1-\delta}{1-\alpha}\varepsilon < f(\mathbf{x}_0)+\frac{\varepsilon}{1-\alpha} = \sup_{\mathbb{E}\|X-f(\mathbf{x}_0)\|\leqslant\varepsilon}\mathrm{CVaR}_\alpha(X),\end{aligned}$$

where the first inequality follows from monotonicity of CVaR, the equality from its translation invariance and positive homogeneity, and the last inequality from $\mathrm{CVaR}_\alpha(X) \leqslant \mathbb{E}[X]/(1-\alpha)$ for all $X \geqslant 0$. This contradicts (22). Thus, $\sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v}) = 1$, i.e., there exists a sequence $\boldsymbol{v}_k$ with $\|\boldsymbol{v}_k\| = 1$ such that $\lim_{k\to\infty}\phi(\boldsymbol{v}_k) = \sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v}) = 1$. This completes the proof of (ii).

(ii) For the "only if" part, by arguments similar to those in the proof of (ii), one can verify that f is Lipschitz continuous with $\mathrm{Lip}(f) \in \mathbb{R}_+$.

Next, consider the "if" part. Without loss of generality, assume that $\mathrm{Lip}(f) = 1$. By (i), it suffices to show that there exist $\mathbf{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^n$ with $\|\boldsymbol{v}\| = 1$ such that $\lim_{m\to\infty}(f(\mathbf{x}_0+m\boldsymbol{v})-f(\mathbf{x}_0))/m = 1$. For any fixed $\mathbf{x}_0 \in \mathbb{R}^n$, define $\phi(\boldsymbol{v}) := \lim_{m\to\infty}(f(\mathbf{x}_0+m\boldsymbol{v})-f(\mathbf{x}_0))/m$ which is well-defined as f is convex. For any $\boldsymbol{v},\boldsymbol{w} \in \mathbb{R}^n$, we have

$$\begin{aligned}\phi(\boldsymbol{v})-\phi(\boldsymbol{w}) &= \lim_{m\to\infty}\frac{f(\mathbf{x}_0+m\boldsymbol{v})-f(\mathbf{x}_0)}{m} - \lim_{m\to\infty}\frac{f(\mathbf{x}_0+m\boldsymbol{w})-f(\mathbf{x}_0)}{m}\\ &= \lim_{m\to\infty}\frac{f(\mathbf{x}_0+m\boldsymbol{v})-f(\mathbf{x}_0+m\boldsymbol{w})}{m} \leqslant \|\boldsymbol{v}-\boldsymbol{w}\|,\end{aligned}$$

showing that ϕ is 1-Lipschitz. We now claim that $\sup_{\{\boldsymbol{v}:\|\boldsymbol{v}\|=1\}}\phi(\boldsymbol{v}) = \mathrm{Lip}(f) = 1$. Indeed, by definition $\phi(\boldsymbol{v}) \leqslant 1$ for all $\boldsymbol{v}$ with $\|\boldsymbol{v}\| = 1$, so $\sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v}) \leqslant 1$. Conversely, by the convexity of f, for any $\mathbf{x} \in \mathbb{R}^n$, $\boldsymbol{v}$ with $\|\boldsymbol{v}\| = 1$, and $g \in \partial f(\mathbf{x})$, it holds that

$$f(\mathbf{x}+m\boldsymbol{v}) \geqslant f(\mathbf{x}) + mg^\top\boldsymbol{v}.$$

Since $\lim_{m\to\infty}(f(\mathbf{x}+m\boldsymbol{v})-f(\mathbf{x}))/m = \phi(\boldsymbol{v})$ for all $\mathbf{x} \in \mathbb{R}^n$, it follows upon dividing by m and taking limits that $\sup_{\{\boldsymbol{v}:\|\boldsymbol{v}\|=1\}}\phi(\boldsymbol{v}) \geqslant \sup_{\{\boldsymbol{v}:\|\boldsymbol{v}\|=1\}} g^\top\boldsymbol{v} = \|g\|_*$. Taking the supremum over $\mathbf{x}$ and $g \in \partial f(\mathbf{x})$ yields

$$\sup_{\{\boldsymbol{v}:\|\boldsymbol{v}\|=1\}}\phi(\boldsymbol{v}) \geqslant \sup_{\mathbf{x}}\sup_{g\in\partial f(\mathbf{x})}\|g\|_* = \mathrm{Lip}(f) = 1.$$

Thus, $\sup_{\|\boldsymbol{v}\|=1}\phi(\boldsymbol{v}) = 1$. Since ϕ is Lipschitz continuous and the unit sphere $\{\boldsymbol{v} : \|\boldsymbol{v}\| = 1\}$ is compact, the supremum is attained at some $\boldsymbol{v}_0$. Therefore, $\sup_{\{\boldsymbol{v}:\|\boldsymbol{v}\|=1\}}\phi(\boldsymbol{v}) = \lim_{m\to\infty}(f(\mathbf{x}_0+$

$m\boldsymbol{v}_0) - f(\mathbf{x}_0))/m = 1$. This completes the proof. □

A.1 Proofs for Section 4

To prove Theorem 4, we need the following lemma from Wu et al. (2022).

Lemma A3. *Let $p \in (1,\infty)$, t, $\varepsilon > 0$, and $\eta \in (0,\varepsilon)$. For $V \in L^p$, the following statements hold.*

(i) If $\|V\|_p \leqslant \varepsilon$, then $\mathbb{E}\left[(|V|+t)^p\right] \leqslant (\varepsilon+t)^p$.

(ii) If $\|V\|_p \leqslant \varepsilon$ and $\mathbb{E}[|V|] \leqslant \varepsilon - \eta$, then there exists $\Delta > 0$ that only depends on p, t, ε, η such that $\mathbb{E}\left[(|V|+t)^p\right] \leqslant (\varepsilon+t)^p - \Delta$. In particular, if p is an integer, then $\mathbb{E}\left[(|V|+t)^p\right] \leqslant (\varepsilon+t)^p - pt^{p-1}\eta$.

Proof of Theorem 4. For the "if" part, by Theorem 3, we have

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \left(\mathbb{E}^F\left[f^p(\boldsymbol{\xi})\right]\right)^{1/p} = \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \left(\mathbb{E}^F\left[(g(\boldsymbol{\xi}))_+^p\right]\right)^{1/p} = \sup_{G\in\mathbb{B}_p(G_{g(\boldsymbol{\zeta})},c_f\varepsilon)} \left(\mathbb{E}^G[X_+^p]\right)^{1/p},$$

where $g(\mathbf{x}) := \max_{i\in I}\{c_f\boldsymbol{\beta}_i^\top\mathbf{x} + b_i\}$, and $G_{g(\boldsymbol{\zeta})}$ is the distribution of $g(\boldsymbol{\zeta})$ with $\boldsymbol{\zeta} \sim F_0$. Note that $\mathbb{B}_p(G_{g(\boldsymbol{\zeta})}, c_f\varepsilon)$ is a one-dimensional Wasserstein ball with center $G_{g(\boldsymbol{\zeta})}$ and thus, by Theorem 4 of Wu et al. (2022), we have

$$\sup_{G\in\mathcal{C}_p(g|F_0,c_f\varepsilon)} \left(\mathbb{E}^G[(X)_+^p]\right)^{1/p} = \left(\mathbb{E}^{F_0}\left[(g(\boldsymbol{\zeta}))_+^p\right]\right)^{1/p} + c_f\varepsilon.$$

Combining the two equations yields (23).

For the "only if" part, if $c_f = 0$, then (23) is equivalent to

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p]\leqslant\varepsilon^p} \left(\mathbb{E}\left[f^p\left(\boldsymbol{\xi}+\boldsymbol{\zeta}\right)\right]\right)^{1/p} = \left(\mathbb{E}^{F_0}\left[f^p\left(\boldsymbol{\zeta}\right)\right]\right)^{1/p}. \tag{A20}$$

We next show that $\|\nabla f(\mathbf{x})\|_* = 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$. Suppose, for contradiction, that there exists $\mathbf{x}_0$ such that $\|\nabla f(\mathbf{x}_0)\|_* > 0$. Then, there exists $\boldsymbol{v}_{\mathbf{x}_0}$ with $\|\boldsymbol{v}_{\mathbf{x}_0}\| = 1$ such that $\boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\|_*$. Let $F_0 := \delta_{\mathbf{x}_0}$ and consider $\boldsymbol{\xi} \sim \delta_{\varepsilon\boldsymbol{v}_{\mathbf{x}_0}}$ for some sufficiently small $\varepsilon > 0$. Then $\mathbb{E}[\|\boldsymbol{\xi}\|^p] = \varepsilon^p$, and we have

$$\begin{aligned}\sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p]\leqslant\varepsilon^p} \left(\mathbb{E}\left[f^p\left(\boldsymbol{\xi}+\mathbf{x}_0\right)\right]\right)^{1/p} &\geqslant \mathbb{E}\left[f^p(\varepsilon\boldsymbol{v}_{\mathbf{x}_0}+\mathbf{x}_0)\right]^{1/p} \\ &\geqslant f(\mathbf{x}_0) + \varepsilon\boldsymbol{v}_{\mathbf{x}_0}^\top\nabla f(\mathbf{x}_0) = f(\mathbf{x}_0) + \varepsilon\|\nabla f(\mathbf{x}_0)\|_* > f(\mathbf{x}_0),\end{aligned}$$

where the second inequality follows from that f is convex, and the strict inequality follows from $\|\nabla f(\mathbf{x}_0)\|_* > 0$ and $\varepsilon > 0$. This yields a contradiction with (A20). Thus, we have $\|\nabla f(\mathbf{x})\|_* = 0$ for all $\mathbf{x} \in \mathbb{R}^n$, which implies that f is constant. Thus, there exists $b \geqslant 0$ such that $f(\mathbf{x}) \equiv b$ for $\mathbf{x} \in \mathbb{R}^n$. Now consider the case $c_f > 0$. We assume without loss of generality that $c_f = 1$. Note that, (23) is equivalent to

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left(\mathbb{E}\left[f^p\left(\boldsymbol{\xi} + \boldsymbol{\zeta}\right)\right]\right)^{1/p} = \left(\mathbb{E}^{F_0}\left[f^p\left(\boldsymbol{\zeta}\right)\right]\right)^{1/p} + \varepsilon \tag{A21}$$

for any $\boldsymbol{\zeta} \sim F_0$ and $\varepsilon > 0$. We next show that for any $\mathbf{x} \in \mathbb{R}^n$, the subgradient of f at $\mathbf{x}$ satisfies $\|\nabla f\|_* \leqslant 1$. Suppose, for contradiction, that there exists $\mathbf{x}_0 \in \mathbb{R}^n$ such that $\|\nabla f(\mathbf{x}_0)\|_* > 1$. Then, by the definition of the dual norm, there exists $\boldsymbol{\eta}_{\mathbf{x}_0} \in \mathbb{R}^n$ with $\|\boldsymbol{\eta}_{\mathbf{x}_0}\| = 1$ such that $\boldsymbol{\eta}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\|_*$. For $\varepsilon > 0$, let $\boldsymbol{\xi}_0 := \varepsilon \boldsymbol{\eta}_{\mathbf{x}_0}$. One can verify that $\mathbb{E}[\|\boldsymbol{\xi}_0\|^p] = \varepsilon^p$ and

$$\begin{aligned}\sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left(\mathbb{E}\left[f^p\left(\boldsymbol{\xi} + \mathbf{x}_0\right)\right]\right)^{1/p} &\geqslant f(\mathbf{x}_0 + \boldsymbol{\xi}_0) \geqslant f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \boldsymbol{\xi}_0 \\ &= f(\mathbf{x}_0) + \varepsilon \|\nabla f(\mathbf{x}_0)\|_* > f(\mathbf{x}_0) + \varepsilon,\end{aligned}$$

where the second inequality follows from the convexity of f, and the strict inequality from $\|\nabla f(\mathbf{x}_0)\|_* > 1$ and $\varepsilon > 0$. This contradicts (A21). Therefore, $\|\nabla f(\mathbf{x})\|_* \leqslant 1$ for all $\mathbf{x} \in \mathbb{R}^n$. Since f is convex, it follows that f is Lipschitz continuous with $\mathrm{Lip}(f) \leqslant 1$. To establish that f satisfies (24), it suffices to verify that

$$\text{if } \ f(\mathbf{x}) > 0, \quad \text{then } \ \|\nabla f(\mathbf{x})\|_* = 1; \ \text{otherwise } \ f(\mathbf{x}) = 0,$$

or, equivalently,

$$\text{if } \ f(\mathbf{x}) > 0, \quad \text{then } \ \sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \varepsilon, \ \forall \varepsilon > 0; \ \text{otherwise } \ f(\mathbf{x}) = 0. \tag{A22}$$

We assume by contradiction that there exists $\mathbf{x}_0 \in \mathbb{R}$ and $\varepsilon_0 > 0$ such that $f(\mathbf{x}_0) > 0$ and

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon_0} f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0) < \varepsilon_0. \tag{A23}$$

For sufficiently small $\varepsilon > 0$ such that $2\varepsilon \leqslant \varepsilon_0$, define

$$k = \sup_{\|\mathbf{y}\| \leqslant 2\varepsilon} \left\{ \frac{f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0)}{2\varepsilon} \right\}.$$

By the strict inequality of (A23) and the convexity of f, it follows that $k < 1$. Observe that

$$\sup_{\|\mathbf{y}\| \leqslant 2\varepsilon} \left\{ \frac{f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0)}{\|\mathbf{y}\|} \right\} \leqslant \sup_{\|\mathbf{y}\| \leqslant 2\varepsilon} \left\{ \frac{f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0)}{2\varepsilon} \right\} = k,$$

where the inequality again follows from the convexity of f. This implies that $f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0) \leqslant k\|\mathbf{y}\|$ for all $\|\mathbf{y}\| \leqslant 2\varepsilon$. Further, note that

$$\left(f\left(\mathbf{x}_0\right) + \|\mathbf{y}\|\right)^p - \left(f\left(\mathbf{x}_0\right) + k\|\mathbf{y}\|\right)^p \geqslant p f^{p-1}\left(\mathbf{x}_0\right)(1-k)\|\mathbf{y}\|, \quad \forall\, \mathbf{y} \in \mathbb{R}^n. \tag{A24}$$

Therefore,

$$\begin{aligned}
&\sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \mathbb{E}\left[f^p\left(\mathbf{x}_0 + \boldsymbol{\xi}\right)\right] \\
&= \sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left\{ \mathbb{E}\left[f^p\left(\mathbf{x}_0 + \boldsymbol{\xi}\right) \mathbb{1}_{\{\|\boldsymbol{\xi}\| \leqslant 2\varepsilon\}}\right] + \mathbb{E}\left[f^p\left(\mathbf{x}_0 + \boldsymbol{\xi}\right) \mathbb{1}_{\{\|\boldsymbol{\xi}\| > 2\varepsilon\}}\right] \right\} \\
&\leqslant \sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left\{ \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + k\|\boldsymbol{\xi}\|\right)^p \mathbb{1}_{\{\|\boldsymbol{\xi}\| \leqslant 2\varepsilon\}}\right] + \mathbb{E}\left[f^p\left(\mathbf{x}_0 + \boldsymbol{\xi}\right) \mathbb{1}_{\{\|\boldsymbol{\xi}\| > 2\varepsilon\}}\right] \right\} \\
&\leqslant \sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left\{ \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + k\|\boldsymbol{\xi}\|\right)^p \mathbb{1}_{\{\|\boldsymbol{\xi}\| \leqslant 2\varepsilon\}}\right] + \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + \|\boldsymbol{\xi}\|\right)^p \mathbb{1}_{\{\|\boldsymbol{\xi}\| > 2\varepsilon\}}\right] \right\} \\
&= \sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left\{ \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + \|\boldsymbol{\xi}\|\right)^p\right] - \mathbb{E}\left[\left(\left(f\left(\mathbf{x}_0\right) + \|\boldsymbol{\xi}\|\right)^p - \left(f\left(\mathbf{x}_0\right) + k\|\boldsymbol{\xi}\|\right)^p\right) \mathbb{1}_{\{\|\boldsymbol{\xi}\| \leqslant 2\varepsilon\}}\right] \right\} \\
&\leqslant \sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p] \leqslant \varepsilon^p} \left\{ \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + \|\boldsymbol{\xi}\|\right)^p\right] - p f^{p-1}\left(\mathbf{x}_0\right)(1-k)\mathbb{E}\left[\|\boldsymbol{\xi}\| \mathbb{1}_{\{\|\boldsymbol{\xi}\| \leqslant 2\varepsilon\}}\right] \right\} \\
&= \sup_{\mathbb{E}[|V|^p] \leqslant \varepsilon^p} \left\{ \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + |V|\right)^p\right] - p f^{p-1}\left(\mathbf{x}_0\right)(1-k)\mathbb{E}\left[|V| \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}\right] \right\} \\
&=: I,
\end{aligned}$$

where the second inequality holds because f is nonnegative with $\mathrm{Lip}(f) \leqslant 1$; the third inequality follows from (A24), and the penultimate equality holds since the objective function in the optimization problem depends only on $\|\boldsymbol{\xi}\|$. Define

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

we can rewrite $I = \max\{I_1, I_2\}$ with

$$I_i = \sup_{V \in \mathcal{V}_i} \left\{ \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right) + |V|\right)^p\right] - p f^{p-1}\left(\mathbf{x}_0\right)(1-k)\mathbb{E}\left[|V| \mathbb{1}_{\{|V| \leqslant 2\varepsilon\}}\right] \right\}, \quad i = 1, 2.$$

One can verify that

$$\mathbb{E}[|V|] = \mathbb{E}[|V|\mathbb{1}_{\{|V|>2\varepsilon\}}]+\mathbb{E}[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}] \leqslant \varepsilon 2^{-p/q}+\frac{\left(1-2^{-p/q}\right)\varepsilon}{2} = \frac{\left(1+2^{-p/q}\right)\varepsilon}{2} < \varepsilon, \quad \forall\, V \in \mathcal{V}_1, \tag{A25}$$

where the first inequality follows from the Hölder's inequality, Markov's inequality, and the definition of $\mathcal{V}_1$. It holds that

$$I_1 \leqslant \sup_{V\in\mathcal{V}_1} \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right)+|V|\right)^p\right] < \left(f\left(\mathbf{x}_0\right)+\varepsilon\right)^p, \tag{A26}$$

where the strict inequality follows from (A25) and Statement (ii) of Lemma A3 by noting that $f\left(\mathbf{x}_0\right) > 0$. For I_2, we have

$$\begin{aligned} I_2 &\leqslant \sup_{V\in\mathcal{V}_2} \mathbb{E}\left[\left(f\left(\mathbf{x}_0\right)+|V|\right)^p\right] - \inf_{V\in\mathcal{V}_2} pf^{p-1}\left(\mathbf{x}_0\right)(1-k)\mathbb{E}\left[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}\right] \\ &\leqslant \left(f\left(\mathbf{x}_0\right)+\varepsilon\right)^p - pf^{p-1}\left(\mathbf{x}_0\right)(1-k)\inf_{V\in\mathcal{V}_2} \mathbb{E}\left[|V|\mathbb{1}_{\{|V|\leqslant 2\varepsilon\}}\right] \\ &\leqslant \left(f\left(\mathbf{x}_0\right)+\varepsilon\right)^p - pf^{p-1}\left(\mathbf{x}_0\right)(1-k)\frac{\left(1-2^{-p/q}\right)\varepsilon}{2} \\ &< \left(f\left(\mathbf{x}_0\right)+\varepsilon\right)^p, \end{aligned} \tag{A27}$$

where the second inequality follows from Statement (i) of Lemma A3, and the third inequality is due to the definition of $\mathcal{V}_2$. Combining (A26) and (A27), we have

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}\|^p]\leqslant\varepsilon} \left(\mathbb{E}[f^p\left(\mathbf{x}_0+\boldsymbol{\xi}\right)]\right)^{1/p} \leqslant I^{1/p} = \max\left\{I_1^{1/p}, I_2^{1/p}\right\} < f\left(\mathbf{x}_0\right)+\varepsilon,$$

which yields a contradiction to (A21). Hence, (A22) holds. With the similar arguments as in the proof of Proposition 3, we have, if $f(\mathbf{x}) > 0$, $f(\mathbf{x}) = \max_{i\in I}\{\boldsymbol{\beta}_i^\top\mathbf{x}+b_i\}$ for some $\boldsymbol{\beta}_i \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}_i\|_* = 1$ and $b_i \in \mathbb{R}$ for $i \in I$; otherwise, $f(\mathbf{x}) = 0$. To show that f satisfies (24), let $g(\mathbf{x}) := \max_{i\in I}\{\boldsymbol{\beta}_i^\top\mathbf{x}+b_i\}$, where $(\boldsymbol{\beta}_i, b_i)$ are as in the representation of f on the set $\{\mathbf{x} : f(\mathbf{x}) > 0\}$. Note that $f(\mathbf{x}) = g(\mathbf{x})$ whenever $f(\mathbf{x}) > 0$. Define $h(\mathbf{x}) := (g(\mathbf{x}))_+$. We first show that $f \leqslant h$. Indeed, if $f(\mathbf{x}) > 0$, then $g(\mathbf{x}) = f(\mathbf{x}) > 0$, so $f(\mathbf{x}) = h(\mathbf{x})$; otherwise, $f(\mathbf{x}) = 0 \leqslant h(\mathbf{x})$. Thus, $f(\mathbf{x}) \leqslant h(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. We next show the reverse inequality $h \leqslant f$. Since f is convex and

satisfies (A22), we have

$$\begin{aligned} f(\mathbf{x}) &= \sup_{\mathbf{z}\in\mathbb{R}^n} \sup_{\boldsymbol{\beta}_\mathbf{z}\in\partial f(\mathbf{z})} \{f(\mathbf{z}) + \boldsymbol{\beta}_\mathbf{z}^\top(\mathbf{x}-\mathbf{z})\} \\ &= \max\left\{ \sup_{\mathbf{z}:f(\mathbf{z})=0} \sup_{\boldsymbol{\beta}_\mathbf{z}\in\partial f(\mathbf{z})} \{\boldsymbol{\beta}_\mathbf{z}^\top(\mathbf{x}-\mathbf{z})\}, \sup_{\mathbf{z}:f(\mathbf{z})>0} \sup_{\boldsymbol{\beta}_\mathbf{z}\in\partial f(\mathbf{z})} \{f(\mathbf{z}) + \boldsymbol{\beta}_\mathbf{z}^\top(\mathbf{x}-\mathbf{z})\} \right\} \\ &\geqslant \max\left\{ \sup_{\mathbf{z}:f(\mathbf{z})=0} \{\boldsymbol{\beta}_\mathbf{z}^\top(\mathbf{x}-\mathbf{z})\}, g(\mathbf{x}) \right\} \geqslant g(\mathbf{x}). \end{aligned}$$

Hence, if $g(\mathbf{x}) > 0$, then $h(\mathbf{x}) = g(\mathbf{x}) \leqslant f(\mathbf{x})$; and if $g(\mathbf{x}) \leqslant 0$, then $h(\mathbf{x}) = 0 \leqslant f(\mathbf{x})$. Thus, $f(\mathbf{x}) \geqslant h(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$. Combining both inequalities, we conclude that $f(\mathbf{x}) = h(\mathbf{x}) = (\max_{i\in I}\{\boldsymbol{\beta}_i^\top\mathbf{x} + b_i\})_+$ for every $\mathbf{x} \in \mathbb{R}^n$. This completes the proof. □

To prove Corollaries 2 and 3, it suffices to establish a more general result. To this end, we first present the following lemma, which extends Lemma EC.8 in Wu et al. (2022) to high-dimensional settings and incorporates the regularization results from Theorem 4. We begin by introducing a broad class of measures, each of which can be expressed in one of the following two forms:

$$\mathcal{V}_p^F(\boldsymbol{\xi}) = \inf_{t\in\mathbb{R}} \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi},t)\right]\right)^{1/p} \quad \text{and} \quad \rho_p^F(\boldsymbol{\xi}) = \inf_{t\in\mathbb{R}} \left\{ t + \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi},t)\right]\right)^{1/p} \right\} \tag{A28}$$

for some loss function $\ell : \mathbb{R}^{n+1} \to \mathbb{R}$.

Lemma A4. *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 $\ell(\mathbf{z},t)$ is nonnegative on $\mathbb{R}^{n+1}$, and convex in t with $\lim_{t\to-\infty} \partial\ell(\mathbf{z},t)/\partial t < -1$ for all $\mathbf{z} \in \mathbb{R}^n$, and Lipschitz continuous in $\mathbf{z}$ for all $t \in \mathbb{R}$ with a uniform Lipschitz constant, i.e., there exists $M > 0$ such that

$$|\ell(\mathbf{z}_1,t) - \ell(\mathbf{z}_2,t)| \leqslant M\|\mathbf{z}_1 - \mathbf{z}_2\|, \quad \forall, \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n,\ t \in \mathbb{R}.$$

Then we have

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \left\{ t + \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi},t)\right]\right)^{1/p} \right\} = \inf_{t\in\mathbb{R}} \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \left\{ t + \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi},t)\right]\right)^{1/p} \right\}.$$

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

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \mathbb{E}^F\left[\ell^p(\boldsymbol{\xi},t)\right] = \inf_{t\in\mathbb{R}} \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathbb{E}^F\left[\ell^p(\boldsymbol{\xi},t)\right].$$

Proof. (i) Denote by $\pi_{1,\ell}(F,t) := t + \left(\mathbb{E}^F\left[\ell^p\left(\boldsymbol{\xi},t\right)\right]\right)^{1/p}$. With the similar arguments as in the proof of Lemma EC.8 in Wu et al. (2022), one can verify that $\pi_{1,\ell}(F,t)$ is concave in F for all $t\in\mathbb{R}$ and convex in t for all $F\in\mathcal{M}_p\left(\mathbb{R}^n\right)$. Moreover, we have $\lim_{t\to\pm\infty}\pi_{1,\ell}(F,t)=\infty$ for all $F\in\mathcal{M}_p\left(\mathbb{R}^n\right)$. Thus, 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 $\{t(F): F\in\mathbb{B}_p\left(F_0,\varepsilon\right)\}$ is a subset of a compact set. For any $F\in\mathbb{B}_p\left(F_0,\varepsilon\right)$ and $t\in\mathbb{R}$, let $\boldsymbol{\xi}\sim F$ and $\boldsymbol{\zeta}\sim F_0$ such that $\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|^p]\leqslant\varepsilon^p$, and we have

$$\begin{aligned}\left|\pi_{1,\ell}(F,t)-\pi_{1,\ell}\left(F_0,t\right)\right| &= \left|\left(\mathbb{E}^F\left[\ell^p\left(\boldsymbol{\xi},t\right)\right]\right)^{1/p}-\left(\mathbb{E}^{F_0}\left[\ell^p\left(\boldsymbol{\zeta},t\right)\right]\right)^{1/p}\right| \\ &\leqslant \left(\mathbb{E}\left[\left|\ell\left(\boldsymbol{\xi},t\right)-\ell\left(\boldsymbol{\zeta},t\right)\right|^p\right]\right)^{1/p} \\ &\leqslant \left(\mathbb{E}\left[M^p\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|^p\right]\right)^{1/p}\leqslant M\varepsilon, \end{aligned} \tag{A29}$$

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

$$\pi_{1,\ell}\left(F,t\left(F_0\right)\right)\leqslant\pi_{1,\ell}\left(F_0,t\left(F_0\right)\right)+M\varepsilon. \tag{A30}$$

Note that $\pi_{1,\ell}\left(F_0,t\right)\to\infty$ as $t\to\pm\infty$. There exists $\Delta>0$ such that $\pi_{1,\ell}\left(F_0,t\right)>\pi_{1,\ell}\left(F_0,t\left(F_0\right)\right)+2M\varepsilon$ for all $t\notin\left[t\left(F_0\right)-\Delta,t\left(F_0\right)+\Delta\right]$. This, combined with (A29), imply that

$$\pi_{1,\ell}(F,t)\geqslant\pi_{1,\ell}\left(F_0,t\right)-M\varepsilon>\pi_{1,\ell}\left(F_0,t\left(F_0\right)\right)+M\varepsilon,\quad\forall t\notin\left[t\left(F_0\right)-\Delta,t\left(F_0\right)+\Delta\right]. \tag{A31}$$

Applying (A30) and (A31), we have $\{t(F): F\in\mathbb{B}_p\left(F_0,\varepsilon\right)\}\subseteq\left[t\left(F_0\right)-\Delta,t\left(G_0\right)+\Delta\right]$. Using a minimax theorem (see e.g., Sion (1958)), it holds that

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

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

(ii) The proof is similar to (i). □

Lemma A5. *For any $p\in[1,\infty)$ and $c>0$, let $\mathcal{V}_p$ and ρ_p be defined as in equation (A28), where the loss function is given by $\ell(\mathbf{z},t) := (\max_{i\in I}\{c\boldsymbol{\beta}_i^\top\mathbf{z}+d_it+b_i\})_+$ with $\boldsymbol{\beta}_i\in\mathbb{R}^n$ such that $\|\boldsymbol{\beta}_i\|_*=1$, and $b_i,d_i\in\mathbb{R}$ for all $i\in I$.*

(i) If $\min_{i\in I} d_i \leqslant 0 \leqslant \max_{i\in I} d_i$*, then for any* $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$ *and* $\varepsilon > 0$*, we have*

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathcal{V}_p^F(\boldsymbol{\xi}) = \mathcal{V}_p^{F_0}(\boldsymbol{\zeta}) + c\varepsilon. \tag{A32}$$

If $\min_{i\in I} d_i > 0$ *or* $\max_{i\in I} d_i < 0$*, we have* $\mathcal{V}_p^F(\boldsymbol{\xi}) = 0$ *for any* $F \in \mathcal{M}_p(\mathbb{R}^n)$*.*

(ii) If $\min_{i\in I} d_i \leqslant -1$*, then for any* $F_0 \in \mathcal{M}_p(\mathbb{R}^n)$ *and* $\varepsilon > 0$*, we have*

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \rho_p^F(\boldsymbol{\xi}) = \rho_p^{F_0}(\boldsymbol{\zeta}) + c\varepsilon. \tag{A33}$$

If $\min_{i\in I} d_i > -1$*, then* $\rho_p^F(\boldsymbol{\xi}) = -\infty$ *for any* $F \in \mathcal{M}_p(\mathbb{R}^n)$*.*

Proof. (i) We establish the result by analyzing the following three cases.

(i.a) For $\min_{i\in I} d_i < 0 < \max_{i\in I} d_i$, to apply Lemma A4, it suffices to verify that the loss function ℓ satisfies the required conditions. Since $\ell(\mathbf{z}, t)$ is defined as the pointwise maximum of affine functions in t for any $\mathbf{z} \in \mathbb{R}^n$, it follows that $\ell(\mathbf{z}, t)$ is convex in $t \in \mathbb{R}$. Note that $\min_{i\in I} d_i < 0 < \max_{i\in I} d_i$. We have $\lim_{t\to\pm\infty} \ell(\mathbf{z}, t) = \infty$ for all $\mathbf{z} \in \mathbb{R}^n$, which implies that $\lim_{t\to-\infty} \partial\ell(\mathbf{z}, t)/\partial t < 0$ and $\lim_{t\to\infty} \partial\ell(\mathbf{z}, t)/\partial t > 0$ for all $\mathbf{z} \in \mathbb{R}^n$. Moreover, $\ell(\mathbf{z}, t)$ is Lipschitz continuous in $\mathbf{z} \in \mathbb{R}^n$ uniformly over $t \in \mathbb{R}$, with Lipschitz constant c. Indeed, for any $t \in \mathbb{R}$ and $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$,

$$|\ell(\mathbf{z}_1, t) - \ell(\mathbf{z}_2, t)| \leqslant c \max_{i\in I}\{\boldsymbol{\beta}_i^\top(\mathbf{z}_1 - \mathbf{z}_2)\} \leqslant c \max_{i\in I} \|\boldsymbol{\beta}_i\|_* \|\mathbf{z}_1 - \mathbf{z}_2\| \leqslant c\|\mathbf{z}_1 - \mathbf{z}_2\|,$$

where the second inequality follows from Hölder's inequality, and the last inequality follows from the assumption $\|\boldsymbol{\beta}_i\|_* = 1$ for $i \in I$. Hence, the function $\ell(\mathbf{z}, t)$ satisfies the conditions in Lemma A4 (ii). It follows that

$$\begin{aligned}
\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathcal{V}_p^F(\boldsymbol{\xi}) &= \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right]\right)^{1/p} \\
&= \inf_{t\in\mathbb{R}} \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right]\right)^{1/p} \\
&= \inf_{t\in\mathbb{R}} \left\{\left(\mathbb{E}^{F_0}\left[\ell^p(\boldsymbol{\zeta}, t)\right]\right)^{1/p} + c\varepsilon\right\} = \mathcal{V}_p^{F_0}(\boldsymbol{\zeta}) + c\varepsilon,
\end{aligned}$$

where the second equality follows from Lemma A4 (ii), and the third follows from Theorem 4.

(i.b) For $\min_{i\in I} d_i = 0$ or $\max_{i\in I} d_i = 0$, note that when $\min_{i\in I} d_i = \max_{i\in I} d_i = 0$, we have $\mathcal{V}_p^F(\boldsymbol{\xi}) = \mathbb{E}^F\left[\left(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top\boldsymbol{\xi} + b_i\}\right)_+^p\right]$. In this case, Theorem 4 implies that (A32) holds. It

therefore remains to consider the situations where $\min_{i\in I} d_i = 0 < \max_{i\in I} d_i$ or $\min_{i\in I} d_i < \max_{i\in I} d_i = 0$. Since these two cases are symmetric, it suffices to analyze $\min_{i\in I} d_i = 0 < \max_{i\in I} d_i$. Denote by $I_0 := \{i \in I : d_i = 0\}$. Since $d_i \geqslant 0$ for all $i \in I$, the function $\ell(\mathbf{z}, t) = (\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \mathbf{z} + d_i t + b_i\})_+$ is increasing in $t \in \mathbb{R}$. It follows that

$$\mathcal{V}_p^F(\boldsymbol{\xi}) = \inf_{t\in\mathbb{R}} \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right]\right)^{1/p} = \lim_{t\to-\infty} \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right]\right)^{1/p} = \left(\mathbb{E}^F[(\max_{i\in I_0}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\})_+^p]\right)^{1/p}.$$

Thus, by Theorem 4,

$$\begin{aligned}\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathcal{V}_p^F(\boldsymbol{\xi}) &= \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \left(\mathbb{E}^F[(\max_{i\in I_0}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\})_+^p]\right)^{1/p} \\ &= \left(\mathbb{E}^{F_0}[(\max_{i\in I_0}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\zeta} + b_i\})_+^p]\right)^{1/p} + c\varepsilon = \mathcal{V}_p^{F_0}(\boldsymbol{\zeta}) + c\varepsilon,\end{aligned}$$

Hence, (A32) holds.

(i.c) For $\min_{i\in I} d_i > 0$ or $\max_{i\in I} d_i < 0$, we have

$$\begin{aligned}\mathcal{V}_p^F(\boldsymbol{\xi}) &= \inf_{t\in\mathbb{R}} (\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right])^{1/p} \\ &= \begin{cases} \lim_{t\to\infty} \left(\mathbb{E}^F\left[\left(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + d_i t + b_i\}\right)_+^p\right]\right)^{1/p} = 0, & \max_{i\in I} d_i < 0, \\ \lim_{t\to-\infty} \left(\mathbb{E}^F\left[\left(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + d_i t + b_i\}\right)_+^p\right]\right)^{1/p} = 0, & \min_{i\in I} d_i > 0. \end{cases}\end{aligned}$$

Combining the above three cases, we complete the proof of (i).

(ii) We establish the result by analyzing the following three cases.

(ii.a) For $\min_{i\in I} d_i < -1$, by similar arguments as in part (i), the loss function $\ell(\mathbf{z}, t)$ is convex in t for any fixed $\mathbf{z} \in \mathbb{R}^n$, and Lipschitz continuous in $\mathbf{z} \in \mathbb{R}^n$ uniformly over $t \in \mathbb{R}$, with Lipschitz constant c. Moreover, we have $\lim_{t\to-\infty} \partial\ell(z, t)/\partial t = \min_{i\in I} d_i < -1$. Therefore, $\ell(\mathbf{z}, t)$ satisfies the assumptions in Lemma A4 (i), and we have

$$\begin{aligned}\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \rho_p^F(\boldsymbol{\xi}) &= \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \left\{t + \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right]\right)^{1/p}\right\} \\ &= \inf_{t\in\mathbb{R}} \left\{t + \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \left(\mathbb{E}^F\left[\ell^p(\boldsymbol{\xi}, t)\right]\right)^{1/p}\right\} \\ &= \inf_{t\in\mathbb{R}} \left\{t + \left(\mathbb{E}^{F_0}\left[\ell^p(\boldsymbol{\zeta}, t)\right]\right)^{1/p} + c\varepsilon\right\} = \rho_p^{F_0}(\boldsymbol{\zeta}) + c\varepsilon,\end{aligned}$$

where the second equality follows from Lemma A4 (i), and the third follows from Theorem 4.

(ii.b) For $\min_{i\in I} d_i = -1$, denote by $I_{-1} = \{i \in I : d_i = -1\}$. Then, for any $t \in \mathbb{R}$, we have

$$\begin{aligned} t + \left(\mathbb{E}^F[\ell^p(\boldsymbol{\xi}, t)]\right)^{1/p} &= t + \left(\mathbb{E}^F[(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + d_i t + b_i\})_+^p]\right)^{1/p} \\ &\geqslant t + \left(\mathbb{E}^F[\max_{i\in I_{-1}}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} - t + b_i\})_+^p]\right)^{1/p} \\ &\geqslant t + \mathbb{E}^F\left[(\max_{i\in I_{-1}}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} - t + b_i\})_+\right] \geqslant \mathbb{E}^F\left[\max_{i\in I_{-1}}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\}\right], \end{aligned}$$

where the second inequality follows from Hölder's inequality, and the last inequality from the fact that $(x)_+ \geqslant x$. On the other hand, let $g(t) := t + (\mathbb{E}[(\ell^p(\boldsymbol{\xi}, t)])^{1/p} = t + (\mathbb{E}[(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + d_i t + b_i\})_+^p])^{1/p}$. Note that

$$g'(t) = 1 + \frac{\mathbb{E}[\ell^{p-1}(\boldsymbol{\xi}, t) d_*]}{(\mathbb{E}[\ell^p(\boldsymbol{\xi}, t)])^{(1-p)/p}} \geqslant 1 + \frac{(\min_{i\in I} d_i)\mathbb{E}[(\ell^{p-1}(\boldsymbol{\xi}, t)]}{(\mathbb{E}[(\ell^p(\boldsymbol{\xi}, t)])^{(1-p)/p}} \geqslant 1 + \min_{i\in I} d_i = 0,$$

where d_* denotes the coefficient d_i corresponding to the maximizer inside $\ell(\boldsymbol{\xi}, t)$ and the first inequality follows from $d_* \geqslant \min_{i\in I} d_i$. Thus, $g(t)$ is increasing in $t \in \mathbb{R}$ and

$$\rho_p^F(\boldsymbol{\xi}) = \lim_{t\to-\infty}\left\{t + \left(\mathbb{E}^F[(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + d_i t + b_i\})_+^p]\right)^{1/p}\right\} = \mathbb{E}^F\left[\max_{i\in I_{-1}}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\}\right].$$

By Theorem 4, it follows that

$$\begin{aligned} \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \rho_p^F(\boldsymbol{\xi}) &= \sup_{F\in\mathbb{B}_p(F_0,\varepsilon)} \mathbb{E}^F[\max_{i\in I_{-1}}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + b_i\}] \\ &= \mathbb{E}^F[\max_{i\in I_{-1}}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\zeta} + b_i\}] + c\varepsilon = \rho_p^{F_0}(\boldsymbol{\zeta}) + c\varepsilon. \end{aligned}$$

Hence, (A33) holds.

(ii.c) For $\min_{i\in I} d_i > -1$, then for any $F \in \mathcal{M}_p(\mathbb{R}^n)$, we have

$$\begin{aligned} \rho_p^F(\boldsymbol{\xi}) &= \inf_{t\in\mathbb{R}}\left\{t + \left(\mathbb{E}^F[\ell^p(\boldsymbol{\xi}, t)]\right)^{1/p}\right\} \\ &= \lim_{t\to-\infty}\left\{t + \left(\mathbb{E}^F\left[\left(\max_{i\in I}\{c\boldsymbol{\beta}_i^\top \boldsymbol{\xi} + d_i t + b_i\}\right)_+^p\right]\right)^{1/p}\right\} = -\infty. \end{aligned}$$

This completes the proof. □

Proof of Corollary 2. Note that

$$\begin{aligned}\ell(f(\boldsymbol{\xi}),t) &= c\left(\max_{i\in I_1}\{\boldsymbol{\beta}_i^\top\boldsymbol{\xi}+b_i\}+\max_{j\in I_2}\{d_j t\}\right)_+ \\ &= c\left(\max_{\{i\in I_1,\, j\in I_2\}}\{\boldsymbol{\beta}_i^\top\boldsymbol{\xi}+b_i+d_j t\}\right)_+ = c\left(\max_{k\in I_3}\{\boldsymbol{\beta}_k^\top\boldsymbol{\xi}+b_k+d_k t\}\right)_+,\end{aligned}$$

where for each $k\in I_3$, we have $(\boldsymbol{\beta}_k,b_k)\in\{(\boldsymbol{\beta}_i,b_i),\, i\in I_1\}$ and $d_k\in\{d_j,\, j\in I_2\}$. Letting the loss function in Corollary A5 be denoted by $\ell_1(\mathbf{z},t):=\ell(f(\mathbf{z}),t)$, we can directly apply Corollary A5 to obtain the desired result. This completes the proof. □

Proof of Corollary 3. Note that

$$\begin{aligned}\ell(f(\boldsymbol{\xi}),t) &= c\left(\{\max\{\boldsymbol{\beta}^\top\boldsymbol{\xi}+b-d_1 t,-\boldsymbol{\beta}^\top\boldsymbol{\xi}-b-d_2 t\}+d_3\right)_+ \\ &= c\left(\{\max\{\boldsymbol{\beta}^\top\boldsymbol{\xi}+b_1-d_1 t,-\boldsymbol{\beta}^\top\boldsymbol{\xi}-b_2-d_2 t\}\right)_+,\end{aligned}$$

where $b_1:=b+d_3$ and $b_2:=b-d_3$. Letting the loss function in Corollary A5 be denoted by $\ell_1(\mathbf{z},t):=\ell(f(\mathbf{z}),t)$, we can directly apply Corollary A5 to obtain the desired result. This completes the proof. □

Proof of Theorem 5. For the "if" part, we assume that $f:\mathbb{R}^n\to\mathbb{R}$ be given by (13). By Proposition 3, we have

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)}\rho_h^F[f(\boldsymbol{\xi})]=\sup_{G\in\mathcal{C}_p(f|F_0,c_f\varepsilon)}\rho_h^G[X].$$

Note that $\mathcal{C}_p(f|F_0,c_f\varepsilon)$ is a one-dimensional Wasserstein ball with center $f(\boldsymbol{\zeta})$ and thus, by Theorem 5 of Wu et al. (2022), we have

$$\sup_{G\in\mathcal{C}_p(f|F_0,c_f\varepsilon)}\rho_h^G[X]=\rho_h^{F_0}(f(\boldsymbol{\zeta}))+c_f\varepsilon\|h'\|_q.$$

Combining the two equations yields (28).

For the "only if" part, we assume that there exists $c_f\geqslant 0$ such that (28) holds for any $F_0\in\mathcal{M}(\mathbb{R}^n)$ and $\varepsilon>0$. If $c_f=0$, then (28) reduces to

$$\sup_{F\in\mathbb{B}_p(F_0,\varepsilon)}\rho_h^F[f(\boldsymbol{\xi})]=\rho_h^{F_0}(f(\boldsymbol{\zeta})),$$

which is equivalent to

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p \leqslant \varepsilon^p} \rho_h(f(\boldsymbol{\xi}+\boldsymbol{\zeta})) = \rho_h^{F_0}(f(\boldsymbol{\zeta})). \tag{A34}$$

We next show that $\|\nabla f(\mathbf{x})\|_* = 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$. Suppose, for contradiction, that there exists $\mathbf{x}_0$ such that $\|\nabla f(\mathbf{x}_0)\|_* > 0$. Then, there exists $\boldsymbol{v}_{\mathbf{x}_0}$ with $\|\boldsymbol{v}_{\mathbf{x}_0}\| = 1$ such that $\boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\|_*$. Let $F_0 := \delta_{\mathbf{x}_0}$ and consider $\boldsymbol{\xi} \sim \delta_{\varepsilon \boldsymbol{v}_{\mathbf{x}_0}}$ for some sufficiently small $\varepsilon > 0$. Then $\mathbb{E}[\|\boldsymbol{\xi}\|^p] = \varepsilon^p$, and we have

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p \leqslant \varepsilon^p} \rho_h(f(\boldsymbol{\xi}+\mathbf{x}_0)) \geqslant \rho_h(f(\varepsilon \boldsymbol{v}_{\mathbf{x}_0}+\mathbf{x}_0)) \geqslant f(\mathbf{x}_0) + \varepsilon \boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = f(\mathbf{x}_0) + \varepsilon \|\nabla f(\mathbf{x}_0)\|_* > f(\mathbf{x}_0),$$

where the second inequality follows from that f is convex, and the strict inequality follows from $\|\nabla f(\mathbf{x}_0)\|_* > 0$ and $\varepsilon > 0$. This yields a contradiction with (A34). Thus, we have $\|\nabla f(\mathbf{x})\|_* = 0$ for all $\mathbf{x} \in \mathbb{R}^n$, which implies that f is constant. Now consider the case $c_f > 0$. By the positive homogeneity of ρ_h, we assume without loss of generality that $c_f = 1$. Note that in this case, equation (28) is equivalent to

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|^p] \leqslant \varepsilon^p} \rho_h(f(\boldsymbol{\xi})) = \rho_h^{F_0}(f(\boldsymbol{\zeta})) + \varepsilon \|h'\|_q,$$

or, equivalently,

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p \leqslant \varepsilon^p} \rho_h(f(\boldsymbol{\xi}+\boldsymbol{\zeta})) = \rho_h^{F_0}(f(\boldsymbol{\zeta})) + \varepsilon \|h'\|_q. \tag{A35}$$

It suffices to show that (A35) implies f satisfies

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) = \varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\, \varepsilon > 0. \tag{A36}$$

To see it, we first show (A35) implies that $\|\nabla f\|_* \leqslant 1$ for any $\mathbf{x} \in \mathbb{R}^n$ and any subgradient $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$. Since otherwise, $\|\nabla f(\mathbf{x}_0)\|_* > 1$ for some $\mathbf{x}_0$. Then, there exists $\boldsymbol{v}_{\mathbf{x}_0}$ with $\|\boldsymbol{v}_{\mathbf{x}_0}\| = 1$ such that $\boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0) = \|\nabla f(\mathbf{x}_0)\|_*$. Denote by U a uniform random variable on $[0,1]$, and define the random vector $\boldsymbol{\xi}_1 := \varepsilon \boldsymbol{v}_{\mathbf{x}_0} (h'(U))^{q/p} / \|h'\|_q^{q/p}$. One can verify that $\mathbb{E}\|\boldsymbol{\xi}_1\|^p \leqslant \varepsilon^p$. For $\varepsilon > 0$ small enough, we have

$$\begin{aligned}\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p \leqslant \varepsilon^p} \rho_h(f(\mathbf{x_0}+\boldsymbol{\xi})) &\geqslant \rho_h(f(\mathbf{x}_0+\boldsymbol{\xi}_1)) \geqslant \rho_h\left(f(\mathbf{x}_0) + \varepsilon \frac{\boldsymbol{v}_{\mathbf{x}_0}^\top \nabla f(\mathbf{x}_0)}{\|h'\|_q^{q/p}} (h'(U))^{q/p}\right) \\ &= f(\mathbf{x}_0) + \frac{\varepsilon \|\nabla f(\mathbf{x}_0)\|_*}{\|h'\|_q^{q/p}} \rho_h\left((h'(U))^{q/p}\right) > f(\mathbf{x}_0) + \varepsilon \|h'\|_q,\end{aligned}$$

where the second inequality follows from the convexity of f, and the strict inequality follows from the assumption $\|\nabla f(\mathbf{x}_0)\|_* > 1$. This yields a contradiction with (A35). It follows that all subgradients of f satisfy $\|\nabla f\|_* \leqslant 1$, which in turn implies that f is Lipschitz continuous with $\mathrm{Lip}(f) \leqslant 1$. We next show that f satisfies (A36). Since $\|\nabla f\|_* \leqslant 1$, suppose for contradiction that there exist some $\mathbf{x}_0 \in \mathbb{R}^n$ and $\varepsilon_0 > 0$ such that

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon_0} f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0) < \varepsilon_0. \tag{A37}$$

We first consider the case $p = \infty$. In this case, we have

$$\sup_{\text{ess-sup}(\|\boldsymbol{\xi}\|) \leqslant \varepsilon_0} \rho_h\left(f(\mathbf{x_0} + \boldsymbol{\xi})\right) = \sup_{\|\mathbf{y}\| \leqslant \varepsilon_0} f(\mathbf{x}_0 + \mathbf{y}) < f(\mathbf{x}_0) + \varepsilon_0,$$

where the first equality follows from $\sup_{\text{ess-sup}(\|\boldsymbol{\xi}\|) \leqslant \varepsilon_0} \rho_h\left(f(\mathbf{x}_0 + \boldsymbol{\xi})\right) \leqslant \rho_h(\sup_{\|\mathbf{y}\| \leqslant \varepsilon_0} f(\mathbf{x}_0 + \mathbf{y}))$. Thus again contradicts with (A35). Now suppose $p \in (1, \infty)$. We first define a set,

$$\mathcal{X}_p = \left\{ F^{-1}(U) : \int_0^1 |F^{-1}(u)|^p \mathrm{d}u < \infty,\ F^{-1}(U) \geqslant 0 \right\}.$$

We have that $\{F_X : X \in L^p\} = \{F_X : X \in \mathcal{X}_p\}$ and the random variables in $\mathcal{X}_p$ are all comonotonic. For sufficiently small $\varepsilon > 0$, define

$$\widetilde{\alpha} = \{\alpha \in [0,1] : h(\alpha) = 0\} < 1 \text{ and } M = \varepsilon \left(\frac{1 - \widetilde{\alpha}}{2}\right)^{-1/p}.$$

For each random variable $X \in \mathcal{X}_p$, by the Markov's inequality, we have

$$\mathbb{P}(X > M) \leqslant \frac{\mathbb{E}[X^p]}{M^p} \leqslant \frac{\varepsilon^p}{M^p} = \frac{1 - \widetilde{\alpha}}{2}. \tag{A38}$$

For small enough $\varepsilon > 0$ such that $M \leqslant \varepsilon_0$, define

$$k = \sup_{\|\mathbf{y}\| \leqslant M} \left\{ \frac{f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0)}{M} \right\}.$$

By the strict inequality in (A37) and the convexity of f, it follows that $k < 1$. Further, one can verify that $f(\mathbf{x}_0 + \mathbf{y}) - f(\mathbf{x}_0) \leqslant k\|\mathbf{y}\|$ for all $\|\mathbf{y}\| \leqslant M$ by using the convexity of f again. Then, we

have

$$
\begin{aligned}
\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p\leqslant\varepsilon^p}\rho_h\left(f(\mathbf{x_0}+\boldsymbol{\xi})\right) &= \sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p\leqslant\varepsilon^p}\{\rho_h\left(f\left(\mathbf{x}_0+\boldsymbol{\xi}\right)\mathbb{1}_{\{\|\boldsymbol{\xi}\|\leqslant M\}}+f\left(\mathbf{x}_0+\boldsymbol{\xi}\right)\mathbb{1}_{\{\|\boldsymbol{\xi}\|>M\}}\right)\}\\
&\leqslant \sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p\leqslant\varepsilon^p}\{\rho_h\left(\left(f(\mathbf{x}_0)+k\|\boldsymbol{\xi}\|\right)\mathbb{1}_{\{\|\boldsymbol{\xi}\|\leqslant M\}}+f\left(\mathbf{x}_0+\boldsymbol{\xi}\right)\mathbb{1}_{\{\|\boldsymbol{\xi}\|>M\}}\right)\}\\
&\leqslant \sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p\leqslant\varepsilon^p}\{\rho_h\left(\left(f(\mathbf{x}_0)+k\|\boldsymbol{\xi}\|\right)\mathbb{1}_{\{\|\boldsymbol{\xi}\|\leqslant M\}}+\left(f\left(\mathbf{x}_0\right)+\|\boldsymbol{\xi}\|\right)\mathbb{1}_{\{\|\boldsymbol{\xi}\|>M\}}\right)\}\\
&= f\left(\mathbf{x}_0\right)+\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p\leqslant\varepsilon^p}\{\rho_h\left(k\|\boldsymbol{\xi}\|\mathbb{1}_{\{\|\boldsymbol{\xi}\|\leqslant M\}}+\|\boldsymbol{\xi}\|\mathbb{1}_{\{\|\boldsymbol{\xi}\|>M\}}\right)\}\\
&= f\left(\mathbf{x}_0\right)+\sup_{\mathbb{E}[X^p]\leqslant\varepsilon^p,\,X\geqslant 0}\{\rho_h\left(kX\mathbb{1}_{\{X\leqslant M\}}+X\mathbb{1}_{\{X>M\}}\right)\}\\
&= f\left(\mathbf{x}_0\right)+\sup_{\mathbb{E}[X^p]\leqslant\varepsilon^p,\,X\in\mathcal{X}_p}\{\rho_h\left(kX\mathbb{1}_{\{X\leqslant M\}}+X\mathbb{1}_{\{X>M\}}\right)\}\\
&= f\left(\mathbf{x}_0\right)+\sup_{\mathbb{E}[X^p]\leqslant\varepsilon^p,\,X\in\mathcal{X}_p}\{\mathbb{E}[(kX\mathbb{1}_{\{X\leqslant M\}}+X\mathbb{1}_{\{X>M\}})h'(U)]\}\\
&= f\left(\mathbf{x}_0\right)+\sup_{\mathbb{E}[X^p]\leqslant\varepsilon^p,\,X\in\mathcal{X}_p}\{\mathbb{E}[(X-(1-k)X\mathbb{1}_{\{X\leqslant M\}})h'(U)]\}\\
&= f\left(\mathbf{x}_0\right)+\sup_{\mathbb{E}[X^p]\leqslant\varepsilon^p,\,X\in\mathcal{X}_p}\{\rho_h(X)-(1-k)\mathbb{E}[X\mathbb{1}_{\{X\leqslant M\}})h'(U)]\}\\
&:= f\left(\mathbf{x}_0\right)+I,
\end{aligned}
$$

where the second inequality follows from $\|\nabla f\|\leqslant 1$ for all $\mathbf{x}\in\mathbb{R}^n$, the forth equality follows from the law invariance of ρ_h and the fifth equality holds because $X\in\mathcal{X}_p$ implies that $kX\mathbb{1}_{\{X\leqslant M\}}+X\mathbb{1}_{\{X>M\}}$ and $h'(U)$ are comonotonic. For $\eta>0$, define

$$
\mathcal{V}_1=\left[X\in\mathcal{X}_p:\mathbb{E}[X^p]\leqslant\varepsilon^p,\,\mathbb{E}[X\mathbb{1}_{\{X\leqslant M\}}]\leqslant\eta\right\},\quad \mathcal{V}_2=\{X\in\mathcal{X}_p:\mathbb{E}[X^p]\leqslant\varepsilon^p\}\setminus\mathcal{V}_1,
$$

and write $I=\max\{I_1,I_2\}$, where

$$
I_i=\sup_{X\in\mathcal{V}_i}\{\rho_h(X)-(1-k)\mathbb{E}[X\mathbb{1}_{\{X\leqslant M\}})h'(U)]\},\quad i=1,2.
$$

Below we aim to demonstrate that $I_i < \eta \left\|h'\right\|_q$ for $i = 1, 2$ by selecting an appropriate ε. Note that

$$\begin{aligned} I_1 \leqslant \sup_{X\in\mathcal{V}_1} \rho_h(X) &= \sup_{X\in\mathcal{V}_1} \left\{\mathbb{E}\left[X\mathbb{1}_{\{X>M\}}h'(U)\right] + \mathbb{E}\left[X\mathbb{1}_{\{X\leqslant M\}}h'(U)\right]\right\} \\ &\leqslant \sup_{X\in\mathcal{V}_1} \left\|h'(U)\mathbb{1}_{\{X>M\}}\right\|_q \|X\|_p + \eta \\ &\leqslant \sup_{X\in\mathcal{V}_1} \left\|h'(U)\mathbb{1}_{\{X>M\}}\right\|_q \varepsilon + \eta \\ &\leqslant \left(\int_{\frac{1+\widetilde{\alpha}}{2}}^1 \left(h'(s)\right)^q \, \mathrm{d}s\right)^{1/q} \varepsilon + \eta = \left(\left\|h'\right\|_q^q - \int_{\widetilde{\alpha}}^{\frac{1+\widetilde{\alpha}}{2}} \left(h'(s)\right)^q \, \mathrm{d}s\right)^{1/q} \varepsilon + \eta, \quad \text{(A39)} \end{aligned}$$

where the second inequality follows from Hölder's inequality and the definition of $\mathcal{V}_1$, the last inequality is due to (A38), and the last equality follows from the definition of $\widetilde{\alpha}$. Denote by

$$A = \int_{\widetilde{\alpha}}^{\frac{1+\widetilde{\alpha}}{2}} \left(h'(s)\right)^q \, \mathrm{d}s \text{ and } \varepsilon = \left[\left\|h'\right\|_q - \left(\left\|h'\right\|_q^q - A\right)^{1/q}\right] \varepsilon.$$

Recalling the definition of $\widetilde{\alpha}$ again, we have $A > 0$ and $\varepsilon > 0$, and hence, $I_1 < \varepsilon \left\|h'\right\|_q$ whenever $\eta < \varepsilon$. For $\eta < \varepsilon$, we have

$$I_2 = \sup_{X\in\mathcal{V}_2} \left\{\rho_h(X) - (1-k)\mathbb{E}\left[X\mathbb{1}_{\{X\leqslant M\}}h'(U)\right]\right\} \leqslant \sup_{X\in\mathcal{V}_2} \left\|h'\right\|_q \|X\|_p - (1-k)\eta < \varepsilon \left\|h'\right\|_q, \quad \text{(A40)}$$

where the first inequality follows from Hölder's inequality, and the strict inequality is due to $k < 1$. Hence, we have that $\max\{I_1, I_2\} < \varepsilon \left\|h'\right\|_q$. Combining (A39) and (A40), we conclude that

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}\|^p \leqslant \varepsilon^p} \rho_h\left(f(\mathbf{x_0} + \boldsymbol{\xi})\right) \leqslant f(\mathbf{x_0}) + \max\{I_1, I_2\} < f(\mathbf{x}_0) + \varepsilon \|h'\|_q.$$

This leads to a contradiction, thereby establishing (A36). With the similar arguments as in the proof of Proposition 3, we have $f(\mathbf{x}) = \max_{i\in I}\{\boldsymbol{\beta}_i^\top \mathbf{x} + b_i\}$ for some $\boldsymbol{\beta}_i \in \mathbb{R}^n$ with $\|\boldsymbol{\beta}_i\|_* = 1$ and $b_i \in \mathbb{R}$ for $i \in I$. This completes the proof. □

Proof of Proposition 5. (i) Denote by $\widetilde{\ell}(\mathbf{z}, t) := \ell(f(\mathbf{z}), t)$. For any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$ with $\mathbf{z}_1 \neq \mathbf{z}_2$,

$$|\widetilde{\ell}(\mathbf{z}_1, t) - \widetilde{\ell}(\mathbf{z}_2, t)| = |\ell(f(\mathbf{z}_1), t) - \ell(f(\mathbf{z}_2), t)| \leqslant b|f(\mathbf{z}_1) - f(\mathbf{z}_2)| \leqslant b\mathrm{Lip}(f)\|\mathbf{z}_1 - \mathbf{z}_2\|, \quad \forall\, t \in \mathbb{R},$$

where the first inequality follows from the uniform Lipschitz continuity of $\ell(z, t)$ in z, and the second from the Lipschitz continuity of f. Thus, $\widetilde{\ell}(\mathbf{z}, t)$ is Lipschitz continuous in $\mathbf{z}$ for all t, with constant $b\mathrm{Lip}(f)$. Moreover, one can verify that $\widetilde{\ell}(\mathbf{z}, t)$ is convex in t and satisfies $\lim_{t\to-\infty} \partial\widetilde{\ell}(\mathbf{z}, t)/\partial t < 0 < \lim_{t\to\infty} \partial\widetilde{\ell}(\mathbf{z}, t)/\partial t$ for all $\mathbf{z} \in \mathbb{R}^n$. Therefore, $\widetilde{\ell}(\mathbf{z}, t)$ satisfies the assumptions of Lemma A4 (ii).

Hence,

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \mathcal{D}_1^F(f(\boldsymbol{\xi})) = \sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \inf_{t\in\mathbb{R}} \left(\mathbb{E}^F\left[\widetilde{\ell}(\boldsymbol{\xi},t)\right]\right)$$
$$= \inf_{t\in\mathbb{R}} \sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \left(\mathbb{E}^F\left[\widetilde{\ell}(\boldsymbol{\xi},t)\right]\right) = \inf_{t\in\mathbb{R}} \sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \left(\mathbb{E}^F\left[\ell(f(\boldsymbol{\xi}),t)\right]\right). \tag{A41}$$

We next show that

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \left(\mathbb{E}^F\left[\ell(f(\boldsymbol{\xi}),t)\right]\right) = \sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|]\leqslant\varepsilon} \left(\mathbb{E}\left[\ell(f(\boldsymbol{\xi}),t)\right]\right) = \mathbb{E}^{F_0}\left[\ell(f(\boldsymbol{\zeta}),t)\right] + b\mathrm{Lip}(f)\varepsilon.$$

It suffices to establish the second equality. First,

$$\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|]\leqslant\varepsilon} \left(\mathbb{E}\left[\ell(f(\boldsymbol{\xi}),t)\right] - \mathbb{E}\left[\ell(f(\boldsymbol{\zeta}),t)\right]\right) \leqslant b\mathrm{Lip}(f) \sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|]\leqslant\varepsilon} \mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\| \leqslant b\mathrm{Lip}(f)\varepsilon, \tag{A42}$$

where the first inequality follows from the Lipschitz continuity of $\ell(\cdot,t)$ and f. For the reverse inequality, following arguments similar to those in Lemma A2 and Theorem 3, fix $\eta_j := 1/j$, $j\in\mathbb{N}$. We may choose $k_j > K_j$ (for some $K_j > 0$) and $m_j > \max\{m_{j-1}, j, \varepsilon\}$ with $m_j \to \infty$ such that $(f(\mathbf{x}_0+m_j\boldsymbol{v}_{k_j})-f(\mathbf{x}_0))/m_j \geqslant \mathrm{Lip}(f)-\eta_j$. Let U be a uniform random variable on $[0,1]$, independent of $\boldsymbol{\zeta}$, and define

$$\boldsymbol{\xi}_j := \boldsymbol{\zeta} + m_j\boldsymbol{v}_{k_j}\mathbb{1}_{\{U\in A_j\}}, \quad \text{where} \quad A_j := [1-\varepsilon/m_j, 1).$$

One can verify that $\mathbb{E}\,\|\boldsymbol{\xi}_j-\boldsymbol{\zeta}\| = \varepsilon$. For each realization of $\boldsymbol{\zeta}$, denote by $\Delta_j := f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j})-f(\boldsymbol{\zeta})$. By Lipschitz continuity of f and the choice of m_j, k_j,

$$m_j(\mathrm{Lip}(f)-\eta_j) - 2\mathrm{Lip}(f)\|\boldsymbol{\zeta}-\mathbf{x}_0\| \leqslant \Delta_j \leqslant m_j\mathrm{Lip}(f) + 2\mathrm{Lip}(f)\|\boldsymbol{\zeta}-\mathbf{x}_0\|. \tag{A43}$$

Thus $\Delta_j/m_j \to \mathrm{Lip}(f)$ and $\Delta_j \to \infty$. Moreover, note that $\Delta_j \to \infty$, as $j\to\infty$. Recall that for each $t\in\mathbb{R}$ there exists $z_0(t)$ such that $\lim_{m\to\infty}\frac{\ell(z_0(t)+m,t)-\ell(z_0(t),t)}{m} = b$, which implies that $\lim_{m\to\infty}\frac{\ell(z+m,t)-\ell(z,t)}{m} = b$ for all $z,t\in\mathbb{R}$. Then, we have $\lim_{j\to\infty}\frac{\ell(f(\boldsymbol{\zeta})+\Delta_j,t)-\ell(f(\boldsymbol{\zeta}),t)}{\Delta_j} = b$. There-

fore, for any $t \in \mathbb{R}$ and sufficiently large $m_j > \max\{1, \varepsilon\}$,

$$\begin{aligned}
\sup_{\mathbb{E}[\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|]\leqslant\varepsilon} \left(\mathbb{E}\left[\ell(f(\boldsymbol{\xi}),t)-\ell(f(\boldsymbol{\zeta}),t)\right]\right) &\geqslant \mathbb{E}\left[\ell(f(\boldsymbol{\xi}_j),t)-\ell(f(\boldsymbol{\zeta}),t)\right] \\
&= \mathbb{E}\left[\left(\ell(f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j}),t)-\ell(f(\boldsymbol{\zeta}),t)\right)\mathbb{1}_{A_j}\right] \\
&= \mathbb{E}\left[m_j\frac{\Delta_j}{m_j}\frac{(\ell(f(\boldsymbol{\zeta})+\Delta_j,t)-\ell(f(\boldsymbol{\zeta}),t))}{\Delta_j}\mathbb{1}_{A_j}\right] \\
&= \mathbb{E}\left[\frac{\Delta_j}{m_j}\frac{(\ell(f(\boldsymbol{\zeta})+\Delta_j,t)-\ell(f(\boldsymbol{\zeta}),t))}{\Delta_j}\right]\mathbb{E}\left[m_j\mathbb{1}_{A_j}\right] \\
&= \varepsilon\mathbb{E}\left[\frac{\Delta_j}{m_j}\frac{(\ell(f(\boldsymbol{\zeta})+\Delta_j,t)-\ell(f(\boldsymbol{\zeta}),t))}{\Delta_j}\right] \to b\mathrm{Lip}(f)\varepsilon \text{ as } j\to\infty,
\end{aligned} \tag{A44}$$

where the third equality uses the independence of U and $\boldsymbol{\zeta}$, and the limit follows from the dominated convergence theorem, since

$$\left|\frac{\Delta_j}{m_j}\frac{(\ell(f(\boldsymbol{\zeta})+\Delta_j,t)-\ell(f(\boldsymbol{\zeta}),t))}{\Delta_j}\right| \leqslant b\frac{|\Delta_j|}{m_j} \leqslant b(\mathrm{Lip}(f)+2\mathrm{Lip}(f)\|\boldsymbol{\zeta}-\mathbf{x}_0\|),$$

with the last inequality from (A43) and $m_j > 1$. Combining (A44) with (A42) and (A41), we obtain (30), completing the proof of (i).

(ii) For $\widetilde{\ell}(\mathbf{z},t) := \ell(f(\mathbf{z}),t)$, the function $\widetilde{\ell}(\mathbf{z},t)$ is convex in t and satisfies $\lim_{t\to-\infty}\partial\widetilde{\ell}(\mathbf{z},t)/\partial t < -1 < \lim_{t\to\infty}\partial\widetilde{\ell}(\mathbf{z},t)/\partial t$ for all $\mathbf{z}\in\mathbb{R}^n$. Hence, $\lim_{t\to\pm\infty}\left(t+\mathbb{E}^F[\widetilde{\ell}(\boldsymbol{\xi},t)]\right)=\infty$ for all $F\in\mathcal{M}_1(\mathbb{R}^n)$. Moreover, $\widetilde{\ell}(\mathbf{z},t)$ is Lipschitz continuous in $\mathbf{z}$ with constant $b\mathrm{Lip}(f)$ for all t, the map $(t,F)\mapsto t+\mathbb{E}^F[\widetilde{\ell}(\boldsymbol{\xi},t)]$ is concave in F for all $t\in\mathbb{R}$ and convex in t for all $F\in\mathcal{M}_p(\mathbb{R}^n)$. Therefore, by the minimax theorem (see, e.g., Sion (1958)) and arguments analogous to those in the proof of Lemma A4 (i), we obtain

$$\begin{aligned}
\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)}\mathcal{H}_1^F(f(\boldsymbol{\xi})) &= \sup_{F\in\mathbb{B}_1(F_0,\varepsilon)}\inf_{t\in\mathbb{R}}\left(t+\mathbb{E}^F\left[\widetilde{\ell}(\boldsymbol{\xi},t)\right]\right) \\
&= \inf_{t\in\mathbb{R}}\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)}\left(t+\mathbb{E}^F\left[\widetilde{\ell}(\boldsymbol{\xi},t)\right]\right) = \inf_{t\in\mathbb{R}}\left(t+\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)}\mathbb{E}^F\left[\ell(f(\boldsymbol{\xi}),t)\right]\right).
\end{aligned} \tag{A45}$$

Then, by arguments analogous to those in the proof of (i),

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)}\left(\mathbb{E}^F\left[\ell(f(\boldsymbol{\xi}),t)\right]\right) = \mathbb{E}^{F_0}\left[\ell(f(\boldsymbol{\zeta}),t)\right]+b\mathrm{Lip}(f)\varepsilon.$$

Therefore, combining this with (A45), we obtain (31). This completes the proof. □

Proof of Proposition 6. Without loss of generality, assume $\mathrm{Lip}(f) = 1$. Note that (32) is equivalent to

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|\leqslant\varepsilon} \rho_h\left(f(\boldsymbol{\xi})\right) = \rho_h^{F_0}(f(\boldsymbol{\zeta})) + \varepsilon\|h'\|_\infty.$$

Then, to show (32), it suffices to prove that for any $\boldsymbol{\zeta} \sim F_0$ and $\varepsilon \geqslant 0$,

$$\sup_{\mathbb{E}\|\boldsymbol{\xi}\|\leqslant\varepsilon} \rho_h\left(f(\boldsymbol{\xi}+\boldsymbol{\zeta})\right) = \rho_h^{F_0}(f(\boldsymbol{\zeta})) + \varepsilon\|h'\|_\infty. \tag{A46}$$

To see it, first note that

$$\begin{aligned}\sup_{\mathbb{E}\|\boldsymbol{\xi}\|\leqslant\varepsilon} \rho_h(f(\boldsymbol{\xi}+\boldsymbol{\zeta})) &\leqslant \sup_{\mathbb{E}\|\boldsymbol{\xi}\|\leqslant\varepsilon} \rho_h(f(\boldsymbol{\zeta})+\|\boldsymbol{\xi}\|) \leqslant \rho_h(f(\boldsymbol{\zeta})) + \sup_{\mathbb{E}\|\boldsymbol{\xi}\|\leqslant\varepsilon} \rho_h(\|\boldsymbol{\xi}\|) \\ &\leqslant \rho_h(f(\boldsymbol{\zeta})) + \sup_{\mathbb{E}\|\boldsymbol{\xi}\|\leqslant\varepsilon} \left\|h'\right\|_\infty \mathbb{E}\|\boldsymbol{\xi}\| \leqslant \rho_h(\ell(Z)) + \left\|h'\right\|_\infty \varepsilon,\end{aligned}$$

where the first inequality follows from $f(\boldsymbol{\xi}+\boldsymbol{\zeta}) - f(\boldsymbol{\zeta}) \leqslant \|\boldsymbol{\xi}\|$ and the monotonicity of ρ_h, the second inequality follows from the subadditivity of ρ_h, and the third inequality is due to Hölder's inequality. Let us now verify the other direction. With arguments similar to those in the proof of Lemma A2 and Theorem 3, for each $\eta_j := 1/j$, $j \in \mathbb{N}$, we can choose $k_j > K_j$ for some $K_j > 0$ and $m_j > \max\{m_{j-1}, j, \varepsilon/(1-\alpha)\}$, with $m_j \to \infty$, such that $(f(\mathbf{x}_0 + m_j\boldsymbol{v}_{k_j}) - f(\mathbf{x}_0))/m_j \geqslant 1-\eta_j$. Denote by U a uniform random variable on $[0,1]$ such that U and $f(\boldsymbol{\zeta})$ are comonotonic. For chosen η_j, k_j, and m_j, define

$$A_j := \{\omega : 1 - \varepsilon/m_j < U(\omega) \leqslant 1\} \quad \text{and} \quad \mathbf{V}_j := m_j\boldsymbol{v}_{k_j}\mathbb{1}_{A_j}.$$

One can verify that $\mathbb{E}\,\|\mathbf{V}_j\| = \varepsilon$. Moreover, we have

$$\begin{aligned}\rho_h\left(f\left(\boldsymbol{\zeta}+\mathbf{V}_j\right)\right) &= \rho_h\left(f(\boldsymbol{\zeta})\mathbb{1}_{A_j^c} + f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j})\mathbb{1}_{A_j}\right) \\ &= \rho_h\left(f(\boldsymbol{\zeta}) + (f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}))\mathbb{1}_{A_j}\right) \\ &\geqslant \mathbb{E}\left[\left(f(\boldsymbol{\zeta}) + (f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}))\mathbb{1}_{A_j}\right)h'(U)\right] \\ &= \rho_h(f(\boldsymbol{\zeta})) + \mathbb{E}\left[(f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}))\mathbb{1}_{A_{m,\eta}}h'(U)\right],\end{aligned}$$

where the inequality follows from the dual representation of ρ_h (see e.g., Theorem 4.79 of Föllmer and Schied (2016)). Note that $f(\boldsymbol{\zeta})$ must be uniformly bounded on A_j for all sufficiently large

$m > \varepsilon/(1-\eta)$ as $f(\boldsymbol{\zeta})$ and U are comonotonic. Then, for m_j sufficiently large, we have

$$\begin{aligned}\sup_{\mathbb{E}\|\boldsymbol{\xi}\|\leqslant\varepsilon} \rho_h(f(\boldsymbol{\xi}+\boldsymbol{\zeta})) - \rho_h(f(\boldsymbol{\zeta})) &\geqslant \mathbb{E}\left[(f(\boldsymbol{\zeta}+m_j\boldsymbol{v}_{k_j}) - f(\boldsymbol{\zeta}))\mathbb{1}_{A_j}h'(U)\right] \\ &\geqslant (1-\eta_j)\mathbb{E}\left[m_j\mathbb{1}_{A_j}h'(U)\right] \\ &= (1-\eta_j)\frac{\int_{1-\varepsilon/m_j}^1 h'(s)\mathrm{d}s}{\varepsilon/m_j}\varepsilon \to \left\|h'\right\|_\infty \varepsilon \text{ as } j\to\infty.\end{aligned}$$

This completes the proof. □

Proof of Proposition 7. Note that we can rewrite $\mathrm{ex}_\alpha^F(X) = \max\{x\in\mathbb{R} : \mathbb{E}^F[\ell_\alpha(X-x)] \geqslant 0\}$, where $\ell_\alpha(x) := \alpha x_+ - (1-\alpha)x_-$, $x\in\mathbb{R}$. Denote by $c_f = \mathrm{Lip}(f)$. Therefore we have

$$\begin{aligned}\sup_{G\in\mathcal{C}_1(f|F_0,c_f\varepsilon)} \mathrm{ex}_\alpha^G(X) &= \sup_{G\in\mathcal{C}_1(f|F_0,c_f\varepsilon)} \max\left\{x\in\mathbb{R} : \mathbb{E}^G[\ell_\alpha(X-x)] \geqslant 0\right\} \\ &= \max\left\{x\in\mathbb{R} : \sup_{G\in\mathcal{C}_1(f|F_0,c_f\varepsilon)} \mathbb{E}^G[\ell_\alpha(X-x)] \geqslant 0\right\} \\ &= \max\left\{x : \mathbb{E}^{F_0}[\ell_\alpha(f(\boldsymbol{\zeta})-x)] + \alpha c_f\varepsilon \geqslant 0\right\},\end{aligned}$$

where the last equality follows from the regularization result of a convex Lipschitz continuous function over a Wasserstein ball, i.e., $\sup_{G\in\mathbb{B}_1(G_0,\varepsilon)} \mathbb{E}^G[\ell_\alpha(X-x)] = \mathbb{E}^{G_0}[\ell_\alpha(X-x)] + \alpha\varepsilon$. This implies

$$\sup_{F\in\mathbb{B}_1(F_0,\varepsilon)} \mathrm{ex}_\alpha^F\left(f\left(\boldsymbol{\xi}\right)\right) = \max\left\{x : \mathbb{E}^{F_0}[\ell_\alpha(f(\boldsymbol{\zeta})-x)] + \alpha c_f\varepsilon \geqslant 0\right\},$$

or equivalently, the unique solution to $\mathbb{E}^{F_0}[\ell_\alpha(f(\boldsymbol{\zeta})-x)] + \alpha c_f\varepsilon = 0$. This completes the proof. □

A.2 Proofs for Section 5

Proof of Proposition 8. Note that the implication (ii) ⇒ (i) is trivial. We only give the proof of (iii) ⇒ (ii) and (i) ⇒ (iii).

For (iii) ⇒ (ii), let $c_f := \mathrm{Lip}(f)$. If $\mathrm{Lip}(f) = 0$, then f is constant. There exists $b\in\mathbb{R}$ such that $f(\mathbf{x}) \equiv b$ for $\mathbf{x}\in\mathbb{R}^n$. In this case, for any $F_0 \in \mathcal{M}(\Xi)$ with $(Y_0, \mathbf{X}_0) \in F_0$, denote by F_{Y_0} the marginal distribution of Y_0. Then we have

$$\{F_{bY} :\ F_{(Y,\mathbf{X})} \in \overline{\mathbb{B}}_p(F_0,\varepsilon)\} = \{F_{bY_0}\} = \overline{\mathcal{C}}_p(f|F_0,0),$$

for any $F_0 \in \mathcal{M}(\mathbb{R}^n)$ and $\varepsilon > 0$, where F_{bY_0} is the distribution of bY_0 with $Y_0 \sim F_{Y_0}$. If $\mathrm{Lip}(f) > 0$,

we assume without loss generality that $\mathrm{Lip}(f) = 1$. Then, since f satisfies (9), it follows that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$,

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leqslant \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Now let $\boldsymbol{\xi} = (Y, \mathbf{X})$, $\boldsymbol{\zeta} = (Y_0, \mathbf{X}_0)$ be such that $\mathbb{E}[d(\boldsymbol{\xi}, \boldsymbol{\zeta})^p] \leqslant \varepsilon^p$, where and the cost d is given by (34). Then

$$\mathbb{E}\left[|Y f(\mathbf{X}) - Y_0 f(\mathbf{X}_0)|^p\right] = \mathbb{E}\left[|f(\mathbf{X}) - f(\mathbf{X}_0)|^p\right] \leqslant \mathbb{E}[\|\mathbf{X} - \mathbf{X}_0\|^p] \leqslant \varepsilon^p,$$

where the first equality follows from the definition in (34) which implies $Y = Y_0$ almost surely whenever $\mathbb{E}\left[d\left((Y, \mathbf{X}), (Y_0, \mathbf{X}_0)\right)^p\right] \leqslant \varepsilon^p$ and $|Y_0| = 1$ a.s., and the inequality follows from the 1-Lipschitz continuity of f. It follows that $\{F_{Yf(\mathbf{X})} : F_{(Y,\mathbf{X})} \in \overline{\mathbb{B}}_p(F_0, \varepsilon)\} \subseteq \overline{\mathcal{C}}_p(f|F_0, \varepsilon)$, To prove the reverse inclusion, it suffices to show that for any $G \in \overline{\mathcal{C}}_p(f|F_0, \varepsilon)$ and $Z \sim G$ with $\mathbb{E}[|Z - Y_0 f(\mathbf{X}_0)|^p] \leqslant \varepsilon^p$, there exists a random vector $(Y, \mathbf{X})$ with distribution $F_{(Y,\mathbf{X})} \in \overline{\mathbb{B}}_p(F_0, \varepsilon)$ such that $Y f(\mathbf{X}) = Z$ almost surely. To this end, let $T := Z/Y_0 - f(\boldsymbol{\zeta})$. Then

$$\mathbb{E}[|T|^p] = \mathbb{E}[|\frac{Z}{Y_0} - f(\boldsymbol{\zeta})|^p] = \mathbb{E}[|Z - Y_0 f(\mathbf{X}_0)|^p] \leqslant \varepsilon^p.$$

By Rieder (1978) and following similar arguments as in the proof of Theorem 1 (iii) $\Rightarrow$ (ii), there exist measurable mappings $\mathbf{V}_1$ and $\mathbf{V}_2$ such that

$$\mathbf{V}_1(\omega) \in \underset{\|\mathbf{y}\| \leqslant |T(\omega)|}{\arg\max}\ f\left(\boldsymbol{\zeta}(\omega) + \mathbf{y}\right) \quad \text{and} \quad \mathbf{V}_2(\omega) \in \underset{\|\mathbf{y}\| \leqslant |T(\omega)|}{\arg\min}\ f\left(\boldsymbol{\zeta}(\omega) + \mathbf{y}\right), \quad \omega \in \Omega.$$

Denote by $A_+ := \{\omega : T(\omega) \geqslant 0\}$ and $A_- := \{\omega : T(\omega) < 0\}$. We define $Y = Y_0$ and $\mathbf{X} = (\mathbf{X}_0 + \mathbf{V}_1)\mathbb{1}_{A_+} + (\mathbf{X}_0 + \mathbf{V}_2)\mathbb{1}_{A_-}$. Then by (8), for each realization,

$$Y(\omega) f(\mathbf{X}(\omega)) = Y_0(\omega) f(\mathbf{X}_0(\omega)) + Y_0(\omega) T(\omega) = Z(\omega), \quad \omega \in A_+,$$

and

$$Y(\omega) f(\mathbf{X}(\omega)) = Y_0(\omega) f(\mathbf{X}_0(\omega)) - Y_0(\omega)(-T(\omega)) = Z(\omega), \quad \omega \in A_-.$$

Moreover,

$$\mathbb{E}\left[d\left((Y, \mathbf{X}), (Y_0, \mathbf{X}_0)\right)^p\right] = \mathbb{E}[\|\mathbf{X} - \mathbf{X}_0\|^p] \leqslant \mathbb{E}[(\max\{\|\mathbf{V}_1\|, \|\mathbf{V}_2\|\})^p] \leqslant \mathbb{E}[|T|^p] \leqslant \varepsilon^p,$$

where the first inequality follows from the definition of $\mathbf{X}$. This implies $F_{(Y,\mathbf{X})} \in \overline{\mathbb{B}}_p(F_0, \varepsilon)$. Therefore, $\overline{\mathcal{C}}_p(f|F_0, \varepsilon) \subseteq \{F_{Yf(\mathbf{X})} : F_{(Y,\mathbf{X})} \in \overline{\mathbb{B}}_p(F_0, \varepsilon)\}$ and thus, (38) holds, completing the proof of this direction.

For (i) $\Rightarrow$ (iii), suppose that there exists $c_f \geqslant 0$ such that (37) holds for any $\varepsilon > 0$ and $F_0 \in \mathcal{M}(\Xi)$. If $c_f = 0$, take $\rho = \mathrm{VaR}_\alpha$ with $\alpha \in [0, 1)$, and let $F_{\mathbf{X}_0} \in \mathcal{M}(\mathbb{R}^n)$ be any marginal distribution. Define F_0 such that $(Y_0, \mathbf{X}_0) \sim F_0$ with $Y_0 \equiv 1$ almost surely and $\mathbf{X}_0 \sim F_{\mathbf{X}_0}$. Then, by arguments analogous to those used in the proof of Theorem 1 (i) $\Rightarrow$ (iii), it follows that f must be constant. Now consider $c_f > 0$ and take $\rho = \mathrm{VaR}_\alpha$ with $\alpha \in [0, 1)$. By the positive homogeneity of VaR, we assume without loss of generality that $c_f = 1$. Consider any marginal distribution $F_{\mathbf{X}_0} \in \mathcal{M}(\mathbb{R}^n)$, and define F_0 such that $(Y_0, \mathbf{X}_0) \sim F_0$ with $Y_0 \equiv 1$ almost surely and $\mathbf{X}_0 \sim F_{\mathbf{X}_0}$. In this case, (37) reduces to

$$\sup_{F \in \mathbb{B}_p(F_{\mathbf{X}_0}, \varepsilon)} \mathrm{VaR}_\alpha^F(f(\mathbf{X})) = \sup_{G \in \mathcal{C}_p(f|F_{\mathbf{X}_0}, \varepsilon)} \mathrm{VaR}_\alpha^G(X),$$

which holds for any $\varepsilon > 0$ and $F_{\mathbf{X}_0} \in \mathcal{M}(\mathbb{R}^n)$. By Lemma A1 and the proof of Theorem 1 (i) $\Rightarrow$ (iii), it follows that

$$\sup_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \varepsilon, \tag{A47}$$

for any $\mathbf{x} \in \mathbb{R}^n$ and $\varepsilon > 0$. Similarly, define F_0 such that $Y_0 \equiv -1$ almost surely and $\mathbf{X}_0 \sim F_{\mathbf{X}_0}$. Then, we have

$$\sup_{F \in \mathbb{B}_p(F_{\mathbf{X}_0}, \varepsilon)} \mathrm{VaR}_\alpha^F(-f(\mathbf{X})) = \sup_{G \in \mathcal{C}_p(-f|F_{\mathbf{X}_0}, \varepsilon)} \mathrm{VaR}_\alpha^G(X),$$

which holds for any $\varepsilon > 0$ and $F_{\mathbf{X}_0} \in \mathcal{M}(\mathbb{R}^n)$. By Lemma A1, one can verify that

$$f(\mathbf{x}) - \inf_{\|\mathbf{y}\| \leqslant \varepsilon} f(\mathbf{x} + \mathbf{y}) = \varepsilon, \tag{A48}$$

for any $\mathbf{x} \in \mathbb{R}^n$ and $\varepsilon > 0$. Combined with (A47) and (A48) yields that f satisfies (9). This completes the proof. □

Proof of Proposition 9. For the "if" part, by Proposition 8, we can take $\rho = \mathrm{VaR}_\alpha$ for some $\alpha \in [0, 1)$, which implies that (39) holds.

For the "only if" part, applying arguments similar to those in the proof of Proposition 8, direction (i) $\Rightarrow$ (iii), the desired result follows. This completes the proof. □

Proof of Corollary 4. By Proposition 1, Proposition 2, Proposition 8, and Theorem 5, the conclu-

sion follows immediately. □

References

Aolaritei, L., Lanzetti, N., Chen, H., and Dörfler, F. (2023). Distributional uncertainty propagation via optimal transport *arXiv preprint arXiv:2205.00343*

Bellini, F., Klar, B., Müller, A., and Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54, 41–8.

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.

Chen, Z., Kuhn, D., and Wiesemann, W. (2024). Data-driven chance constrained programs over Wasserstein balls. *Operations Research*, **72**(1), 410–424.

Chen, Z., and Xie, W. (2021). Sharing the value-at-risk under distributional ambiguity. *Mathematical Finance*, **31**(1), 531–559.

Clarkson, J. A. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*, **40**(3), 396-414.

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

Filipović, D. and Svindland, G. (2007). Convex risk measures on Lp. *Preprint, www. math. lmu. de/filipo/PAPERS/crmlp*.

Gao, R., Chen, X., and Kleywegt, A. J. (2024). Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, **72**(3), 1177–1191.

Ho-Nguyen, N., and Wright, S. J. (2023). Adversarial classification via distributional robustness with Wasserstein ambiguity. *Mathematical Programming*, **198**(2), 1411-1447.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research and Management Science in the age of analytics*. Informs, 2019. 130-166.

Kusuoka, S. (2001). On law invariant coherent risk measures. *Advances in mathematical economics*. Tokyo: Springer Japan, 2001. 83-95.

Liu, F., Mao, T., Wang, R., and Wei, L. (2022). Inf-convolution, optimal allocations, and model uncertainty for tail risk measures. *Mathematics of Operations Research*, **47**(3), 2494–2519.

Mao, T., Wang, R., and Wu, Q. (2022). Model aggregation for risk evaluation and robust optimization. arXiv preprint arXiv:2201.06370.

Pflug, G. C., Pichler, A., and Wozabal, D. (2012). The 1/N investment strategy is optimal under high model ambiguity. *Journal of Banking and Finance*, **36**(2), 410–417.

Rieder, U. (1978). Measurable selection theorems for optimization problems. *manuscripta mathematica*, **24**(1), 115–131.

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

Santambrogio, F. (2015). Optimal transport for applied mathematicians.

Shafieezadeh-Abadeh, S., Kuhn, D., Esfahani, P.M. (2019) Regularization via mass transportation. *Journal of Machine Learning Research*, **20**(103), 1–68

Shapiro, A. (2013). On Kusuoka representation of law invariant risk measures. *Mathematics of Operations Research*, **38**(1), 124–152.

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

Wu, Q., Li, J. Y. M., and Mao, T. (2022). On generalization and regularization via wasserstein distributionally robust optimization. *arXiv preprint arXiv:2212.05716*.

Xie, W. (2021). On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming*, **186**(1), 115–155.