https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–10
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

A Short and General Duality Proof for Wasserstein Distributionally Robust Optimization

Luhao Zhang,[a] Jincheng Yang,[b] Rui Gao[c,*]

[a] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland 21218; [b] Department of Mathematics, University of Chicago, Chicago, Illinois 60637; [c] Department of Information, Risk and Operations Management, The University of Texas at Austin, Austin, Texas 78712

*Corresponding author

Contact: luhao.zhang@jhu.edu, https://orcid.org/0000-0001-8568-3581 (LZ); jincheng@uchicago.edu, https://orcid.org/0000-0002-3581-9425 (JY); rui.gao@mccombs.utexas.edu, https://orcid.org/0000-0003-0145-8577 (RG)

Received: March 20, 2023
Revised: April 13, 2024
Accepted: May 21, 2024
Published Online in Articles in Advance: July 15, 2024

Area of Review: Optimization

https://doi.org/10.1287/opre.2023.0135

Abstract. We present a general duality result for Wasserstein distributionally robust optimization that holds for any Kantorovich transport cost, measurable loss function, and nominal probability distribution. Assuming an interchangeability principle inherent in existing duality results, our proof only uses one-dimensional convex analysis. Furthermore, we demonstrate that the interchangeability principle holds if and only if certain measurable projection and weak measurable selection conditions are satisfied. To illustrate the broader applicability of our approach, we provide a rigorous treatment of duality results in distributionally robust Markov decision processes and distributionally robust multistage stochastic programming. Additionally, we extend our analysis to other problems such as infinity-Wasserstein distributionally robust optimization, risk-averse optimization, and globalized distributionally robust counterpart.

Funding: L. Zhang acknowledges the support of Xunyu Zhou and the Nie Center for Intelligent Asset Management at Columbia University.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2023.0135.

Keywords: Wasserstein metric • distributionally robust optimization • duality

1. Introduction

In this paper, we consider the following problem

$$\mathcal{L}(\rho) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] : \mathcal{K}_c(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\}, \qquad \text{(P)}$$

where $\rho \in [0,\infty)$, $\mathcal{P}(\mathcal{X})$ is the set of all probability distributions on a data space $\mathcal{X}$, $f:\mathcal{X}\to\mathbb{R}$ is a loss function, X is a random variable on $\mathcal{X}$ having a nominal distribution $\widehat{\mathbb{P}}$, and $\mathcal{K}_c$ denotes the Kantorovich transport cost, defined as

$$\mathcal{K}_c(\widehat{\mathbb{P}},\mathbb{P}) = \inf_{\gamma\in\Gamma(\widehat{\mathbb{P}},\mathbb{P})} \mathbb{E}_{(\widehat{X},X)\sim\gamma}[c(\widehat{X},X)], \qquad (1)$$

where $\Gamma(\widehat{\mathbb{P}},\mathbb{P})$ denotes the set of all probability distributions on $\mathcal{X}\times\mathcal{X}$ with marginals $\widehat{\mathbb{P}}$ and $\mathbb{P}$, and $c:\mathcal{X}\times\mathcal{X}\to[0,\infty]$ is a transport cost function. Note that $\mathcal{K}_c$ is a distance on $\mathcal{P}(\mathcal{X})$ when c is a metric. The function $\mathcal{L}$ represents the robust loss hedging against deviations of data within ρ-neighborhood of the nominal distribution $\widehat{\mathbb{P}}$. When $c=d^p$, where d is a metric on $\mathcal{X}$ and $p\in[1,\infty)$, Problem (P) is the inner worst-case problem in p-Wasserstein distributionally robust optimization (DRO), which has raised much interest recently (Kuhn et al. 2019, Blanchet et al. 2021).

A central question of interest is developing the dual problem for (P). In this paper, we present a novel proof that yields results held in more general settings than those found in the literature. Assuming the interchangeability principle—a condition inherent in existing results—our proof is significantly shorter. The key idea is to view the robust loss as a function of the radius of the uncertainty set and then apply the Legendre transform to it twice. The concavity of the robust loss enables us to establish strong duality directly through the Legendre transformation. Compared with existing duality proofs that relied on convexity duality in conic programming (Mohajerin Esfahani and Kuhn 2018, Zhao and Guan 2018, Zhen et al. 2023) or vector spaces (Blanchet and Murthy 2019), our proof only uses one-dimensional convex analysis. For detailed comparisons, we refer to Table 1 and the discussion following Theorem 1. Furthermore, we explore the interchangeability principle in depth and provide an equivalent condition, which has a strong connection to the measurable projection theorem and the measurable selection theorem. As a result, establishing the duality boils down to verifying the measurability conditions, which have been studied in broader settings.

Table 1. Comparison with Existing Duality Results

	Mohajerin Esfahani and Kuhn (2018)	Zhao and Guan (2018)	Blanchet and Murthy (2019)	Gao and Kleywegt (2023)	Zhen et al. (2023)	This paper
c	Norm	Continuous	Lower semicontinuous	Power of metric	Convex	Arbitrary
f	Piecewise concave	Bounded	Upper semicontinuous	Arbitrary	Piecewise concave	Arbitrary
$\widehat{\mathbb{P}}$	Empirical	Empirical	Borel	Borel	Empirical	(Interchangeability principle)
$\mathcal{X}$	Convex subset of $\mathbb{R}^d$	Convex compact	Polish	Polish	Convex subset of $\mathbb{R}^d$	
Proof	Conic duality	Conic duality	Approximation argument	Constructive	Slater condition	Legendre transform

To showcase its potential applications, we develop duality results for distributionally robust Markov decision processes, distributionally robust multistage stochastic programming, infinity-Wasserstein distributionally robust optimization, risk-averse optimization, and globalized distributionally robust counterpart.

The remainder of this paper is structured as follows. In Section 2, we present our main proof. In Section 3, we provide a verification result of the interchangeability principle. We provide several examples in Section 4 and extend our results to other distributional robust problems in Section 5. Finally, we conclude the paper in Section 6.

2. Model Formulation and Main Result

In this section, we state and prove our main duality results under general assumptions. Proofs of Auxiliary lemmas are listed in Appendix A.

2.1. Main Assumptions

Notations. We denote by $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ the extended reals and adopt the convention that $0 \cdot \infty = \infty$. For a probability space $(\mathcal{X}, \mathscr{F}, \widehat{\mathbb{P}})$, we say an extended real-valued function on $(\mathcal{X}, \mathscr{F})$ is measurable if it is $(\mathscr{F}, \mathscr{B}(\bar{\mathbb{R}}))$-measurable, where $\mathscr{B}(\bar{\mathbb{R}})$ is the Borel σ-algebra on $\bar{\mathbb{R}}$, and we say an extended real-valued function on $\mathcal{X}$ is $\widehat{\mathbb{P}}$-measurable if it is measurable with respect to the completion of $\mathscr{F}$ under the measure $\widehat{\mathbb{P}}$ (Ambrosio et al. 2000, definition 1.11). We allow the expectation to take values in $\bar{\mathbb{R}}$—recall that the integration of a measurable function under a measure is well defined whenever the positive part or the negative part of the integrand has a finite integral. When $\mathcal{X}$ is a metric space, we denote by $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ its metric. Let $\bar{\mathcal{P}}(\mathcal{X})$ denote the set of probability measures $\mathbb{P}$ on $(\mathcal{X}, \mathscr{F})$ satisfying $\mathcal{K}_c(\widehat{\mathbb{P}}, \mathbb{P}) < \infty$. For a function $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, we denote by $h^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ its Legendre transform $h^*(\lambda) := \sup_{\rho \in \mathbb{R}} \{\lambda\rho - h(\rho)\}$. If $h : \mathbb{R} \to \bar{\mathbb{R}}$ attains $-\infty$ somewhere, then $h^* \equiv +\infty$. In addition to Problem (P), we also study its soft-penalty counterpart

$$\sup_{\mathbb{P} \in \bar{\mathcal{P}}} \{\mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \lambda \mathcal{K}_c(\widehat{\mathbb{P}}, \mathbb{P})\}, \tag{P-soft}$$

where $\lambda \in [0, \infty)$.

We assume the following situation.

Assumption 1. *Let $(\mathcal{X}, \mathscr{F}, \widehat{\mathbb{P}})$ be a probability space, $f : \mathcal{X} \to \mathbb{R}$ be a measurable function with $\mathbb{E}_{\widehat{\mathbb{P}}}[f] > -\infty$, and $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ be a measurable transport cost function satisfying $c(x, x) = 0$ for all $x \in \mathcal{X}$.*

The following lemma shows some useful properties of the worst-case loss $\mathcal{L}(\cdot)$ defined in (P).

Lemma 1. *Assume Assumption 1 holds. Then $\mathcal{L}(\cdot)$ is lower bounded by $\mathbb{E}_{\widehat{\mathbb{P}}}[f]$, monotonically increasing, and concave on $[0, \infty)$.*

As we will see, the interchangeability principle discussed next is essential for strong duality, being both a necessary and sufficient condition. This principle is often mentioned in stochastic programming literature, as seen in references like Shapiro et al. (2021, section 9.3.4) and Shapiro (2017). It facilitates the swapping of expectation and minimization operators in our analysis.

Interchangability Principle (IP). *We say an $(\mathscr{F} \otimes \mathscr{F})$-measurable function $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ satisfies the interchangeability principle if the function $\widehat{x} \mapsto \sup_{x \in \mathcal{X}} \phi(\widehat{x}, x)$ is $\widehat{\mathbb{P}}$-measurable and it holds that*

$$\mathbb{E}_{\widehat{X} \sim \widehat{\mathbb{P}}}\left[\sup_{x \in \mathcal{X}} \phi(\widehat{X}, x)\right] = \sup_{\gamma \in \Gamma_{\widehat{\mathbb{P}}}} \mathbb{E}_{(\widehat{X}, X) \sim \gamma}[\phi(\widehat{X}, X)],$$

where $\Gamma_{\widehat{\mathbb{P}}}$ is the set of probability distributions on $(\mathcal{X} \times \mathcal{X}, \mathscr{F} \otimes \mathscr{F})$ with first marginal $\widehat{\mathbb{P}}$.

2.2. Main Result and Its Proof

Using Lemma 1, we derive the dual of (P) as follows.

Theorem 1. *Assume Assumption 1 holds. Let $\lambda, \rho > 0$. Then (P-soft) is equivalent to*

$$(-\mathcal{L})^*(-\lambda) = \sup_{\gamma \in \Gamma_{\widehat{\mathbb{P}}}} \mathbb{E}_{(\widehat{X}, X) \sim \gamma}[f(X) - \lambda c(\widehat{X}, X)],$$

and (P) is equivalent to

$$\mathcal{L}(\rho) = \min_{\lambda \geq 0} \left\{ \lambda\rho + \sup_{\gamma \in \Gamma_{\widehat{\mathbb{P}}}} \{\mathbb{E}_{(\widehat{X}, X) \sim \gamma}[f(X) - \lambda c(\widehat{X}, X)]\} \right\}.$$

In addition, for $\lambda > 0$, if and only if the function $\phi_\lambda(\widehat{x}, x) := f(x) - \lambda c(\widehat{x}, x)$ satisfies (IP), it holds that

$$(-\mathcal{L})^*(-\lambda) = \mathbb{E}_{\widehat{X} \sim \widehat{\mathbb{P}}}\left[\sup_{x \in \mathcal{X}} \{f(x) - \lambda c(\widehat{X}, x)\}\right],$$

and if and only if ϕ_λ satisfies (IP) for every $\lambda > 0$, it holds that

$$\mathcal{L}(\rho) = \min_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}} \left[\sup_{x\in\mathcal{X}} \{f(x) - \lambda c(\hat{X}, x)\} \right] \right\}, \quad \forall \rho > 0. \tag{D}$$

Remark 1 (p-Wasserstein Distance). Recall that the p-Wasserstein distance $\mathcal{W}_p(\hat{\mathbb{P}}, \mathbb{P})$, $p \in [1, \infty)$, is defined by

$$\mathcal{W}_p(\hat{\mathbb{P}}, \mathbb{P}) = \inf_{\gamma\in\Gamma(\hat{\mathbb{P}}, \mathbb{P})} \|d\|_{L^p(\mathcal{X}\times\mathcal{X};\gamma)} = \inf_{\gamma\in\Gamma(\hat{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{(\hat{X}, X)\sim\gamma}[d(\hat{X}, X)^p]^{\frac{1}{p}}.$$

By setting $c(\hat{x}, x) = d(\hat{x}, x)^p$ we have $\mathcal{K}_c(\hat{\mathbb{P}}, \mathbb{P}) = \mathcal{W}_p^p(\hat{\mathbb{P}}, \mathbb{P})$. Thereby (D) corresponds to the dual formulation of the p-Wasserstein DRO

$$\mathcal{L}(\rho^p) = \min_{\lambda\geq 0} \left\{ \lambda\rho^p + \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}} \left[\sup_{x\in\mathcal{X}} \{f(x) - \lambda d(\hat{X}, x)^p\} \right] \right\}.$$

We will handle the case $p = \infty$ separately in Section 5.1.

Remark 2 (Necessity of (IP)). The second part of Theorem 1 discusses the necessity of the IP. As will be elaborated on in the next section, it ensures the measurability of the supremum function in (D) and the existence of approximately worst-case distributions. To our knowledge, (IP) is weaker than the assumptions in all existing results that enable the expression (D).

Remark 3 (Continuity at $\rho = 0$). In general, (D) does not hold at $\rho = 0$. Indeed, the right-hand side of (D) is continuous in $\rho \in [0, \infty)$, but $\mathcal{L}(\rho)$ may be not right-continuous at zero. For instance, if $\mathcal{X} = \mathbb{R}$, $c(\hat{x}, x) = |\hat{x} - x|$, $f(x) = \mathbf{1}\{x \neq 0\}$, and $\hat{\mathbb{P}} = \boldsymbol{\delta}_0$, the Dirac measure at zero, then $\mathcal{L}(\rho) = 1$ for any $\rho > 0$ and $\mathcal{L}(0) = 0$. A sufficient condition ensuring the right-continuity of $\mathcal{L}(\rho)$ at zero is the following: There exists a continuous concave function $\varphi : [0, \infty) \to [0, \infty)$ with $\varphi(0) = 0$ such that $f(x) - f(\hat{x}) \leq \varphi \circ c(\hat{x}, x)$ for all $x \in \mathcal{X}$ and $\hat{\mathbb{P}}$-a.e. $\hat{x} \in \mathcal{X}$ with $c(\hat{x}, x) < \infty$. Indeed, under this condition, for any $\epsilon > 0$ and $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ with $\mathcal{K}_c(\hat{\mathbb{P}}, \mathbb{P}) \leq \epsilon$, there exists a $\gamma \in \Gamma_{\hat{\mathbb{P}}}$ such that $\mathbb{E}_\gamma[c(\hat{X}, X)] \leq 2\epsilon$; hence, $\mathbb{E}_\gamma[f(X) - f(\hat{X})] \leq \mathbb{E}_\gamma[\varphi \circ c] \leq \varphi(\mathbb{E}_\gamma[c])$ by Jensen's inequality. Therefore, $\mathcal{L}(\epsilon) \leq \mathbb{E}_{\hat{\mathbb{P}}}[f] + \varphi(\mathbb{E}_\gamma[c]) \leq \mathcal{L}(0) + \varphi(\mathbb{E}_\gamma[c])$, which converges to $\mathcal{L}(0)$ as $\epsilon \to 0$. When $c = d^p$, where d is a metric on $\mathcal{X}$, this condition is related to the growth condition imposed in Gao and Kleywegt (2023) and the upper semicontinuity of f.

Proof of Theorem 1. Fix $\lambda > 0$. By definition, $\mathcal{L}(\rho) = -\infty$ for $\rho < 0$. Taking the Legendre transform of $-\mathcal{L}(\cdot)$ gives that

$$\begin{aligned}
&(-\mathcal{L})^*(-\lambda) \\
&= \sup_{\rho\geq 0}\{(-\lambda)\rho - (-\mathcal{L}(\rho))\} \\
&= \sup_{\rho\geq 0}\{\mathcal{L}(\rho) - \lambda\rho\} \\
&= \sup_{\rho\geq 0} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\rho : \mathcal{K}_c(\hat{\mathbb{P}}, \mathbb{P}) \leq \rho\} \\
&= \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \sup_{\rho\geq 0} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\rho : \mathcal{K}_c(\hat{\mathbb{P}}, \mathbb{P}) \leq \rho\} \\
&= \sup_{\mathbb{P}\in\bar{\mathcal{P}}}\{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\mathcal{K}_c(\hat{\mathbb{P}}, \mathbb{P})\},
\end{aligned}$$

which gives (P-soft). Using the definition of the Kantorovich transport cost (1), it follows that

$$\begin{aligned}
(-\mathcal{L})^*(-\lambda) &= \sup_{\mathbb{P}\in\bar{\mathcal{P}}}\{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\mathcal{K}_c(\hat{\mathbb{P}}, \mathbb{P})\} \\
&= \sup_{\mathbb{P}\in\bar{\mathcal{P}}} \left\{ \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda \inf_{\gamma\in\Gamma(\hat{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{(\hat{X}, X)\sim\gamma}[c(\hat{X}, X)] \right\} \\
&= \sup_{\mathbb{P}\in\bar{\mathcal{P}}, \gamma\in\Gamma(\hat{\mathbb{P}}, \mathbb{P})} \left\{ \mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\mathbb{E}_{(\hat{X}, X)\sim\gamma}[c(\hat{X}, X)] : \mathbb{E}_\gamma[c] < \infty \right\} \\
&= \sup_{\gamma\in\Gamma_{\hat{\mathbb{P}}}} \mathbb{E}_{(\hat{X}, X)\sim\gamma}[f(X) - \lambda c(\hat{X}, X)].
\end{aligned}$$

Observe that $(-\mathcal{L})^*(-\lambda) \geq \sup_{\rho\geq 0}\{(-\lambda)\rho + \mathcal{L}(0)\} = +\infty$ for $\lambda < 0$. From Lemma 1, we have that $\mathcal{L}(\cdot)$ is bounded from below, increasing and concave in $[0, \infty)$, so either $\mathcal{L}(\rho) < +\infty$ for every $\rho \geq 0$ or $\mathcal{L}(\rho) = +\infty$ for every $\rho > 0$. In the former case, by the involution property of Legendre transform (Rockafellar 1970, theorem 12.2), $(-\mathcal{L})^{**}$ equals to the lower semicontinuous convex envelope of $-\mathcal{L}$. Because $\mathcal{L}(\cdot)$ is concave in $[0, \infty)$, $-\mathcal{L}(\rho) = -\mathcal{L}^{**}(\rho)$ on the interior of the set $\{\rho \in \mathbb{R} : -\mathcal{L}(\rho) < +\infty\}$, which is $(0, \infty)$. Hence, for every $\rho > 0$,

$$\begin{aligned}
\mathcal{L}(\rho) &= -(-\mathcal{L})^{**}(\rho) = -\max_{\lambda\geq 0}\{(-\lambda)\rho - (-\mathcal{L})^*(-\lambda)\} \\
&= \min_{\lambda\geq 0}\{\lambda\rho + (-\mathcal{L})^*(-\lambda)\} \\
&= \min_{\lambda\geq 0} \left\{ \lambda\rho + \sup_{\gamma\in\Gamma_{\hat{\mathbb{P}}}}\{\mathbb{E}_{(\hat{X}, X)\sim\gamma}[f(X) - \lambda c(\hat{X}, X)]\} \right\}.
\end{aligned}$$

We switched from supremum to maximum because $(-\mathcal{L})^*(-\lambda)$ is lower semicontinuous and bounded from below for $\lambda \geq 0$ by Lemma A.1. Therefore, $(-\lambda)\rho - (-\mathcal{L})^*(-\lambda)$ can be arbitrarily small as $\lambda \to +\infty$; hence,

the maximum is attainable. In the latter case, $(-\mathcal{L})^*(-\lambda) = +\infty$ for any $\lambda \in \mathbb{R}$, so the previous is also true. This proves the first part of the theorem.

For the second part, for $\lambda > 0$, $\phi_\lambda = f - \lambda c$ satisfies (IP) means

$$\begin{aligned}(-\mathcal{L})^*(-\lambda) &= \sup_{\gamma \in \Gamma_{\hat{\mathbb{P}}}} \{\mathbb{E}_{(\hat{X},X)\sim\gamma}[f(X) - \lambda c(\widehat{X}, X)]\} \\ &= \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x\in\mathcal{X}}\{f(x) - \lambda c(\widehat{X}, x)\}\right] =: \mathcal{G}(\lambda),\end{aligned}$$

where $\mathcal{G}(\lambda)$ is defined previously for $\lambda \geq 0$ and $\mathcal{G}(\lambda) := (-\mathcal{L})^*(-\lambda) = +\infty$ for $\lambda < 0$. By Lemma A.1, $(-\mathcal{L})^*(-\lambda)$ and $\mathcal{G}(\lambda)$ are lower bounded by $\mathbb{E}_{\hat{\mathbb{P}}}[f]$, monotonically decreasing, convex, and lower semicontinuous on $[0,\infty)$, implying the right continuity at $\lambda = 0$. Hence, the following equivalence holds:

$$\begin{aligned}&\phi_\lambda \text{ satisfies (IP) for all } \lambda \in (0,\infty) \\ &\iff (-\mathcal{L})^*(-\lambda) = \mathcal{G}(\lambda) \text{ for all } \lambda \in (0,\infty) \\ &\iff (-\mathcal{L})^*(-\lambda) = \mathcal{G}(\lambda) \text{ for all } \lambda \in [0,\infty) \\ &\iff (-\mathcal{L})^*(\lambda) = \mathcal{G}(-\lambda) \text{ for all } \lambda \in \mathbb{R}.\end{aligned}$$

Here we separately consider the cases $\mathcal{L}(\rho) < +\infty$ and $\mathcal{L}(\rho) \equiv +\infty$ for $\rho \geq 0$. If $\mathcal{L}(\rho) < +\infty$ for every $\rho \geq 0$, then $(-\mathcal{L})^* \not\equiv +\infty$ by Lemma A.2, so by the involution property of Legendre transform, we have

$$\begin{aligned}&(-\mathcal{L})^*(\lambda) = \mathcal{G}(-\lambda) \text{ for all } \lambda \in \mathbb{R} \\ &\iff (-\mathcal{L})^{**}(\rho) = (\mathcal{G}(-\cdot))^*(\rho) = \mathcal{G}^*(-\rho) \text{ for all } \rho \in \mathbb{R} \\ &\iff (-\mathcal{L})^{**}(\rho) = \mathcal{G}^*(-\rho) \text{ for all } \rho \in [0,\infty) \\ &\iff (-\mathcal{L})^{**}(\rho) = \mathcal{G}^*(-\rho) \text{ for all } \rho \in (0,\infty) \\ &\iff \mathcal{L}(\rho) = \mathcal{G}^*(-\rho) = \inf_{\lambda\in\mathbb{R}}\{\lambda\rho + \mathcal{G}(\lambda)\} \\ &\quad = \min_{\lambda\geq 0}\{\lambda\rho + \mathcal{G}(\lambda)\} \text{ for all } \rho \in (0,\infty).\end{aligned}$$

Here we used the properties that $(-\mathcal{L})^{**}$ and $\mathcal{G}^*(-\cdot)$ are both right continuous at zero and are both $+\infty$ on $(-\infty, 0)$ from Lemma A.2. In the last step, the minimum is attainable because $\lambda\rho + \mathcal{G}(\lambda)$ is bounded from below and lower semicontinuous and becomes arbitrarily large as $\lambda \to \infty$. If $\mathcal{L}(\rho) \equiv +\infty$ for $\rho > 0$, then $(-\mathcal{L})^*(-\lambda) = +\infty$ for all $\lambda \in \mathbb{R}$, so

$$\begin{aligned}&(-\mathcal{L})^*(\lambda) = \mathcal{G}(-\lambda) \text{ for all } \lambda \in \mathbb{R} \\ &\iff \mathcal{G}(\lambda) = +\infty \text{ for all } \lambda \in \mathbb{R} \\ &\iff \mathcal{G}(\lambda) = +\infty \text{ for all } \lambda \geq 0 \\ &\iff \min_{\lambda\geq 0}\{\lambda\rho + \mathcal{G}(\lambda)\} = +\infty \text{ for all } \rho \in (0,\infty) \\ &\iff \min_{\lambda\geq 0}\{\lambda\rho + \mathcal{G}(\lambda)\} = \mathcal{L}(\rho) \text{ for all } \rho \in (0,\infty).\end{aligned}$$

In conclusion, ϕ_λ satisfies (IP) for all $\lambda \in (0,\infty)$ if and only if $\mathcal{L}(\rho) = \min_{\lambda\geq 0}\{\lambda\rho + \mathcal{G}(\lambda)\}$ for all $\rho \in (0,\infty)$. □

Let us compare our proof technique with existing duality results in the literature. Proofs in Mohajerin Esfahani and Kuhn (2018), Blanchet and Murthy (2019), Zhao and Guan (2018), Sinha et al. (2018), and Zhen et al. (2023) rely on advanced convex duality theory. More specifically, Mohajerin Esfahani and Kuhn (2018), Zhao and Guan (2018), and Zhen et al. (2023) exploit conic duality (Shapiro 2001) for the problem of moments. This approach requires the nominal distribution $\widehat{\mathbb{P}}$ to be finitely supported and the space $\mathcal{X}$ to be convex, along with additional assumptions on the transport cost c and the loss function f. Blanchet and Murthy (2019) uses an approximation argument that represents the Polish space $\mathcal{X}$ as an increasing sequence of compact subsets. This enables duality for any Borel distribution $\widehat{\mathbb{P}}$, based on the Fenchel conjugate on vector spaces (Luenberger 1997), under semicontinuity assumptions on the transport cost function c and loss function f. Using the same infinite-dimensional convex duality, Sinha et al. (2018, theorem 5) streamlines the analysis by assuming the function $(\widehat{x}, x) \mapsto \lambda c(\widehat{x}, x) - f(x)$ is a normal integrand (Rockafellar and Wets 2009). Compared with these nonconstructive duality proofs, our (nonconstructive) proof uses only the Legendre transform, namely, the convex duality for univariate real-valued functions. The constructive proof developed by Gao and Kleywegt (2023) achieves a similar level of generality as Blanchet and Murthy (2019), but without relying on convex duality theory. They construct an approximately worst-case distribution using the first-order optimality condition of the weak dual problem. Although both their approach and ours avoid using advanced minimax theorems, our analysis is notably shorter.

3. Discussion on the IP

In this section, we first show that the IP in Section 2 is intricately linked to the measurable projection and measurable selection conditions and then prove (IP) holds in the Wasserstein DRO setting.

In (D), the function f belongs to a family of loss functions, and the transport cost function c is chosen contingent on the specific application. Consequently, from a pragmatic standpoint, we aim for (IP) to be applicable to functions within the family

$$\left\{\phi : \mathcal{X}\times\mathcal{X} \to \mathbb{R}\cup\{-\infty\} \;\middle|\; \begin{array}{l}\phi(\widehat{x}, x) = f(x) - \lambda c(\widehat{x}, x),\ f : \mathcal{X} \to \mathbb{R} \text{ measurable},\ \lambda \geq 0 \\ c : \mathcal{X}\times\mathcal{X} \to [0,\infty] \text{ measurable},\ c(x,x) = 0 \text{ for all } x \in \mathcal{X}\end{array}\right\}.$$

Observe that the nonnegativity and reflexivity of c imply the following *diagonally dominant* property of functions in this family (see Lemma A.3).

Definition 1. We say a $(\mathscr{F}\otimes\mathscr{F})$-measurable function $\phi:\mathcal{X}\times\mathcal{X}\to\mathbb{R}\cup\{-\infty\}$ is *diagonally dominant* if $\phi(\hat{x},x)\le\phi(x,x)$ for every $\hat{x},x\in\mathcal{X}$. We say a $(\mathscr{F}\otimes\mathscr{F})$-measurable set $A\subset\mathcal{X}\times\mathcal{X}$ is *diagonally dominant* if for every $(\hat{x},x)\in A$, it holds that $(x,x)\in A$. We say a set function $E:\mathcal{X}\to\mathscr{F}\setminus\{\emptyset\}$ with $(\mathscr{F}\otimes\mathscr{F})$-measurable graph is *diagonally dominant* if $x\in E(\hat{x})$ implies $x\in E(x)$.

With this definition, we show that (IP) is equivalent to the *measurable projection* and the weak *measurable selection* conditions. We denote by $\mathscr{F}_{\hat{\mathbb{P}}}$ the completion of $\mathscr{F}$ under $\hat{\mathbb{P}}$ (Kallenberg 1997, p. 13).

Proposition 1. (IP) *holds for all $(\mathscr{F}\otimes\mathscr{F})$-measurable diagonally dominant functions $\phi:\mathcal{X}\times\mathcal{X}\to\mathbb{R}\cup\{-\infty\}$ if and only if $(\mathcal{X},\mathscr{F},\hat{\mathbb{P}})$ satisfies the following two conditions:*

(Proj) *[Measurable Projection] For any diagonally dominant set $A\in\mathscr{F}\otimes\mathscr{F}$,*

$$\mathrm{Proj}_{\hat{x}}(A):=\{\hat{x}\in\mathcal{X}:(\hat{x},x)\in A \text{ for some } x\in\mathcal{X}\}\in\mathscr{F}_{\hat{\mathbb{P}}}.$$

(Sel*) *[Weak Measurable Selection] For any diagonally dominant set-valued function $E:\mathcal{X}\to\mathscr{F}\setminus\{\emptyset\}$ with a measurable graph*

$$\mathrm{Graph}(E):=\{(\hat{x},x)\in\mathcal{X}\times\mathcal{X}:x\in E(\hat{x})\}\in\mathscr{F}\otimes\mathscr{F},$$

there exists a probability measure $\gamma\in\Gamma_{\hat{\mathbb{P}}}$, such that $\mathrm{supp}\,\gamma\subset\mathrm{Graph}(E)$.

Remark 4. Measurable projection and measurable selection are often seen in the literature on stochastic control (Bertsekas and Shreve 1996, Rockafellar and Wets 2009) and stochastic programming (Shapiro 2017). (Proj) states that the projection operator $\mathrm{Proj}_{\hat{x}}:(\hat{x},x)\mapsto\hat{x}$ maps measurable sets in $\mathscr{F}\otimes\mathscr{F}$ to $\hat{\mathbb{P}}$-measurable sets. We refer (Sel*) to as the weak measurable selection, because it is weaker than the measurable selection in the literature (Rockafellar and Wets 2009, Shapiro et al. 2021). The measurable selection therein involves a *deterministic* selection which, in our context, reads as

(Sel) [Measurable Selection] *For any set-valued function $E:\mathcal{X}\to\mathscr{F}\setminus\{\emptyset\}$ with a measurable graph, there exists an $(\mathscr{F}_{\hat{\mathbb{P}}},\mathscr{F})$-measurable map $T:\mathcal{X}\to\mathcal{X}$ such that $T(\hat{x})\in E(\hat{x})$, $\forall\hat{x}\in\mathcal{X}$.*

Rockafellar and Wets (2009, theorem 14.60) (see also Shapiro et al. 2021, section 9.3.4) shows that measurable projection and measurable selection together imply the following:

$$\mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x\in\mathcal{X}}\phi(\hat{X},x)\right]=\sup_{T\in\mathcal{T}}\mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}[\phi(\hat{X},T(\hat{X}))],$$

where $\mathcal{T}$ denotes the set of $(\mathscr{F}_{\hat{\mathbb{P}}},\mathscr{F})$-measurable maps. Note that $(\mathrm{Id}\otimes T)_{\#}\hat{\mathbb{P}}\in\Gamma_{\hat{\mathbb{P}}}$, so the previous expression is stronger than (IP). Comparatively, the weak measurable selection condition (Sel*) allows a *random* selection, represented by the conditional distribution of $\gamma_{x|\hat{x}}$ supported on $E(\hat{x})$. This indicates that (Sel*) is weaker than (Sel). By Proposition 1, (IP) is equivalent to (Sel*) and (Proj).

The next result shows that (IP) holds when the transport cost function corresponds to the p-Wasserstein DRO, even if f is not $\mathscr{F}$-measurable but merely $\hat{\mathbb{P}}$-measurable. This will be used in Example 4.

Proposition 2. *Let $(\mathcal{X},d)$ be a metric space equipped with Borel σ-algebra $\mathscr{F}$, $\hat{\mathbb{P}}$ be a tight measure, f be $\hat{\mathbb{P}}$-measurable with $\mathbb{E}_{\hat{\mathbb{P}}}[f]>-\infty$. Let $p\in[1,\infty)$, $\lambda\ge0$. Then the function $\phi(\hat{x},x)=f(x)-\lambda d(\hat{x},x)^p$ is $(\mathscr{F}\otimes\mathscr{F}_{\hat{\mathbb{P}}})$-measurable and satisfies* (IP).

4. Examples

In this section, we offer several examples that demonstrate how our findings not only align with existing research but also are useful for important applications in the area of distributionally robust sequential decision making.

The following two examples illustrate that existing results in the literature are based on assumptions that are strictly stronger than (Proj) and (Sel*). Consequently, our results strictly generalize existing findings in the literature.

Example 1 (Empirical Distribution). If $(\mathcal{X},\mathscr{F}_{\hat{\mathbb{P}}})$ is a discrete measurable space, that is, $\mathscr{F}_{\hat{\mathbb{P}}}=2^{\mathcal{X}}$, then (Proj) and (Sel) always holds, because every subset of $\mathcal{X}$ is $\mathscr{F}_{\hat{\mathbb{P}}}$-measurable, and every map $T:\mathcal{X}\to\mathcal{X}$ is $\hat{\mathbb{P}}$-measurable. For instance, when $\mathcal{X}$ is a measurable space equipped with a Borel σ-algebra and $\hat{\mathbb{P}}$ is finitely supported, then $\mathscr{F}_{\hat{\mathbb{P}}}=2^{\mathcal{X}}$ is the collection of all subsets of $\mathcal{X}$, which is the discrete σ-algebra on $\mathcal{X}$. Thus our result covers the results in Mohajerin Esfahani and Kuhn (2018), Zhao and Guan (2018), and Zhen et al. (2023), which studies the case where $\mathcal{X}$ is a convex subset of $\mathbb{R}^d$.

Example 2 (Polish Space). If $\mathcal{X}$ is a Polish (complete and separable metric) space and $\mathscr{F}$ is its Borel σ-algebra, then (Proj) holds due to Aubin and Frankowska (2009, theorem 8.3.2), and (Sel) holds due to Aliprantis and Border (2006, theorem 18.26). Thereby our result covers the results in Blanchet and Murthy (2019) and Gao and Kleywegt (2023).

Example 2 can be generalized as follows.

Example 3 (Suslin Space). A Hausdorff topological space is Suslin (also known as analytic) if it is the continuous image of a Borel set in a Polish space. If $\mathcal{X}$ is a Suslin space and $\mathscr{F}$ is its Borel σ-algebra, then (Proj) holds due to Castaing and Valadier (1977, theorem III.23), and (Sel) holds due to Castaing and Valadier (1977, theorem III.22).

The following examples showcase the broad applicability of our results.

Example 4 (Distributionally Robust Markov Decision Process). Consider a finite-horizon Markov decision process. The standard working horse (Shreve and Bertsekas 1979, Bertsekas and Shreve 1996) involves the Borel space (i.e., a Borel subset of a complete and separable metric space). To avoid measurability issues, existing literature often assumes a finite or countable state space. Nonetheless, for general Borel state space, we can still verify (IP) by Proposition 2. Let the state space $\mathcal{S}$ and the action space $\mathcal{A}$ be nonempty Borel spaces. Let $\{\mathcal{A}(s)\}_{s\in\mathcal{S}}$ be a family of nonempty feasible action sets $\mathcal{A}(s)\subset\mathcal{A}$ such that the corresponding constraint set $\mathrm{K}=\{(s,a): s\in\mathcal{S}, a\in\mathcal{A}(s)\}$ is an analytic subset of $\mathcal{S}\times\mathcal{A}$. Let the one-stage cost g_t be a lower semianalytic function on K bounded from below. Suppose we have a nominal transition kernel $\{\widehat{\mathbb{P}}(\cdot|s,a)\}_{(s,a)\in\mathrm{K}}$ that is a Borel measurable stochastic kernel on $\mathcal{S}$ given (s,a). Consider the following uncertainty sets defined for every state-action pair $(s,a)\in\mathrm{K}$:

$$\mathfrak{M}(s,a)=\left\{\mathbb{P}\in\mathcal{P}(\mathcal{S}): \mathcal{K}_c(\widehat{\mathbb{P}}(\cdot|s,a),\mathbb{P})\le\rho(s,a)\right\},$$

where the positive radius function ρ is lower semianalytic on K, and the transport cost c associated with $\mathcal{K}_c$ is d^p, $p\in[1,\infty)$ where d is the metric on $\mathcal{S}$. The distributionally robust counterpart of the value iteration (Yang 2017, Wang et al. 2022) is given by $V_{T+1}\equiv 0$, and for $t=1,\dots,T$,

$$V_t(s)=\inf_{a\in\mathcal{A}(s)}\left\{g_t(s,a)+\sup_{\mathbb{P}\in\mathfrak{M}(s,a)}\mathbb{E}_{s'\sim\mathbb{P}}[V_{t+1}(s')]\right\}.$$

Then by induction, we can prove the duality

$$V_t(s)=\inf_{\substack{a\in\mathcal{A}(s)\\ \lambda\ge 0}}\left\{g_t(s,a)+\lambda\rho(s,a)+\mathbb{E}_{\hat{s}'\sim\widehat{\mathbb{P}}(\cdot|s,a)}\left[\sup_{s'\in\mathcal{S}}\{V_{t+1}(s')-\lambda c(\hat{s}',s')\}\right]\right\}. \quad (2)$$

In the inductive step, V_{t+1} is $\widehat{\mathbb{P}}$-measurable and lower semianalytic. The strong duality holds thanks to Proposition 2. Because $\{\widehat{\mathbb{P}}(\cdot|s,a)\}_{(s,a)\in\mathrm{K}}$ is a Borel measurable kernel, the robust loss is also lower semianalytic. Taking the infimum yields a lower semianalytic value function V_t, which completes the induction. We refer to Online Appendix EC.3 for details.

Example 5 (Data-Driven Robust Multistage Stochastic Programming). Consider a multistage stochastic programming problem (Shapiro et al. 2021). Let $(x_1,\dots,x_T)$ be a T-stage random data process, which is assumed to be stagewise independent, namely, $\{x_t\}_{t=1}^T$ are mutually independent. At each stage, after the current-stage uncertainty x_t is realized, the decision maker seeks a nonanticipative decision u_t from a feasible set $\mathcal{U}_t(u_{t-1},x_t)\subset\mathbb{R}^{d_t}$ dependent on the previous-stage decision u_{t-1} and the current-stage uncertainty x_t, where $\mathcal{U}_t$ is a measurable multifunction (i.e., set-valued function). The (random) cost of taking a decision u_t at stage t is $f_t(u_t,x_t)$, and the goal is to minimize the cumulative cost. Suppose the uncertainty x_t has an unknown distribution on a Suslin space $\mathcal{X}_t$ equipped with the Borel σ-algebra, and one formulates an uncertainty set based on a nominal Borel distribution $\widehat{\mathbb{P}}_t$ with a radius $\rho_t>0$:

$$\mathfrak{M}_t=\{\mathbb{P}\in\mathcal{P}(\mathcal{X}_t): \mathcal{K}_c(\widehat{\mathbb{P}}_t,\mathbb{P})\le\rho_t\},\qquad t=1,\dots,T.$$

Following the convention in stochastic programming, we assume there is no uncertainty in the first stage, that is, $\rho_1=0$ and $\widehat{\mathbb{P}}_1$ is a Dirac measure $\widehat{\mathbb{P}}_1=\boldsymbol{\delta}_{x_1}$, where $x_1\in\mathcal{X}_1$, and we set $u_0=\varnothing$. It would be natural to consider the following distributionally robust counterpart of Bellman recursion: $Q_{T+1}\equiv 0$, and for $t=1,\dots,T$,

$$Q_t(u_{t-1},x_t)=\inf_{u\in\mathcal{U}_t(u_{t-1},x_t)}\left\{f_t(u,x_t)+\sup_{\mathbb{P}\in\mathfrak{M}_{t+1}}\mathbb{E}_{\mathbb{P}}[Q_{t+1}(u,x_{t+1})]\right\}.$$

Under this setting, (Proj) and (Sel) follow from Example 3. Assume that $f_t:\mathbb{R}^{d_t}\times\mathcal{X}_t\to\mathbb{R}$ is bounded from below and random lower semicontinuous, that is, the epigraph multifunction $x_t\longmapsto \mathrm{epi} f_t(\cdot,x_t)$ is measurable and closed valued (Shapiro et al. 2021, definition 9.47). Assume the multifunction $\mathcal{U}_t:\mathbb{R}^{d_{t-1}}\times\mathcal{X}_t\rightrightarrows\mathbb{R}^{d_t}$ is measurable. Moreover, for each $x_t\in\mathcal{X}_t$, assume $\mathcal{U}_t(\cdot,x_t)$ is nonempty and has closed graph. Assume further that there exists a bounded set containing $\mathcal{U}_t(u_{t-1},x_t)$ for all u_{t-1} and x_t. Then by induction (see Online Appendix EC.3 for details), we have the duality

$$Q_t(u_{t-1},x_t)=\inf_{\substack{u\in\mathcal{U}_t(u_{t-1},x_t)\\ \lambda\ge 0}}\left\{f_t(u,x_t)+\lambda\rho_t+\mathbb{E}_{\widehat{\mathbb{P}}_{t+1}}\left[\sup_{x\in\mathcal{X}_{t+1}}\{Q_{t+1}(u,x)-\lambda c(\hat{x}_{t+1},x)\}\right]\right\}.$$

Example 6 (Chance Constraint). Let $(\mathcal{X},d)$ be a metric space and let $\mathscr{F}$ be its Borel σ-algebra. Consider a distributionally robust chance constraint (Xie 2021, Yang and Gao 2022, Chen et al. 2023):

$$\inf_{\mathbb{P}\in\mathcal{P}(\mathcal{X})}\{\mathbb{P}(\mathcal{S}): \mathcal{W}_p(\widehat{\mathbb{P}},\mathbb{P})\le\rho\}\ge 1-\beta, \quad (3)$$

which ensures that an event $\mathcal{S}\in\mathscr{F}$ happens with probability higher than $1-\beta$, $\beta\in(0,1)$, with respect to every distribution within the uncertainty set. Denote by $\mathcal{S}^c$ the complement of the set $\mathcal{S}$. In Online Appendix EC.3, we show that if $\phi(\hat{x},x)=\mathbf{1}_{\mathcal{S}^c}(x)-\lambda d(\hat{x},x)^p$ satisfies (IP) for every $\lambda>0$, then using Theorem 1, the distributionally robust chance constraint (3) is equivalent to a

worst-case conditional value-at-risk constraint

$$\rho^p \le -\beta \mathbb{CV}@\mathrm{R}_\beta^{\widehat{\mathbb{P}}}(-d(\widehat{X},\mathcal{S}^c)^p).$$

Here, $\mathbb{CV}@\mathrm{R}_\beta^{\widehat{\mathbb{P}}}(\cdot)$ represents the conditional value-at-risk at risk level β. The constraint is infeasible if $\rho \ge \mathbb{E}_{\hat{X}\sim\widehat{\mathbb{P}}}[d(\widehat{X},\mathcal{S}^c)^p]^{1/p}$. Similar results have been obtained by Chen et al. (2023), Yang and Gao (2022), and Xie (2021) but with less generality. In Chen et al. (2023), the ambient space $\mathcal{X}$ is Euclidean, and $\widehat{\mathbb{P}}$ is empirical. In Yang and Gao (2022), the ambient space is a subset of a Euclidean space, and $\widehat{\mathbb{P}}$ is a Borel probability measure. In Xie (2021), $\mathcal{X}$ is a totally bounded Polish space. Both Chen et al. (2023) and Xie (2021) consider only 1-Wasserstein distance. In comparison, the previous result requires $(\mathcal{X},d)$ only to be a metric space and (IP) for φ and has no further requirement on the nominal probability distribution $\widehat{\mathbb{P}}$ nor the set $\mathcal{S}$ beyond measurability.

5. Extensions

In this section, we extend our results to several other problems. The proofs are based on similar techniques that we developed in Section 2.

5.1. Infinity-Wasserstein DRO and Maximum Transport Cost

Recall the ∞-Wasserstein distance

$$\mathcal{W}_\infty(\widehat{\mathbb{P}},\mathbb{P}) = \inf_{\gamma\in\Gamma(\widehat{\mathbb{P}},\mathbb{P})} \|d\|_{L^\infty(\mathcal{X}\times\mathcal{X};\gamma)} = \inf_{\gamma\in\Gamma(\widehat{\mathbb{P}},\mathbb{P})} \gamma\text{-ess sup}\, d(\hat{x},x),$$

where the result in Section 2 only covers the Wasserstein distance of a finite order. To study the case $p=\infty$, we introduce the *maximum transport cost* as

$$\bar{\mathcal{K}}_c(\widehat{\mathbb{P}},\mathbb{P}) := \inf_{\gamma\in\Gamma(\widehat{\mathbb{P}},\mathbb{P})} \gamma\text{-ess sup}\, c(\hat{x},x),$$

where c is a transport cost function satisfying Assumption 1. We define the maximum transport cost robust loss by

$$\bar{\mathcal{L}}(\rho) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] : \bar{\mathcal{K}}_c(\widehat{\mathbb{P}},\mathbb{P}) \le \rho\}.$$

Similar to the results in Section 2, the soft-constrained counterpart can be viewed as the Legendre transform of negative hard-constrained robust loss

$$(-\bar{\mathcal{L}})^*(-\lambda) = \sup_{\rho\ge 0}\{(-\lambda)\rho - (-\bar{\mathcal{L}}(\rho))\} = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\bar{\mathcal{K}}_c(\widehat{\mathbb{P}},\mathbb{P})\}.$$

With this definition, we cover the ∞-Wasserstein DRO by setting $c=d$. Here, we first establish a duality result when the constraint of the uncertainty set is a strict inequality.

Proposition 3. *Let f and c satisfy Assumption 1. Define*

$$\bar{\mathcal{L}}^\circ(\rho) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] : \bar{\mathcal{K}}_c(\widehat{\mathbb{P}},\mathbb{P}) < \rho\}.$$

Suppose the function $\psi_\rho(\hat{x},x) := f(x) - \infty\mathbf{1}\{c(\hat{x},x)\ge\rho\}$ satisfies (IP) for some $\rho>0$. Then

$$\bar{\mathcal{L}}^\circ(\rho) = \mathbb{E}_{\hat{X}\sim\widehat{\mathbb{P}}}\left[\sup_x\{f(x) : c(\widehat{X},x) < \rho\}\right].$$

Suppose the ψ_ρ satisfies (IP) for every $\rho>0$. Then

$$(-\bar{\mathcal{L}})^*(-\lambda) = \sup_{\rho>0}\left\{\mathbb{E}_{\hat{X}\sim\widehat{\mathbb{P}}}\left[\sup_x\{f(x) : c(\widehat{X},x) < \rho\} - \lambda\rho\right]\right\}.$$

Remark 5. Unlike the results in Section 2, we have to distinguish two cases: $\bar{\mathcal{L}}^\circ(\rho)$ that involves the strict inequality constraint and $\bar{\mathcal{L}}(\rho)$ that involves the nonstrict inequality constraint. Indeed, for Problem (P) in Section 2, the value of $\mathcal{L}(\rho)$ in (P) would not be affected if we replace the nonstrict inequality by the strict inequality thanks to the concavity, and thus continuity, of $\mathcal{L}$ with respect to $\rho\in(0,\infty)$. However, for the problem with the maximum transport cost, neither $\bar{\mathcal{L}}^\circ$ nor $\bar{\mathcal{L}}$ is necessarily continuous.

Without additional assumptions, the duality does not hold for the equality constraint, even when the cost function c is also a metric. For instance, consider $\mathcal{X}=[0,1]\times[0,1]$, $\widehat{\mathbb{P}}$ is a uniform distribution on $\{0\}\times[0,1]$, and $f(x_1,x_2)=\mathbf{1}\{x_1=1\}$. We introduce the following cost function:

$$c((x_1,x_2),(y_1,y_2)) = \begin{cases} 0, & x_1=y_1, x_2=y_2,\\ 101+|x_1-y_1|, & x_1\ne y_1, x_2=y_2,\\ 100+|x_1-y_1|+|x_2-y_2|, & x_2\ne y_2. \end{cases}$$

Then $c(x,y)=0$ iff $x=y$, and c is symmetric. Triangular inequality holds because the distance between any two distinct points is between 100 and 102. If $\mathbb{P}$ is a uniform distribution on $\{1\}\times[0,1]$, then $\bar{\mathcal{K}}_c(\widehat{\mathbb{P}},\mathbb{P})=101$; thus,

$$\bar{\mathcal{L}}(101) = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] : \bar{\mathcal{K}}_c(\widehat{\mathbb{P}},\mathbb{P}) \le 101\} = 1.$$

However, because $d((0,t),(1,s))>101$ for any $t,s\in[0,1]$, we would have

$$\mathbb{E}_{\hat{X}\sim\widehat{\mathbb{P}}}\left[\sup_x\{f(x) : c(\widehat{X},x)\le 101\}\right] = 0 \ne 1.$$

The following result shows that, with additional assumptions on the space and the transport cost function, we can obtain the duality result. The detailed proofs for Proposition 3 and Theorem 2 can be found in Online Appendix EC.4.

Theorem 2. *Suppose $\mathcal{X}$ is a Polish space and $c : \mathcal{X} \times \mathcal{X} \to [0,\infty)$ is continuous. Let f satisfy Assumption 1. Then*

$$\bar{\mathcal{L}}(\rho) = \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_x\{f(x) : c(\hat{X},x) \le \rho\}\right],$$

$$(-\bar{\mathcal{L}})^*(-\lambda) = \sup_{\rho\ge 0}\left\{\mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_x\left\{f(x) : c(\hat{X},x)\le\rho\right\} - \lambda\rho\right]\right\}.$$

Example 7 (Chance Constraint (Continued)). Consider the chance constraint introduced in Example 6 but with ∞-Wasserstein distance. If $\mathcal{X}$ is complete and separable, then the robust chance constraint becomes $\hat{\mathbb{P}}(d(\hat{X},\mathcal{S}^c) \le \rho) \le \beta$. In particular, it is infeasible if $\rho \ge d(\mathrm{supp}\ \hat{\mathbb{P}}, \mathcal{S}^c)$. Compared with Yang and Gao (2022), which assumes $\mathcal{X}$ is a normed space, we only require $(\mathcal{X}, d)$ to be Polish.

5.2. Risk-Averse Optimization

Recall $(\mathcal{X}, \mathscr{F}, \hat{\mathbb{P}})$ is a probability space. Consider a concave risk measure $J : \mathcal{P}(\mathcal{X}) \to \bar{\mathbb{R}}$ of the following form:

$$J(\mathbb{P}) := \inf_{\alpha\in A} \mathbb{E}_{X\sim\mathbb{P}}[f_\alpha(X)], \tag{4}$$

where $f_\alpha : \mathcal{X} \to \mathbb{R}$ are a family of measurable functions indexed by $\alpha \in A$, a subset of a linear topological space.

Given a nominal distribution $\hat{\mathbb{P}}$, define

$$\mathcal{L}_J(\rho) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})}\{J(\mathbb{P}) : \mathcal{K}_c(\hat{\mathbb{P}},\mathbb{P}) \le \rho\},$$

$$\bar{\mathcal{L}}_J(\rho) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{X})}\{J(\mathbb{P}) : \bar{\mathcal{K}}_c(\hat{\mathbb{P}},\mathbb{P}) \le \rho\}.$$

We assume that there exists a compact set $A' \subset A$ such that for all distributions $\mathbb{P}$ in the distributional uncertainty set, it holds that

$$\inf_{\alpha\in A}\mathbb{E}_{\mathbb{P}}[f_\alpha] = \min_{\alpha\in A'}\mathbb{E}_{\mathbb{P}}[f_\alpha]. \tag{5}$$

This enables the exchange of sup over $\mathbb{P}$ and inf over α using Sion's minimax theorem. We have the following result.

Theorem 3. *Let f_α and c satisfy Assumption 1 and let $\alpha \longmapsto f_\alpha(x)$ be lower semicontinuous and convex for each x. Assume for every compact subset $A' \subset A$, $\inf_{\alpha\in A', x\in\mathcal{X}} f_\alpha(x) > -\infty$.*

I. *Suppose $\phi_{\lambda,\alpha}(\hat{x},x) = f_\alpha(x) - \lambda c(\hat{x},x)$ satisfies (IP) for every $\lambda > 0$ and $\alpha \in A$. Assume there exists a compact subset $A' \subset A$ such that (5) holds for every $\mathbb{P}$ satisfying $\mathcal{K}_c(\hat{\mathbb{P}},\mathbb{P}) \le \rho$. Then*

$$\mathcal{L}_J(\rho) = \inf_{\lambda\ge 0,\alpha\in A}\left\{\lambda\rho + \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x\in\mathcal{X}}\{f_\alpha(x) - \lambda c(\hat{X},x)\}\right]\right\}. \tag{6}$$

II. *Suppose $\mathcal{X}$ is a Polish space, and $c : \mathcal{X}\times\mathcal{X} \to [0,\infty)$ is continuous. Assume there exists a compact subset $A' \subset A$ such that (5) holds for every $\mathbb{P}$ satisfying $\bar{\mathcal{K}}_c(\hat{\mathbb{P}},\mathbb{P}) \le \rho$. Then*

$$\bar{\mathcal{L}}_J(\rho) = \inf_{\alpha\in A}\mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_x\{f_\alpha(x) : c(\hat{X},x) \le \rho\}\right]. \tag{7}$$

Here we list several examples where $X \in \mathbb{R}$ represents a random loss and $A = \mathbb{R}$.

- Conditional value-at-risk $\mathbb{CV@R}^{\mathbb{P}}_\beta(X)$ at risk level $\beta \in (0,1)$: $f_\alpha(x) = \alpha + (x-\alpha)_+/\beta$ (Rockafellar and Uryasev 2000, Föllmer and Schied 2010).
- Variance $\mathbb{Var}^{\mathbb{P}}(X)$: $f_\alpha(x) = (x-\alpha)^2$.
- Mean absolute deviation (around median) $\mathbb{MAD}^{\mathbb{P}}(X)$: $f_\alpha(x) = |x-\alpha|$.
- Entropic risk measure $\mathbb{Ent}^{\mathbb{P}}_\theta(X) = \log\mathbb{E}[e^{\theta X}]/\theta$ with risk-aversion parameter $\theta > 0$: $f_\alpha(x) = \alpha + (e^{\theta(x-\alpha)} - 1)/\theta$.

In Online Appendix EC.5, we can verify (5), and other assumptions of Theorem 3 hold for all these risk measures.

Example 8. Let $Z \in \mathcal{Z} = (\mathbb{R}^d, \|\cdot\|)$ be the vector of random loss of d assets with nominal distribution $\hat{\mathbb{Q}}$ and $b \in \mathbb{R}^d$ be a portfolio weight vector on these assets. The portfolio loss is thus $b^\top Z$. We have the following results on the robust risk of portfolio loss under various risk measures. Their proofs are given in Online Appendix EC.5.

- Conditional value-at-risk: for $p \in [1,\infty]$:

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{CV@R}^{\mathbb{Q}}_\beta(b^\top Z) : \mathcal{W}_p(\hat{\mathbb{Q}},\mathbb{Q}) \le \rho\}$$
$$= \mathbb{CV@R}^{\hat{\mathbb{Q}}}_\beta(b^\top\hat{Z}) + (1-\beta)^{-\frac{1}{p}}\|b\|_*\rho.$$

- Variance: for $p = 2$,

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{Var}^{\mathbb{Q}}(b^\top Z) : \mathcal{W}_2(\hat{\mathbb{Q}},\mathbb{Q}) \le \rho\}$$
$$= (\mathbb{Var}^{\hat{\mathbb{Q}}}(b^\top\hat{Z})^{\frac{1}{2}} + \|b\|_*\rho)^2.$$

For $p = \infty$,

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{Var}^{\mathbb{Q}}(b^\top Z) : \mathcal{W}_\infty(\hat{\mathbb{Q}},\mathbb{Q}) \le \rho\}$$
$$= \min_{\alpha\in\mathbb{R}}\mathbb{E}_{\hat{Z}\sim\hat{\mathbb{Q}}}[(|\hat{Z}-\alpha| + \|b\|_*\rho)^2].$$

For $1 \le p < 2$, the robust loss is positive infinity.

- Mean average deviation: for $1 \le p \le \infty$,

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{MAD}^{\mathbb{Q}}(b^\top Z) : \mathcal{W}_p(\hat{\mathbb{Q}},\mathbb{Q}) \le \rho\}$$
$$= \mathbb{MAD}^{\hat{\mathbb{Q}}}(b^\top\hat{Z}) + \|b\|_*\rho.$$

- Entropic risk measure: for $p = \infty$,

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{Ent}^{\mathbb{Q}}_\theta(b^\top Z) : \mathcal{W}_\infty(\hat{\mathbb{Q}},\mathbb{Q}) \le \rho\}$$
$$= \mathbb{Ent}^{\hat{\mathbb{Q}}}_\theta(b^\top\hat{Z}) + \|b\|_*\rho.$$

For $1 \le p < \infty$, the robust loss is positive infinity.

5.3. Globalized Distributionally Robust Counterpart

The globalized distributionally robust counterpart (Liu et al. 2023) studies the following problem:

$$\sup_{\mathbb{P},\widetilde{\mathbb{P}}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\mathcal{K}_c(\widetilde{\mathbb{P}},\mathbb{P}) : \mathcal{K}_{\tilde{c}}(\widehat{\mathbb{P}},\widetilde{\mathbb{P}}) \le \theta\}. \tag{G}$$

We also consider its hard- and soft-constrained variants:

$$\mathcal{L}_G(\rho,\theta) := \sup_{\mathbb{P},\widetilde{\mathbb{P}}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] : \mathcal{K}_c(\widetilde{\mathbb{P}},\mathbb{P}) \le \rho, \mathcal{K}_{\tilde{c}}(\widehat{\mathbb{P}},\widetilde{\mathbb{P}}) \le \theta\}, \tag{G-hard}$$

$$\sup_{\mathbb{P},\widetilde{\mathbb{P}}\in\mathcal{P}(\mathcal{X})} \{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\mathcal{K}_c(\widetilde{\mathbb{P}},\mathbb{P}) - \mu\mathcal{K}_{\tilde{c}}(\widehat{\mathbb{P}},\widetilde{\mathbb{P}})\}. \tag{G-soft}$$

The following result extends the work of Liu et al. (2023), which is based on the assumption that $\mathcal{X}$ is a subset of Euclidean space and the transport cost is defined by the 1-Wasserstein distance. The proof can be found in Online Appendix EC.6.

Proposition 4. *Let the loss function f and two cost functions $c,\tilde{c}$ satisfy Assumption 1. For $\rho,\theta>0$, $\lambda,\mu\ge 0$, if $(\tilde{x},x)\longmapsto f(x)-\lambda c(\tilde{x},x)$ satisfies (IP) for every $\lambda\ge 0$, and $(\widehat{x},x)\longmapsto \sup_{\tilde{x}\in\mathcal{X}} f(x)-\lambda c(\tilde{x},x)-\mu\tilde{c}(\widehat{x},\tilde{x})$ satisfies (IP) for every $\lambda,\mu\ge 0$, then for every $\lambda,\mu,\rho,\theta>0$, (G-hard) is equivalent to*

$$\min_{\lambda,\mu\ge 0}\left\{\lambda\rho+\mu\theta + \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x,\tilde{x}\in\mathcal{X}}\{f(x)-\lambda c(\tilde{x},x)-\mu\tilde{c}(\widehat{X},\tilde{x})\}\right]\right\},$$

(G) is equivalent to

$$(-\mathcal{L}_G(\cdot,\theta))^*(-\lambda) = \min_{\mu\ge 0}\left\{\mu\theta+\mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x,\tilde{x}\in\mathcal{X}}\{f(x)-\lambda c(\tilde{x},x)-\mu\tilde{c}(\widehat{X},\tilde{x})\}\right]\right\},$$

and (G-soft) is equivalent to

$$\mathcal{L}_G^*(-\lambda,-\mu) = \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x,\tilde{x}\in\mathcal{X}}\{f(x)-\lambda c(\tilde{x},x)-\mu\tilde{c}(\widehat{X},\tilde{x})\}\right].$$

Here $\mathcal{L}_G^*(-\lambda,\cdot)$ is the dual of the mapping $\theta\longmapsto -(-\mathcal{L}_G(\cdot,\theta))^*(-\lambda)$. Moreover, if $c=\tilde{c}=d$, then (G) is further equivalent to

$$\min_{0\le\mu\le\lambda}\left\{\mu\theta+\mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x\in\mathcal{X}}\{f(x)-\mu d(\widehat{X},x)\}\right]\right\}.$$

6. Concluding Remarks

We developed a new duality proof for Wasserstein distributionally robust optimization. The new result offers an alternative view of duality from the perspective of the interchangeability principle. This suggests that establishing duality in broader settings hinges primarily on verifying the interchangeability property, which has been more extensively explored.

Appendix A. Auxiliary Results

Lemma A.1. *Assume Assumption 1 holds. Recall for $\lambda\ge 0$,*

$$(-\mathcal{L})^*(-\lambda) = \sup_{\mathbb{P}\in\widehat{\mathcal{P}}}\{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \lambda\mathcal{K}_c(\widehat{\mathbb{P}},\mathbb{P})\},$$

$$\mathcal{G}(\lambda) = \mathbb{E}_{\hat{X}\sim\hat{\mathbb{P}}}\left[\sup_{x\in\mathcal{X}}\{f(x)-\lambda c(\widehat{X},x)\}\right].$$

Then $(-\mathcal{L})^*(-\cdot)$ and $\mathcal{G}(\cdot)$ are lower bounded by $\mathbb{E}_{\hat{\mathbb{P}}}[f]$, monotonically decreasing, convex, and lower semi-continuous on $[0,\infty)$.

Lemma A.2. *If $f:\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ is a monotonically decreasing convex function with $f(\rho)=+\infty$ for $\rho<0$ and $f\not\equiv+\infty$, then $f^*(-\lambda)=\sup_{\rho\in\mathbb{R}}\{-\lambda\rho-f(\rho)\}$ is a lower semicontinuous, monotonically decreasing convex function of λ with $f^*(-\lambda)=+\infty$ for $\lambda<0$ and $f^*\not\equiv+\infty$.*

Lemma A.3. *The function $\phi:\mathcal{X}\times\mathcal{X}\to\mathbb{R}\cup\{-\infty\}$ is diagonally dominant if and only if there exists a measurable function $f:\mathcal{X}\to\mathbb{R}\cup\{-\infty\}$, a constant $\lambda\ge 0$, and a measurable function $c:\mathcal{X}\times\mathcal{X}\to[0,\infty]$ vanishing on diagonal such that $\phi(\widehat{x},x)=f(x)-\lambda c(\widehat{x},x)$.*

References

Aliprantis CD, Border K (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide* (Springer Science & Business Media, Boston).

Ambrosio L, Fusco N, Pallara D (2000) *Functions of Bounded Variation and Free Discontinuity Problems* (Oxford University Press, Oxford, UK).

Aubin JP, Frankowska H (2009) *Set-Valued Analysis* (Springer Science & Business Media, Boston).

Bertsekas DP, Shreve SE (1996) *Stochastic Optimal Control: The Discrete-Time Case*, vol. 5 (Athena Scientific, Belmont, MA).

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

Blanchet J, Murthy K, Nguyen VA (2021) Statistical analysis of Wasserstein distributionally robust estimators. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications* (INFORMS), 227–254.

Castaing C, Valadier M (1977) Measurable multifunctions. *Convex Analysis and Measurable Multifunctions* (Springer, Berlin), 59–90.

Chen Z, Kuhn D, Wiesemann W (2023) On approximations of data-driven chance constrained programs over Wasserstein balls. *Oper. Res. Lett.* 51(3):226–233.

Föllmer H, Schied A (2010) Convex and coherent riskmeasures. *Encyclopedia of Quantitative Finance*, 355–363.

Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* 48(2):603–655.

Kallenberg O (1997) *Foundations of Modern Probability*, vol. 2 (Springer, Berlin).

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research &*

Management Science in the Age of Analytics (INFORMS, Catonsville, MD), 130–166.

Liu F, Chen Z, Wang S (2023) Globalized distributionally robust counterpart. *INFORMS J. Comput.* 35(5):1120–1142.

Luenberger DG (1997) *Optimization by Vector Space Methods* (John Wiley & Sons, New York).

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.

Rockafellar RT (1970) *Convex Analysis*, Number 28 in Princeton Mathematical Series (Princeton University Press, Princeton, NJ).

Rockafellar RT, Uryasev S (2000) Optimization of conditional value at-risk. *J. Risk* 2:21–42.

Rockafellar RT, Wets RJB (2009) *Variational Analysis*, vol. 317 (Springer Science & Business Media, Boston).

Shapiro A (2001) On duality theory of conic linear problems. *Semi-Infinite Programming* (Springer, Berlin), 135–165.

Shapiro A (2017) Interchangeability principle and dynamic equations in risk averse stochastic programming. *Oper. Res. Lett.* 45(4):377–381.

Shapiro A, Dentcheva D, Ruszczynski A (2021) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM, Philadelphia).

Shreve SE, Bertsekas DP (1979) Universally measurable policies in dynamic programming. *Math. Oper. Res.* 4(1):15–30.

Sinha A, Namkoong H, Duchi J (2018) Certifying some distributional robustness with principled adversarial training. *Proc. Internat. Conf. Learn. Representations* (ICLR, Appleton, WI).

Wang J, Gao R, Zha H (2022) Reliable off-policy evaluation for reinforcement learning. *Oper. Res.* 72(2):699–716.

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

Yang I (2017) A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Lett.* 1(1):164–169.

Yang Z, Gao R (2022) *Wasserstein Regularization for 0-1 Loss* (Optimization Online).

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46(2):262–267.

Zhen J, Kuhn D, Wiesemann W (2023) A unified theory of robust and distributionally robust optimization via the primal-worst-equals-dual-best principle. *Oper. Res.*, ePub ahead of print September 26, https://doi.org/10.1287/opre.2021.0268.

Luhao Zhang is joining Johns Hopkins University as an assistant professor in the Department of Applied Mathematics and Statistics. She is a postdoctoral research scientist in the Department of Industrial Engineering and Operations Research at Columbia University.

Jincheng Yang is a Dickson instructor in the Department of Mathematics at the University of Chicago.

Rui Gao is an assistant professor in the Department of Information, Risk, and Operations Management at the McCombs School of Business at the University of Texas at Austin.