# Decision-making with Side Information: A Causal Transport Robust Approach

Jincheng Yang
Department of Applied Mathematics and Statistics, Johns Hopkins University, jincheng@jhu.edu

Luhao Zhang
Department of Applied Mathematics and Statistics, Johns Hopkins University, luhao.zhang@jhu.edu

Ningyuan Chen
Department of Management, University of Toronto Mississauga,
Rotman School of Management, University of Toronto, ningyuan.chen@utoronto.ca

Rui Gao
Department of Information, Risk and Operations Management, The University of Texas at Austin,
rui.gao@mccombs.utexas.edu

Ming Hu
Rotman School of Management, University of Toronto, ming.hu@utoronto.ca

We consider stochastic optimization with side information where, prior to decision-making, covariate data are available to inform better decisions. To hedge against data uncertainty while capturing the information structure revealed from the conditional distribution of random problem parameters given the covariate values, we propose a distributionally robust formulation based on causal transport distance. We derive a dual reformulation for evaluating the worst-case expected cost and show that the worst-case distribution in a causal transport distance ball preserves the conditional information structure from the nominal distribution. When optimizing over affine decision rules, we identify cases in which the overall problem can be solved via convex programming. When optimizing over all (non-parametric) decision rules, we identify a new class of robust optimal decision rules when the cost function is convex with respect to a one-dimensional decision variable.

*Key words*: Distributionally robust optimization; optimal transport; end-to-end learning; adjustable robust optimization

## 1. Introduction

Stochastic optimization with side information, also known as contextual stochastic optimization or conditional stochastic optimization, addresses the following problem:

$$\min_{w\in\mathcal{D}} \mathbb{E}[\Psi(w,Z)\,|\,X=x], \tag{1}$$

where the goal is to select a decision $w$ from a feasible set $\mathcal{D}$ that minimizes the conditional expectation of the cost $\Psi(w,Z)$, dependent on both the decision $w$ and a random variable $Z$, given some side information, represented by a covariate $X$. The increasing use

of side information from covariate data has significantly enhanced decision-making across e-commerce and online platforms, enabling more personalized, informed strategies. The performance evaluation often encompasses the entire covariate population; e.g., the manager of an e-commerce company cares about overall performance across all customer types. As such, we are interested in finding a decision rule that minimizes the expected cost over the joint distribution of the covariate $X$ and the random variable $Z$:

$$\min_{f\in\mathcal{F}} \mathbb{E}[\Psi(f(X),Z)]. \tag{2}$$

The decision rule $f$ offers an end-to-end map from the covariate space $\mathcal{X}$ to the decision space $\mathcal{D}$, chosen from a family $\mathcal{F}$ of functions—parametric or non-parametric—on $\mathcal{X}$. Such an end-to-end approach is often computationally efficient at deployment time, since the decision for a new context $x$ can be obtained by directly evaluating $f(x)$. Moreover, the complexity of the policy class $\mathcal{F}$ can be adjusted flexibly—ranging from simple parametric forms such as affine decision rules to rich nonparametric families—allowing practitioners to strike a desired balance between interpretability, flexibility, and statistical efficiency.

The formulation (2) covers many contextual optimization problems in operations research and machine learning. For instance, suppose $\Psi(w,z)=h(w-z)_+ + b(z-w)_+$, where $w$ is the order quantity decision, $z$ represents the demand of a product, $(\cdot)_+ = \max(\cdot,0)$, and $h,b\geq 0$ represent the overage cost and the underage cost, respectively, then (2) is known as the big-data newsvendor model (Ban and Rudin, 2019). If $\mathcal{F}$ is the set of all measurable functions on $\mathcal{X}$, then the optimal order quantity equals the conditional critical fractile $f^\star(x)=F_x^{-1}(\frac{b}{h+b})$, where $F_x$ is the conditional cumulative distribution function of demand $Z$ given $X=x$; and if $\mathcal{F}$ is the set of affine functions on $\mathcal{X}$, then (2) finds the optimal affine decision rule for the big-data newsvendor. As another example, when $\Psi(w,z)=(w-z)^2$ and $\mathcal{F}$ is the set of all measurable functions on $\mathcal{X}$, the optimal solution to (2) is $f^\star(x)=\mathbb{E}[Z\mid X=x]$, and thus the formulation (2) finds the conditional mean of $Z$ given $X$. More examples will be given in Section 2.2. We note that this is not the only formulation of contextual decision-making, and we discuss related work in Section 1.3.

Similar to classical stochastic optimization, the underlying joint distribution $\mathbb{P}_{\mathsf{true}}$ of $(X,Z)$ is often not known exactly; instead, historical data from the distribution are available. In this case, a natural approach is to replace the unknown underlying distribution of demands and features with their empirical counterpart. However, if the function class $\mathcal{F}$

contains all measurable functions, then the resulting empirical risk minimization problem yields a degenerate decision rule that is defined only on the set of historical observations of the contextual values. To overcome the degeneracy issue, it is reasonable to consider a data-driven, distributionally robust contextual decision-making framework

$$\min_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathfrak{M}} \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)], \tag{3}$$

a minimax formulation that hedges the data uncertainty, where $\mathfrak{M}$ is a chosen uncertainty set. At the core of the distributionally robust formulation is the choice of the uncertainty set, and the presence of the side information adds new challenges beyond those for classic stochastic optimization. Below, in Section 1.1, we review some existing choices of uncertainty sets and discuss their potential issues.

### 1.1. Discussion on Some Existing Uncertainty Sets

To begin with, we would like to focus on distance-based uncertainty sets, as the other popular choice—moment-based uncertainty sets—generally lacks statistical consistency.

Two classes of distance-based uncertainty sets have been widely studied in the literature. The first class is the divergence family, deeply rooted in statistics, information theory, and physics. Consider the following example.

Example 1 (KL robust solution is degenerate). Suppose $\mathfrak{M}$ is a Kullback–Leibler (KL) divergence ball, centered at the empirical distribution $\widehat{\mathbb{P}}$ constructed from $K$ independently and identically distributed (i.i.d.) samples from a continuous underlying distribution. Then with probability one, $\widehat{\mathbb{P}}$ can be represented as $\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\delta}_{(\widehat{x}_k,\widehat{z}_k)}$, where $K$ is the sample size and all $(\widehat{x}_k,\widehat{z}_k)$'s are different from each other. Let $\mathcal{F}$ be the set of all measurable functions on $\mathcal{X}$. Then, we claim that the KL robust optimal solution would satisfy

$$f_{\mathsf{kl}}(x)=\begin{cases}\arg\min_{w\in\mathcal{D}}\ \Psi(w,\widehat{z}_k), & \text{if } x=\widehat{x}_k,\ k=1,\ldots,K,\\ \text{arbitrary value}, & \text{otherwise.}\end{cases}$$

Indeed, every distribution in the KL ball is supported only on the data points from $\widehat{\mathbb{P}}$, but may differ from it in the probability weights. On an in-sample data point $\widehat{x}_k$, regardless of its weight, the optimal decision would always be the minimizer of $\Psi(\cdot,\widehat{z}_k)$ due to the interchangeability principle (Shapiro et al., 2014, Section 9.3.4). Furthermore, since the KL robust cost depends only on the function values on the in-sample data, the robust optimal solution can take any value on out-of-sample data without changing the objective value. ♣

Example 1 shows that the KL robust optimal decision rule is degenerate with probability one when the underlying distribution is continuous, regardless of the size of the uncertainty set, the sample size, or the objective function. A similar phenomenon also holds for all other divergence measures due to the structure of the worst-case distribution (Bayraksan and Love, 2015).

The second class is Wasserstein, or transport cost distance, family. It is well-known that the resulting uncertainty set avoids some degeneracy issues of the divergence sets in stochastic optimization (Kuhn et al., 2019; Gao and Kleywegt, 2023). Nonetheless, it faces new challenges when additional side information is presented. Let us consider the following toy example.

EXAMPLE 2 (WASSERSTEIN SET CANNOT CAPTURE CONDITIONAL INFORMATION). In Figure 1, $\widehat{\mathbb{P}}$ and $\mathbb{P}$ are two uniform distributions supported respectively on the blue and green line segments with a common endpoint with $x$-entry $\widehat{x}$. The angle between the two line segments is $\varepsilon$ radian. Notably, the conditional distribution $\mathbb{P}_{Z|X=x}$ is a Dirac measure

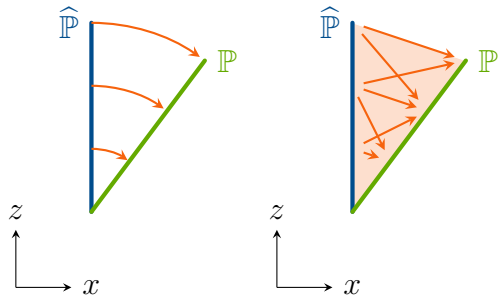


**Figure 1** **$\widehat{\mathbb{P}}$ and $\mathbb{P}$ have completely different conditional information structures but with $O(\varepsilon)$ Wasserstein distance. If we restrict transport plans to causal ones, then two distributions are distinguished by an $O(1)$ distance.**

for $x > \widehat{x}$, which is apparently very different from the conditional distribution $\widehat{\mathbb{P}}_{Z|X=\widehat{x}}$ that is uniform on the blue line segment. As will be calculated in Section 2, the Wasserstein distance between $\widehat{\mathbb{P}}$ and $\mathbb{P}$ is $O(\varepsilon)$, and the optimal transport map is a rotation. This means a Wasserstein ball centered at $\widehat{\mathbb{P}}$ would always contain a distribution that has a different conditional information structure than that of $\widehat{\mathbb{P}}$, regardless of the value of $\varepsilon$. On the other hand, as will be revisited in Section 2.1, by restricting to the causal transport map (shown in the right plot) which, in this case, is the independent joint distribution

$\widehat{\mathbb{P}} \otimes \mathbb{P}$, distributions with a different conditional information structure will be ruled out from the uncertainty set. ♣

In practice, the following situation is often observed in the data: the conditional distribution can be estimated accurately for a number of covariate values, but is largely unobserved for other values. For example, historical data may provide a good estimate of the conditional demand distribution for the product sold at deployed vending machines, but demand at new locations remains unexplored. Nonetheless, it is conceivable that the conditional demand distribution should resemble that of similar locations. In such cases, it would be reasonable to expect that the conditional distributions $\mathbb{P}_{Z|X=x}$ and $\mathbb{P}_{Z|X=\widehat{x}}$ corresponding to two similar values $x$ and $\widehat{x}$ should be close in a certain way. Therefore, we would like to choose an uncertainty set containing distributions with a similar conditional information structure to the nominal distribution. Example 2 demonstrates that the Wasserstein uncertainty set fails to preserve the conditional information structure and, in fact, the same phenomenon also holds for the worst-case distribution, as will be shown in Section 3.2. This raises concerns about the conservatism of the Wasserstein formulation.

### 1.2. Our Contributions

To capture the conditional information, in this paper, we consider a distributional uncertainty set based on *causal transport distance*, a notion that is related to Wasserstein distance but imposes an additional assumption on the transport plan; see Section 2.1 for its definition and a more in-depth discussion. The causal transport distance uncertainty set brings new computational challenges to the inner optimization over probability distributions in (3), which require new analysis of tractable reformulations and interpretations. Moreover, when the outer minimization over the class of decision rules is performed over a nonparametric class, additional computational challenges arise due to the associated infinite-dimensional functional optimization. Our main contributions are as follows.

(I) We develop a strong duality reformulation for computing the worst-case loss of a fixed decision rule (Section 3.1). Our proof is based on a new analysis of the worst-case distribution, which demonstrates how our choice of distributional uncertainty set captures the conditional information structure of the random variable given the side information (Section 3.2).

(II) We study tractable reformulations for finding the optimal decision rule when optimizing over (i) the affine class, (ii) reproducing kernel Hilbert space, and (iii) all

(non-parametric) decision rules. We provide convex reformulations when $\Psi(w, z)$ is convex in the decision variable $w$, in paticular for (i) when $\Psi$ is linear in $w$ or bilinear/quadratic in $w$ and $z$ (Section 4.1), for (ii) when $\Psi$ is Lipschitz in $w$ (Section 4.2), and for (iii) when the decision $w$ is one-dimensional. This provides a new class of decision rule with no sub-optimality gap for adjustable robust optimization (Section 4.3). We illustrate our results with conditional mean estimation, feature-based newsvendor, personalized pricing, and contextual linear optimization.

(III) We conduct numerical experiments to demonstrate that the causal transport distance uncertainty set effectively utilizes conditional information, as compared to the Wasserstein uncertainty set, and compare the performance of different classes of decision rules (Section 5).

### 1.3. Related Literature

*On stochastic optimization with side information.* In the literature, the frameworks for contextual optimization (with an offline data set) can be broadly classified into three categories (Qi and Shen, 2022; Sadana et al., 2024): *sequential learning and optimization*, *integrated prediction and optimization*, and *decision rule optimization*.

(I) Sequential learning and optimization (or two-step optimization) involves first estimating the conditional distribution of $Z$ given a new context $X = x$ and then optimizing for the conditional expectation $\min_{w\in\mathcal{D}} \mathbb{E}[\Psi(w, Z) \mid X = x]$ (e.g., Toktay and Wein (2001); Zhu and Thonemann (2004)). This approach has theoretical guarantees discussed in El Balghiti et al. (2019); Hu et al. (2022). Recent advancements in this area include approaches using Dirichlet processes (Hannah et al., 2010), Nadaraya–Watson kernel regression (Hanasusanto and Kuhn, 2013; Ban and Rudin, 2019; Srivastava et al., 2021), local regression and classification (Bertsimas and Kallus, 2020; Bertsimas and McCord, 2019), trees and forests (Bertsimas and Kallus, 2020; Bertsimas and McCord, 2019; Ban et al., 2019), and empirical residuals (Kannan et al., 2025, 2023). Additionally, robustness in optimization and regularization have been explored in Tulabandhula and Rudin (2013); Zhu et al. (2022); Bertsimas and Van Parys (2021); Loke et al. (2022); Esteban-Pérez and Morales (2022); Van Parys and Bennouna (2022); Van Parys et al. (2021); Chenreddy et al. (2022); Nguyen et al. (2025); Perakis et al. (2023). However, as noted in Liyanage and Shanthikumar (2005); Ban and Rudin (2019), statistical estimation errors and

model misspecifications can propagate through the optimization process, resulting in suboptimal performance, while in the case of contextual linear optimization, Hu et al. (2022) shows that the plug-in approach actually achieves faster regret convergence rates than methods that directly optimize downstream decision performance.

(II) Integrated prediction and optimization avoids estimating the conditional distribution by optimizing the parameterization of a contextual predictor to minimize downstream expected costs, $\mathbb{E}[\Psi(w, Z) \mid X = x]$. Various estimation techniques are explored here, including models based on conditional distributions (Kallus and Mao, 2023; Qi et al., 2025), regret minimization (Estes, 2021), smart prediction-then-optimization (Elmachtoub and Grigas, 2022; El Balghiti et al., 2019; Elmachtoub et al., 2020; Ho-Nguyen and Kılınç-Karzan, 2022), bilevel optimization (Muñoz et al., 2022; Cao and Gao, 2021; Hu et al., 2024), inverse optimization, etc. This approach requires solving an optimization over the decisions for each individual context.

(III) Decision rule optimization is an end-to-end formulation that finds a decision rule prescribing the decision for every possible context. Due to the computational difficulty of this infinite-dimensional optimization, typically, the policies are parameterized by a finite-dimensional vector, such as coefficients in an affine function of features (Brandt et al., 2009; Ban and Rudin, 2019; Bazier-Matte and Delage, 2020; Bertsimas et al., 2022) or in a reproducing kernel Hilbert space (Bertsimas and Koduri, 2022), and weight matrices in a neural network (Oroojlooyjadid et al., 2020; Qi et al., 2023; Liu et al., 2021; Rychener et al., 2023).

Our formulation falls into the third category, but our results in Section 4 do not necessarily restrict the class of decision rules to a parametric family. In this respect, the closest work to ours is Zhang et al. (2024), which considers robust optimization over decision rules with the Wasserstein uncertainty set; see the last paragraph of the literature review for a detailed comparison. We remark that in the online setting, stochastic optimization with side information has also been considered under the umbrella of contextual bandits and reinforcement learning, and there have been some studies of decision-dependent uncertainty (Basciftci et al., 2021; Vayanos et al., 2025; Yu and Shen, 2022). These are beyond the scope of this paper.

*On transport-distance based distributionally robust optimization.* Distributionally robust optimization (DRO) has received significant attention recently as a tool for decision-making under uncertainty, and different approaches mainly differ in how the uncertainty set is constructed. We refer to Rahimian et al. (2019) for a thorough review of choices of uncertainty set. Our choice of uncertainty set is aligned with DRO with transport distance, such as Wasserstein distance (Pflug and Wozabal, 2007; Wozabal, 2012; Esfahani and Kuhn, 2018; Blanchet and Murthy, 2019; Blanchet et al., 2019; Gao and Kleywegt, 2023; Gao et al., 2024b; Gao, 2023) and nested distance (Analui and Pflug, 2014; Pichler and Shapiro, 2021; Rüschendorf, 1985)—a symmetrized analogue of causal transport distance. The origin of causal transport could be traced back to the Yamada–Watanabe criterion for stochastic differential equations (Yamada and Watanabe, 1971; Jean, 1980; Kurtz, 2014). In optimal transport theory, Lassalle (2013) investigated the transport problem in continuous time under the causal constraints, and Backhoff et al. (2017) studied a discrete-time analogue. Causal transport has been applied to continuous-time stochastic optimization in Acciaio et al. (2020), as well as other areas such as stochastic control (Acciaio et al., 2019) and machine learning (Xu et al., 2020). In discrete time stochastic programming, the nested distance has been exploited to study the stability and sensitivity of multistage stochastic programming (Pflug, 2010; Pflug and Pichler, 2012, 2014, 2015, 2016; Bartl and Wiesel, 2023).

Our problem can be viewed as a two-stage DRO with causal transport distance. After our paper's first draft appeared online, several works studied DRO with causal transport distance. Gao et al. (2024a) studied the dynamic programming reformulation for multistage DRO with nested distance. Jiang (2024) derives duality for the DRO problem with causal transport penalty. Compared with their methodology, our constructive proof of duality enables the characterization of the worst-case distribution, and we develop tractable reformulations for decision-rule optimization.

Another related work is Shen et al. (2024), which proposes a 2-Wasserstein distributionally robust contextual bandit framework that places a Wasserstein ball around the context distribution together with a family of Wasserstein balls for the conditional reward distributions. From this perspective, our dual reformulation is closely related. However, their ambiguity sets impose hard constraints with a shared radius across all contexts, leading to a two-level nested optimization.

*On decision-rule approach in adjustable robust optimization.* In the literature for adjustable robust optimization, different choices of decision rules have been thoroughly investigated, including affine families (Chen et al., 2008; Bertsimas et al., 2010, 2011; Bertsimas and Goyal, 2012; Iancu et al., 2013; El Housni and Goyal, 2021; Bertsimas et al., 2022; Georghiou et al., 2025), k-adaptability (Hanasusanto et al., 2015, 2016; Subramanyam et al., 2019), iterative splitting of uncertainty sets (Postek and Hertog, 2016), binary decision rules (Bertsimas and Georghiou, 2015), non-parametric Markovian stopping rules (Sturt, 2023), etc. Most of these works do not consider side information in their problem formulations. Bertsimas et al. (2022) considers dynamic decision-making with side information using affine decision rules, whereas we consider general decision rules in a static setting; and Zhang et al. (2024) considers the newsvendor problem with Wasserstein distance, whereas we consider a different uncertainty set, and we adopt a completely different proof strategy and obtain a broader class of optimal policies for adjustable robust optimization that encapsulates the Shapley policy proposed therein.

The rest of the paper proceeds as follows. We introduce the causal transport distance and corresponding robust model in Section 2. In Section 3, we develop a duality result for evaluating the worst-case expected cost by exploiting the structure of the worst-case distribution. In Section 4, we consider the outer optimization over affine decision rules and over all decision rules. Finally, we present numerical results in Section 5 and conclude the paper in Section 6. Proofs and additional results are deferred to the Appendices.

## 2. Distributionally Robust Optimization with Causal Transport Distance

In this section, we briefly introduce notation and provide some background on distributionally robust optimization with causal transport distance.

**Notation**. Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$ be subsets of normed vector spaces. For notational simplicity, the subscripts in $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Z}}$ will be omitted as long as they can be inferred from the context. Let $p \in [1, \infty)$ and denote by $q$ its Hölder conjugate number, i.e., $\frac{1}{p} + \frac{1}{q} = 1$. We denote by $\mathcal{P}_p(\mathcal{Z})$ the set of probability measures of $\mathcal{Z}$ with finite $p$-th moment, namely, $\mathbb{Q} \in \mathcal{P}_p(\mathcal{Z})$ if and only if $\mathbb{E}_{Z\sim\mathbb{Q}}[\|Z\|^p] < \infty$. The support of a distribution is denoted by $\operatorname{supp}\mathbb{Q}$. The set of all possible transport plans between the given marginals $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}(\mathcal{X} \times \mathcal{Z})$, on the product space $(\mathcal{X} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Z})$, is denoted as $\Gamma(\mathbb{Q}_1, \mathbb{Q}_2)$.

### 2.1. Causal Transport Distance

Our motivation to adopt the causal transport distance in DRO is illustrated by the following example. Consider the feature-based newsvendor problem, where historical demand for a product in a vending machine is influenced by covariates such as location, weather, and economic conditions. In such a scenario, the causal relationship is directed: the distributional uncertainty of the features can lead to uncertainty in the demand, but not vice versa. Therefore, if we consider a data perturbation map

$$T : (\widehat{X}, \widehat{Z}) \mapsto T(\widehat{X}, \widehat{Z}) = \big(T_1(\widehat{X}, \widehat{Z}), T_2(\widehat{X}, \widehat{Z})\big),$$

the perturbation of features (e.g., location, weather, economic state) should not depend on the demand, but the perturbation of demand can be affected by the perturbation of features. In other words, the perturbation map should have the form

$$T(\widehat{X}, \widehat{Z}) = \big(T_1(\widehat{X}), T_2(\widehat{X}, \widehat{Z})\big),$$

where $\widehat{X}$ is transported to $T_1(\widehat{X})$, and given $\widehat{X}$, $X = T_1(\widehat{X})$ is a constant. This implies that $X$ is conditionally independent of $\widehat{Z}$, represented as $X \perp \widehat{Z} \mid \widehat{X}$. Extending upon this notion of conditional independence, we introduce the following definition of causal transport plan and causal transport distance.

Definition 1 (Causal Transport Distance). A joint distribution $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$ is called a *causal transport plan* if for $((\widehat{X}, \widehat{Z}), (X, Z)) \sim \gamma$, $X$ and $\widehat{Z}$ are conditionally independent given $\widehat{X}$:

$$X \perp \widehat{Z} \mid \widehat{X}.$$

We denote by $\Gamma_c(\widehat{\mathbb{P}}, \mathbb{P})$ the set of all transport plans $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$ that are causal. Let $p \in [1, \infty)$. The *p-causal transport distance* between $\widehat{\mathbb{P}}$ and $\mathbb{P}$ is defined as

$$\mathsf{C}_p(\widehat{\mathbb{P}}, \mathbb{P}) := \left( \inf_{\gamma \in \Gamma_c(\widehat{\mathbb{P}}, \mathbb{P})} \mathbb{E}_{((X,Z),(\widehat{X},\widehat{Z})) \sim \gamma} \left[ \|X - \widehat{X}\|^p + \|Z - \widehat{Z}\|^p \right] \right)^{1/p}. \qquad \diamondsuit$$

The conditional independence condition in Definition 1 basically means that the destination $X$ of a sample in a causal transport plan should depend only on the origin $\widehat{X}$ but not on the associated information of $\widehat{Z}$. There are other equivalent definitions of a causal transport plan, which are provided in Appendix EC.1.

Like the Wasserstein distance, the causal transport distance finds the minimal transport cost between two distributions, where norms capture the geometry of the data space and the similarity between samples. Nevertheless, causal transport distance differs from Wasserstein distance in the involved class of transport plans: Wasserstein distance considers all transport plans with given marginals, while causal transport distance restricts causal transport plans as defined in Definition 1.

Let us use the following example to visually explain a causal transport plan.

EXAMPLE 3 (CAUSAL TRANSPORT BETWEEN COLORED IMAGES). Let $\mathcal{X} = \{1, 2, \ldots, H\}^2$, where $H$ represents the width of a squared image, and let $\mathcal{Z} = \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$, representing the three color channels, red (R), green (G), and blue (B). One can identify $\mathcal{Z}$ with the three basis vector of $\mathbb{R}^3$. A bitmap image stores the position-color information of an image via an $H \times H \times 3$ tensor $A = (A_{ijk})_{i,j\in\{1,2,\ldots,H\},k\in\{1,2,3\}}$. Its $(i, j, k)$-th entry $A_{ijk} \in \{0, 1, \ldots, 255\}$ represents the 8-bit indexed color at pixel position $(i, j)$ in the $k$-th channel. With a normalizing constant $M = \sum_{i,j,k} A_{ijk}$, the tensor $A/M$ represents a probability mass function on $\mathcal{X} \times \mathcal{Z}$. Let us equip norms $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_{\ell^1}$ and $\|\cdot\|_{\mathcal{Z}} = cH\|\cdot\|_{\ell^\infty}$, where $c$ is a scaling parameter.

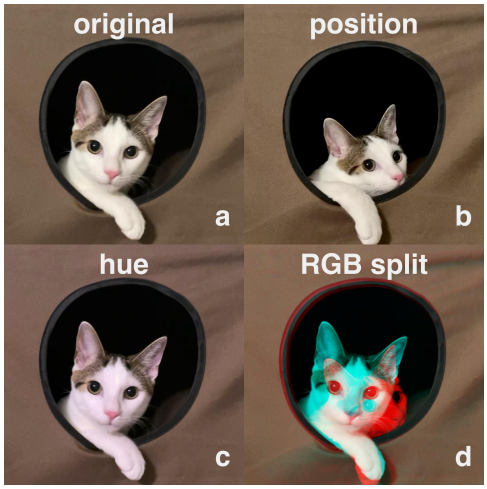


**Figure 2** **An image (a) and its variations by shifting the position (b), adjusting the hue (c), or splitting the RGB channels (d)**

Figure 2 contains four images of a cat: (a)(b)(c) can be viewed as real natural images with different poses or lighting conditions, whereas (d) can be viewed as an artificial image in which the pose exhibited via the red channel is different from that via the green/blue channel.

(I) The movement of the cat yields a causal transport plan from (a) to (b), as under such movement, the destination $(X, Z)$ in (b) of a position-channel pair $(\widehat{X}, \widehat{Z})$ in (a) depends only on its original position $\widehat{X}$ but not on the channel information $\widehat{Z}$, or put it differently, all channels are moved in the same way from $\widehat{X}$ to $X$ without changing the channel value $\widehat{Z}$. This matches precisely the definition of causal transport.

(II) The cats in (a) and (c) have identical poses but different hue values. Changing the hue values of an image would affect its RGB values and, in turn, the distribution on $\mathcal{Z}$. Such color adjustment (changing RGB values while fixing the position) defines a causal transport plan from (a) to (c). Indeed, under such a movement, a position-channel pair $(\widehat{X}, \widehat{Z})$ in (a) keeps its position in $c$, namely, $X = \widehat{X}$, regardless of the value of $\widehat{Z}$. Note that in a causal transport plan, we allow the destination $Z$ of $\widehat{Z}$ to depend on both $\widehat{X}$ and $\widehat{Z}$, that is, at each position in the image, changes in color are permitted.

(III) The green and blue channels of (d) have the same pose as (a), whereas the red channel of (d) has the same pose as (b). If we consider a transport plan that keeps a position-channel pair $(\widehat{X}, \widehat{Z})$ if $\widehat{Z} \in \{\mathrm{G}, \mathrm{B}\}$, and transport it according to the cat's movement if $\widehat{Z} = \mathrm{R}$, then such a transport plan is *not causal*, because given $\widehat{X}$, where this position-channel pair is transported, depends on the channel information $\widehat{Z}$.

**Table 1 Distance between Figure 2(a) and the other three variations**

| Variations | (b) | (c) | (d) |
|---|---|---|---|
| Wasserstein distance | 2.434 | 2.109 | 0.677 |
| Causal transport distance | 2.543 | 2.161 | **7.489** |

In Table 1, we compute the Wasserstein distance and causal transport distance between Fig. 2(a) and the other three variations, with $H = 32$ and $c = 4$. We find that the causal transport distance between Fig. 2(a) and the artificial image Fig. 2(d) is much larger than that between Fig. 2(a) and natural images Fig. 2(b)(c). In contrast, the Wasserstein distance fails to capture such an intuition. ♣

As hinted in Example 3, one of the main advantages of causal transport distance over Wasserstein distance is that it captures the structure of the conditional distribution. To further illustrate this, let us revisit the toy Example 2.

EXAMPLE 2 (REVISITED). We compute the causal transport distance and the Wasserstein distance between $\widehat{\mathbb{P}}$ and $\mathbb{P}$ shown in Example 2. Since the conditional distribution of $\mathbb{P}$ is a Dirac measure for every $x$, the causal transport distance between $\widehat{\mathbb{P}}$ and $\mathbb{P}$ is uniformly bounded from below by a positive constant for all $\varepsilon > 0$. In fact, it is not hard to see that the only causal transport plan is the independent joint distribution $\widehat{\mathbb{P}} \otimes \mathbb{P}$, so

$$\begin{aligned}\mathsf{C}_p(\widehat{\mathbb{P}}, \mathbb{P})^p &= \frac{1}{\sin\varepsilon}\int_0^{\sin\varepsilon} |x-0|^p \,\mathrm{d}x + \frac{1}{\cos\varepsilon}\int_0^1\int_0^{\cos\varepsilon} |\widehat{z}-z|^p \,\mathrm{d}z\,\mathrm{d}\widehat{z} \\ &= \frac{\sin^p\varepsilon}{p+1} + \frac{1+\cos^{p+2}\varepsilon - (1-\cos\varepsilon)^{p+2}}{(p+1)(p+2)\cos\varepsilon} \\ &= \Big((1+p)(1+\frac{p}{2})\Big)^{-\frac{1}{p}} + O(\varepsilon).\end{aligned}$$

As a result, $\mathbb{P}$ would not belong to the uncertainty set induced by the causal transport distance with a small radius. This is consistent with our intuition. In contrast, for the Wasserstein distance, observe that the optimal transport plan is simply the rotation transform, thereby the Wasserstein distance is $(p+1)^{-\frac{1}{p}}(\sin^p\varepsilon + (1-\cos\varepsilon)^p)^{\frac{1}{p}} = O(\varepsilon)$, which is small whenever the angle between the two line segments is small. Consequently, any Wasserstein uncertainty set with a positive radius contains infinitely many distributions with dramatically different conditional information structures from the nominal one, and therefore may lead to an overly conservative solution. ♣

Next, we point out an important property of the uncertainty set constructed using the causal transport distance: for any $\widehat{\mathbb{P}} \in \mathcal{P}(\mathcal{X}\times\mathcal{Z})$ and $\rho > 0$, the set $\mathfrak{M} = \{\mathbb{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Z}) : \mathsf{C}_p(\widehat{\mathbb{P}}, \mathbb{P}) \le \rho\}$ is convex, as indicated in the following lemma.

LEMMA 1 **(Convexity)**. *If $\gamma^{(0)}$ and $\gamma^{(1)}$ are two causal transport plans from $\widehat{\mathbb{P}}$ to $\mathbb{P}^{(0)}$ and $\mathbb{P}^{(1)}$ respectively, then for any $q \in [0,1]$, $\gamma^q := (1-q)\gamma^{(0)} + q\gamma^{(1)}$ is also a causal transport plan from $\widehat{\mathbb{P}}$ to $\mathbb{P}^{(q)} = (1-q)\mathbb{P}^{(0)} + q\mathbb{P}^{(1)}$. Moreover, everything follows even if we replace $q$ by any measurable function $q : \mathcal{X} \to [0,1]$.*

We remark that the direction of the transport plan matters: if $\gamma^{(0)}$ and $\gamma^{(1)}$ are two causal transport plans from $\widehat{\mathbb{P}}^{(0)}$ and $\widehat{\mathbb{P}}^{(1)}$ to $\mathbb{P}$ respectively, we cannot assert that their convex combination $\gamma^{(q)}$ is also a causal transport plan. For a counterexample, please refer to Fig. 1.17 in Pflug and Pichler (2014).

### 2.2. Distributionally Robust Formulation

Based on the definition in the previous subsection, we study the following distributionally robust optimization problem with causal transport distance

$$v_{\mathsf{P}} := \inf_{f\in\mathcal{F}} \max_{\mathbb{P}\in\mathfrak{M}} \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)], \text{ where } \mathfrak{M} = \left\{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Z}):\ \mathsf{C}_p(\widehat{\mathbb{P}},\mathbb{P})\le\rho\right\}. \qquad (\mathsf{P})$$

Below, we list a few examples.

Example 4 (Conditional Mean Estimation). The conditional mean of $Z$ given $X$ can be estimated by minimizing the square loss $(f(X)-Z)^2$. Thus, we consider the following robust conditional mean estimation problem

$$\inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathfrak{M}} \mathbb{E}_{(X,Z)\sim\mathbb{P}}\left[(f(X)-Z)^2\right]. \qquad \clubsuit$$

Example 5 (Feature-based Newsvendor). Let $h$ and $b$ represent the unit overage cost and the unit underage cost, respectively, and let $Z$ be the random demand and $X$ be the covariate features. The goal is to minimize the newsvendor cost function $\Psi(w,z) = h(w-z)_+ + b(z-w)_+$. Consider

$$\inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathfrak{M}} \mathbb{E}_{(X,Z)\sim\mathbb{P}}\left[h(f(X)-Z)_+ + b(Z-f(X))_+\right].$$

Note that this model also serves as the conditional $\frac{b}{b+h}$-quantile estimation. In particular, when $h=b=1$, this is the conditional median estimation. $\clubsuit$

Example 6 (Personalized Pricing). Consider an affine demand model $D(w) = z_1 w + z_2 = Z^\top \begin{pmatrix} w \\ 1 \end{pmatrix}$, where $w$ is the price and $z$ are unknown coefficients, with $z_2 > 0$ representing the demand at zero price and $z_1 < 0$ representing the price sensitivity coefficient, which is the rate at which the price affects the demand. In practice, both coefficients $z_1$ and $z_2$ may exhibit heterogeneity across populations. As such, we model it as a two-dimensional random variable $Z$ that is affected by the contextual information $X$, which the decision maker can use to adjust the price directly or indirectly through personalized promotion. The revenue is calculated as $w(Z_1 w + Z_2)$. Consider revenue maximization with personalized pricing

$$\inf_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathfrak{M}} \mathbb{E}_{(X,Z)\sim\mathbb{P}} \left[-f(X) Z^\top \begin{pmatrix} f(X) \\ 1 \end{pmatrix}\right]. \qquad \clubsuit$$

In the last example, we consider a contextual linear optimization problem where the decision rule is restricted to be affine.

EXAMPLE 7 (CONTEXTUAL LINEAR OPTIMIZATION WITH AFFINE DECISION RULE). Consider a contextual linear optimization problem in which one minimizes the loss function $\Psi(w,z)=w^\top z$. Take a linear policy class $\mathcal{F}_\Theta$ defined by

$$\mathcal{F}_\Theta=\left\{x\mapsto B^\top x+\delta:\ (B,\delta)\in\Theta\right\},\ \text{where}\ \Theta=\left\{(B,\delta)\in\mathbb{R}^{d\times m}\times\mathbb{R}^m:\ B^\top x+\delta\in\mathcal{D},\forall x\in\mathcal{X}\right\}, \tag{4}$$

so that $f(\mathcal{X})\subset\mathcal{D}$ for each $f\in\mathcal{F}_\Theta$. The robust contextual linear optimization problem is given by

$$\inf_{f\in\mathcal{F}_\Theta}\sup_{\mathbb{P}\in\mathfrak{M}}\mathbb{E}_{(X,Z)\sim\mathbb{P}}\left[f(X)^\top Z\right]. \qquad \clubsuit$$

## 3. Evaluating the Worst-case Expectation

In this section, we develop a tractable reformulation for the inner maximization of (P) based on strong duality. As a byproduct of our proof, we also derive the structure of the worst-case distribution, demonstrating that our choice of a distributional uncertainty set based on causal transport distance helps preserve the conditional information structure of the nominal distribution in the worst case.

Throughout this paper, we make the following assumption, which focuses on the data-driven setting where the nominal distribution is discrete, although our proof technique can be extended to a general metric space with additional technical treatment.

ASSUMPTION 1. *$\mathcal{X}$, $\mathcal{Z}$, $\mathcal{D}$ are subsets of normed vector spaces. The cost function $\Psi:\mathcal{D}\times\mathcal{Z}\to\mathbb{R}$ is measurable. The nominal distribution $\widehat{\mathbb{P}}\in\mathcal{P}(\mathcal{X}\times\mathcal{Z})$ is a discrete probability measure*

$$\widehat{\mathbb{P}}=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}\delta_{(\widehat{x}_k,\widehat{z}_{ki})},\qquad \textit{with}\ \sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}=1.$$

### 3.1. Strong Duality Reformulation

We begin by developing a tractable reformulation by deriving its strong dual. For a fixed decision rule $f$, we define the primal problem as

$$v_{\mathsf{P}}^f:=\max_{\mathbb{P}\in\mathfrak{M}}\ \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)], \tag{$\mathsf{P}^f$}$$

and the dual problem as

$$v_{\mathsf{D}}^f := \inf_{\lambda\geq 0}\left\{\lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid \widehat{X}\right]-\lambda\|x-\widehat{X}\|^p\right\}\right]\right\}. \tag{$\mathsf{D}^f$}$$

The dual variable $\lambda$ corresponds to the Lagrangian multiplier of the causal constraint in the primal problem. We will show that ($\mathsf{P}^f$) and ($\mathsf{D}^f$) are equal, leading to the main result of Theorem 1 by taking the infimum over $f$.

To prove the strong duality, we first develop a relatively straightforward weak duality result.

Proposition 1 **(Weak Duality)**. *Let $f:\mathcal{X}\to\mathcal{D}$ be a measurable function. Then $v_{\mathsf{P}}^f \leq v_{\mathsf{D}}^f$.*

*Proof.* The proof is based on an application of Lagrangian weak duality. First, we derive from the Lagrangian weak duality the following

$$\begin{aligned}
v_{\mathsf{P}}^f &= \sup_{\mathbb{P}}\left\{\mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)]:\ \mathsf{C}_p(\widehat{\mathbb{P}},\mathbb{P})^p\leq\rho^p\right\}\\
&= \sup_{\mathbb{P}}\inf_{\lambda\geq 0}\left\{\mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)]-\lambda\left(\mathsf{C}_p(\widehat{\mathbb{P}},\mathbb{P})^p-\rho^p\right)\right\}\\
&\leq \inf_{\lambda\geq 0}\sup_{\mathbb{P}}\left\{\mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)]-\lambda\left(\mathsf{C}_p(\widehat{\mathbb{P}},\mathbb{P})^p-\rho^p\right)\right\}.
\end{aligned}$$

Since for any $\gamma\in\Gamma_c(\widehat{\mathbb{P}},\mathbb{P})$,

$$\mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)] = \mathbb{E}_{((X,Z),(\widehat{X},\widehat{Z}))\sim\gamma}[\Psi(f(X),Z)],$$

so we can write

$$\mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)]-\lambda\left(\mathsf{C}_p(\widehat{\mathbb{P}},\mathbb{P})^p-\rho^p\right) = \lambda\rho^p + \sup_{\gamma\in\Gamma_c(\widehat{\mathbb{P}},\mathbb{P})}\mathbb{E}_\gamma\left[\Psi(f(X),Z)-\lambda\|X-\widehat{X}\|^p-\lambda\|Z-\widehat{Z}\|^p\right].$$

By the tower property,

$$\begin{aligned}
\mathbb{E}_\gamma[\cdot] &= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\gamma_{X|\widehat{X}}}\left[\mathbb{E}_{\gamma_{\widehat{Z}|(\widehat{X},X)}}\left[\mathbb{E}_{\gamma_{Z|(\widehat{X},\widehat{Z},X)}}\left[\cdot\mid\widehat{X},\widehat{Z},X\right]\mid\widehat{X},X\right]\mid\widehat{X}\right]\right]\\
&= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\gamma_{X|\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\mathbb{E}_{\gamma_{Z|(\widehat{X},\widehat{Z},X)}}\left[\cdot\mid\widehat{X},\widehat{Z},X\right]\mid\widehat{X},X\right]\mid\widehat{X}\right]\right]
\end{aligned}$$

where we use $\gamma_{\widehat{Z}|(\widehat{X},X)} = \widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}$ for a.e.-$(\widehat{X}, X)$ because $\gamma$ is causal. Therefore we have

$$
\begin{aligned}
&\mathbb{E}_\gamma \left[\Psi(f(X), Z) - \lambda\|X - \widehat{X}\|^p - \lambda\|Z - \widehat{Z}\|^p\right] \\
&= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\gamma_{X|\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\mathbb{E}_{\gamma_{Z|(\widehat{X},\widehat{Z},X)}}\left[\Psi(f(X), Z) - \lambda\|X - \widehat{X}\|^p - \lambda\|Z - \widehat{Z}\|^p \mid \widehat{X}, \widehat{Z}, X\right] \mid \widehat{X}, X\right] \mid \widehat{X}\right]\right] \\
&\leq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x), z) - \lambda\|z - \widehat{Z}\|^p\right\} \mid \widehat{X}\right] - \lambda\|x - \widehat{X}\|^p\right\}\right].
\end{aligned}
$$

This completes the proof of the weak duality. □

The strong duality result states as follows.

THEOREM 1 **(Strong Duality)**. *Let $f : \mathcal{X} \to \mathcal{D}$ be a measurable function. Then $v_{\mathsf{P}}^f = v_{\mathsf{D}}^f$.*

*Proof Sketch.* The proof idea is to construct a nearly worst-case distribution of the primal problem based on the first-order optimality condition of the weak dual problem ($\mathsf{D}^f$). Conceptually, the duality of Causal transport DRO shares common ideas with the duality proof of Wasserstein DRO (Gao and Kleywegt, 2023): both establish duality by constructing a nearly worst-case scenario via the first-order optimality conditions of the weak dual problem. However, the construction in our setting differs in several key aspects. Unlike standard Wasserstein DRO—where the decision variable typically interacts symmetrically with the joint distribution—our formulation requires handling the asymmetric roles of the variables $X$ and $Z$ under causally ordered constraints. This asymmetry introduces new technical challenges in characterizing admissible perturbations and conditional transport maps. As a result, the worst-case distribution in our framework involves a two-layer construction: first over marginal perturbations, and then over conditional distributions. The dual characterization must therefore simultaneously account for both layers while ensuring measurability and feasibility under causal constraints. This layered structure, combined with the topological and measurability requirements inherent in causal transport, makes our duality proof technically nontrivial and distinct from existing formulations.

The worst-case distribution maximizes the expected loss within a given transport budget. With a fixed dual variable $\lambda$, the worst-case distribution for the soft constraint problem

$$\max_{\mathbb{P}\in\mathcal{P}(\mathcal{X}\times\mathcal{Z})} \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X), Z)] - \lambda \mathsf{C}_p(\widehat{\mathbb{P}}, \mathbb{P})^p$$

is obtained by moving $\widehat{z}_{ki}$ toward the maximizer of the innermost maximization problem of ($\mathsf{D}^f$)

$$\Upsilon(\lambda; x, \widehat{z}_{ki}) := \sup_{z\in\mathcal{Z}}\left\{\Psi(f(x), z) - \lambda\|z - \widehat{z}_{ki}\|^p\right\},$$

and moving $\widehat{x}_k$ toward the maximizer of the maximization problem

$$\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\Upsilon(\lambda;x,\widehat{Z})\mid\widehat{X}=\widehat{x}_k]-\lambda\|x-\widehat{x}_k\|^p\right\}.$$

One can see that such a transport plan is causal: the maximizer $x$ is the perturbation of $\widehat{x}_k$ in the transport plan, and $x$ is solely determined by $\widehat{x}_k$, independent of which $\widehat{z}_{ki}$ it is associated with. If both maximizers over $x$ and over $z$ exist and are unique at the critical $\lambda^*$ dual to the given transport distance $\rho^p$, then the transport plan would induce a worst-case distribution. If the maximizer does not exist or is not unique, two alternative transport plans are considered: one produces a feasible but suboptimal distribution, and the other, although infeasible, achieves a higher objective value. Interpolating between these distributions allows for a near-optimal solution to the primal problem.

As can be seen from the definition of $\Upsilon$, the worst-case distribution for the soft constraint problem is obtained by moving mass with loss-to-distance "efficiency" higher than $\lambda$. Efficiency here refers to the ratio of gain (or loss reduction) to the $p$-th power of distance. Specifically, the efficiency of moving $\widehat{x}$ to $x$ is $\frac{\mathbb{E}[\Upsilon(\lambda;x,\widehat{Z})]-\mathbb{E}[\Upsilon(\lambda;\widehat{x},\widehat{Z})]}{\|x-\widehat{x}\|^p}$, in which $\Upsilon$ already incorporated the efficiency of moving $\widehat{z}$ to $z$, calculated by $\frac{\Psi(f(x),z)-\Psi(f(x),\widehat{z})}{\|z-\widehat{z}\|^p}$.

There are several possibilities where the near-optimal distribution is located, depending on the critical threshold $\lambda^*$ that minimizes ($\mathsf{D}^f$). Indeed, the dual objective function

$$h(\lambda):=\lambda\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\lambda\|x-\widehat{X}\|^p\right\}\right] \tag{5}$$

is an extended-real-valued convex function of $\lambda$. $h(\lambda)-\lambda\rho^p$ is monotonically decreasing in $\lambda$ and coincides with the above soft constraint problem. Let $\kappa\in[0,+\infty]$ be the smallest value such that the dual objective is finite in $(\kappa,+\infty)$. The infimum over $\lambda$ in ($\mathsf{D}^f$) can have several possibilities:

- Case 1: $\kappa=+\infty$, so the dual objective is $+\infty$ for any $\lambda>0$. This means that by transporting an arbitrarily small distance, one can generate an arbitrarily large loss.
- Case 2: $\kappa<+\infty$ and minimization over $\lambda$ in ($\mathsf{D}^f$) is achieved in the interior of $(\kappa,+\infty)$. The dual objective can be arbitrarily large if $\lambda$ is smaller than $\kappa$, but it would require transporting mass that exhausts the transport distance budget. We interpolate two transport plans: moving all the masses with "efficiency" above $\lambda_1<\lambda^*$ (superoptimal but infeasible) v.s. moving all the masses with efficiency above $\lambda_2>\lambda^*$ (feasible but suboptimal).

- Case 3: $\kappa < +\infty$ and $v_{\mathsf{D}}^f$ is minimized at $\kappa$. Moving all the mass with efficiency strictly above $\kappa$ does not exhaust the transport distance budget. This is further divided into
  - Case 3.1: $\kappa = 0$. Any positive $\lambda$ corresponds to a finite soft loss. We simply move all the mass with positive efficiency.
  - Case 3.2: $\kappa > 0$. We again interpolate two transport plans: moving all the masses with efficiency above $\lambda_2 > \lambda^*$ (feasible but suboptimal) v.s. moving *some* of the masses with efficiency above $\kappa_1 < \lambda^*$ (superoptimal but infeasible). We can only move the latter up to some distance, in contrast to Case 2, because moving them all would travel an infinite distance.

We refer to the next subsection for a more detailed construction of a worst-case distribution and Appendix EC.3 for a complete proof. □

Remark 1 (Comparison with Wasserstein DRO). Recall the Wasserstein DRO problem

$$\sup_{\mathbb{P}} \left\{ \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)] : \ \mathsf{W}_p(\widehat{\mathbb{P}},\mathbb{P}) \le \rho \right\},$$

which has the following equivalent dual form (Gao and Kleywegt, 2023; Zhang et al., 2025)

$$\begin{aligned} v_{\mathsf{WD}}^f &:= \inf_{\lambda\ge 0} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}} \left[ \sup_{\substack{x\in\mathcal{X}\\ z\in\mathcal{Z}}} \left\{ \Psi(f(x),z) - \lambda\|z-\widehat{Z}\|^p - \lambda\|x-\widehat{X}\|^p \right\} \right] \right\} \\ &= \inf_{\lambda\ge 0} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[ \sup_{x\in\mathcal{X}} \left\{ \sup_{z\in\mathcal{Z}} \left\{ \Psi(f(x),z) - \lambda\|z-\widehat{Z}\|^p \right\} - \lambda\|x-\widehat{X}\|^p \right\} \mid \widehat{X} \right] \right] \right\}. \end{aligned}$$

Comparing it with the dual problem ($\mathsf{D}^f$) of causal transport distance DRO, the difference is the swap of supremum over $x$ and the conditional expectation of $\widehat{Z}$ given $\widehat{X}$. Hence, if the switching does not change the objective value, which holds, for instance, when the conditional distribution $\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}$ is a Dirac measure for every $\widehat{X}$, then the Wasserstein DRO dual problem and causal transport distance DRO dual problems are equal. From a primal point of view, if $\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}$ is Dirac for every $\widehat{X}$, then every transport plan from $\widehat{\mathbb{P}}$ to $\mathbb{P}$ is causal. In this case, the causal transport distance DRO and Wasserstein DRO coincide. Intuitively, if every conditional distribution $\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}$ is Dirac, then the nominal distribution does not have any meaningful conditional information structure to exploit, and thus the causal transport distance DRO reduces to Wasserstein DRO.

Without considering the causality constraint, the optimal strategy is the *greedy* one. When a unit of mass is moved from $(\widehat{x},\widehat{z})$ to $(x,z)$, it generates a "revenue" of $\Psi(f(x),z) -$

$\Psi(f(\widehat{x}),\widehat{z})$, while incurring a transport distance $\|x-\widehat{x}\|^p+\|z-\widehat{z}\|^p$. Here, the "revenue" refers to the gain (increase in the objective) relative to the cost of moving the mass. The efficiency of this transportation is thus $\frac{\Psi(f(x),z)-\Psi(f(\widehat{x}),\widehat{z})}{\|x-\widehat{x}\|^p+\|z-\widehat{z}\|^p}$. It will move $(\widehat{x},\widehat{z})$ to a destination with the (near-) highest efficiency, and $(\widehat{x},\widehat{z})$ is moved only after all other sources $(\widehat{x}',\widehat{z}')$'s with higher efficiency have been depleted. This greedy strategy is reflected in $v^f_{\mathsf{WD}}$. The dual objective computes the net profit of transporting all the mass with efficiency higher than threshold $\lambda$ with transport cost multiplied by a factor of $\lambda$ (toll rate), and $v^f_{\mathsf{WD}}$ computes the revenue by reimbursing the transport cost $\lambda\rho^p$ and then searches for the critical threshold $\lambda^*$. ♢

### 3.2. Worst-case Distribution

In this subsection, we investigate the structure of the worst-case distribution and its existence conditions. Compared with the results in Section 3.1, in the following result, we require $\mathcal{X}$ and $\mathcal{Z}$ to be finite-dimensional and thus locally compact and require some continuity assumptions on $\Psi$ so that the maximizers are attainable.

THEOREM 2 **(Worst-case Distribution)**. *Suppose $\mathcal{X},\mathcal{Z}$ are finite dimensional, and $\Psi(f(\cdot),\cdot)$ is upper semi-continuous. If the optimal value of $(\mathsf{D}^f)$ is attained at some $\lambda^*>\kappa$ for $\kappa$ specified in Lemma EC.2, then a worst-case distribution exists and has the following form*

$$\mathbb{P}^*=\sum_{k\neq k_0}\sum_{i=1}^{n_k}\widehat{p}_{ki}\delta_{(x^*_k,z^*_{ki})}+\sum_{i=1}^{n_{k_0}}\widehat{p}_{k_0i}\left(q\delta_{(\overline{x}_{k_0},\overline{z}_{k_0i})}+(1-q)\delta_{(\underline{x}_{k_0},\underline{z}_{k_0i})}\right),$$

*where $1\le k_0\le K$, $0\le q\le 1$, $(x^*_k,z^*_{ki})=(\overline{x}_k,\overline{z}_{ki})$, and for every $k$ and $i$,*

$$\overline{x}_k,\underline{x}_k\in\arg\max_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda^*\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}=\widehat{x}_k\right]-\lambda^*\|x-\widehat{x}_k\|^p\right\},$$

$$\overline{z}_{ki}\in\arg\max_{z\in\mathcal{Z}}\left\{\Psi(f(\overline{x}_k),z)-\lambda^*\|z-\widehat{z}_{ki}\|^p\right\},\qquad \underline{z}_{ki}\in\arg\max_{z\in\mathcal{Z}}\left\{\Psi(f(\underline{x}_k),z)-\lambda^*\|z-\widehat{z}_{ki}\|^p\right\}.$$

From Theorem 2, we see that there exists a worst-case distribution $\mathbb{P}^*$ supported on at most $N+n_{k_0}$ points, and its marginal $\mathbb{P}^*_X$ is supported on at most $K+1$ points. We demonstrate the structure of the worst-case distribution in Figure 3 (left). In this plot, the support of $\widehat{\mathbb{P}}$ is represented by '•', and we have $K=3$, $n_k=3$, $k=1,2,3$ and $k_0=2$. These points are transported to '⋆'s, which form the worst-case distribution $\mathbb{P}^*$. For $k=1,3$, we observe that $\widehat{x}_k$ is transported to $x^*_k$, and the conditional distribution $\mathbb{P}^*_{Z|X=x^*_k}$ has the same

structure as the conditional distribution $\widehat{\mathbb{P}}_{\widehat{Z}|X=\widehat{x}_k}$, both supported on 3 points with identical probability mass function $(\widehat{p}_{ki})_{i=1,2,3}$. Furthermore, $\widehat{x}_2$ is split into two values $\overline{x}_2$ and $\underline{x}_2$, and the conditional distributions $\mathbb{P}^*_{Z|X=\underline{x}_2}$, $\mathbb{P}^*_{Z|X=\overline{x}_2}$ have the same structure as the conditional distribution $\widehat{\mathbb{P}}_{\widehat{Z}|X=\widehat{x}_2}$, both supported on 3 points with identical probability mass function $(\widehat{p}_{2i})_{i=1,2,3}$.

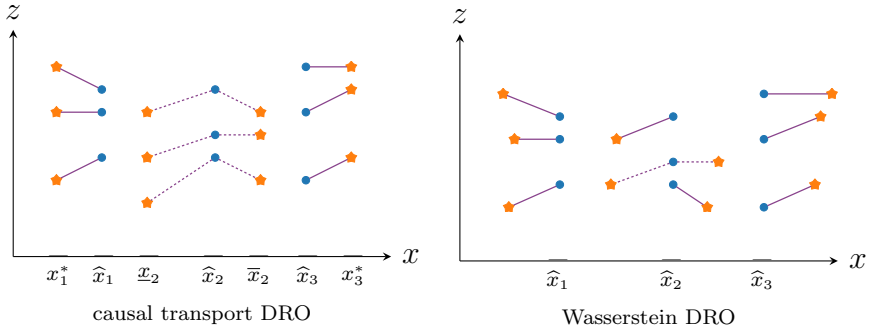


**Figure 3** **Structure of the worst-case distributions**

As a comparison, on the right side of Figure 3, we plot the worst-case distribution resulting from Wasserstein DRO. According to Gao and Kleywegt (2023), the worst-case distribution can be supported on $N+1$ points, and points with the same $x$-value could have different $x$-values after transportation or splitting. The conditional distributions of the worst-case distribution change completely, each of which is a Dirac measure. This example illustrates that the worst-case distribution of the causal transport distance DRO preserves the conditional information structure of the nominal distribution, whereas the Wasserstein DRO fails to do so.

We illustrate the worst-case distributions under Wasserstein DRO and causal transport DRO for the mean estimation problem as follows.

EXAMPLE 4 (REVISITED). Consider the conditional mean estimation problem in Example 4. We compare the worst-case distributions with 2-Wasserstein DRO and 2-causal transport DRO when the decision rule $f = f_{\text{true}}$ is the true conditional mean, and the uncertainty set radius is $\rho = 0.2$. As shown in Figure 4, in the worst-case Wasserstein DRO, the conditional information structure is not preserved. ♣

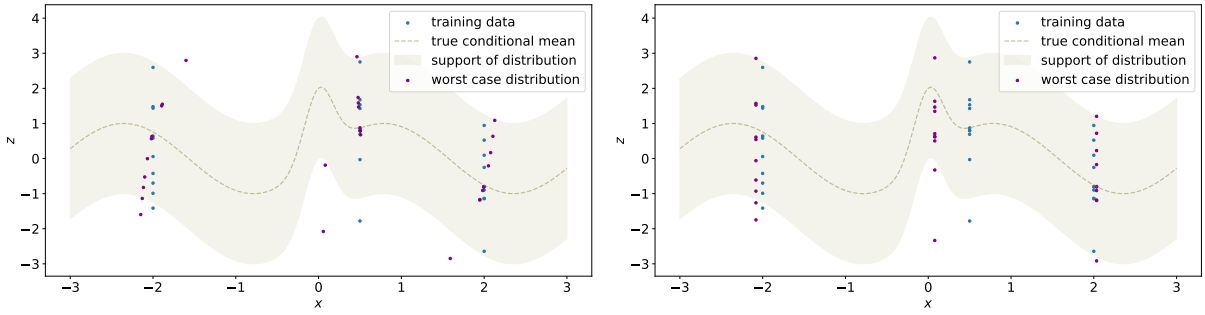


**Figure 4** **Structure of the 2-Wasserstein (left) v.s. causal (right) worst-case distributions for mean estimation.**

## 4. Finding the Optimal Decision Rule

In this section, we study the outer optimization over decision rules in (P). As a direct consequence of Theorem 1, problem (P) is equivalent to the following:

$$v_{\mathsf{D}} := \inf_{\substack{f\in\mathcal{F}\\ \lambda\geq 0}} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[ \sup_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[ \sup_{z\in\mathcal{Z}} \left\{ \Psi(f(x),z) - \lambda\|z-\widehat{Z}\|^p \right\} \mid \widehat{X} \right] - \lambda\|x-\widehat{X}\|^p \right\} \right] \right\}. \tag{D}$$

In particular, if we define $\|z-\widehat{z}\|_{\mathcal{Z}} := \infty \mathbf{1}\{z\neq\widehat{z}\}$, which is often used when the side information is relatively accurate, then (D) is simplified to

$$v_{\mathsf{D}} := \inf_{\substack{f\in\mathcal{F}\\ \lambda\geq 0}} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[ \sup_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[ \Psi(f(x),\widehat{Z}) \mid \widehat{X} \right] - \lambda\|x-\widehat{X}\|^p \right\} \right] \right\}. \tag{6}$$

The tractability of the optimization over $f\in\mathcal{F}$ depends on the class of decision rules $\mathcal{F}$. If $\mathcal{F}$ admits a finite-dimensional parameterization, such as an affine class, then the problem (D) is a finite-dimensional optimization, and we identify cases where the overall problem can be solved by off-the-shelf convex programming solvers (Section 4.1). Otherwise, if $\mathcal{F}$ is a non-parametric class, and particularly the class of all decision rules, then the optimization over $\mathcal{F}$ is an infinite-dimensional functional optimization, yet still, we identify cases where the overall problem can be solved efficiently (Section 4.3).

### 4.1. Optimizing over Affine Decision Rules

In this subsection, we provide tractable formulations when $\mathcal{F}$ is an affine class. Suppose affine functions in $\mathcal{F}$ are parametrized by $\Theta$:

$$\mathcal{F}_\Theta = \left\{ x\mapsto B^\top x+\delta :\ (B,\delta)\in\Theta \right\} \tag{7}$$

where $\Theta$ is a finite-dimensional convex set.

Our first result shows that (6) is tractable when $\Psi$ is affine in the decision variable $w$. The proof can be found in Appendix EC.4.

COROLLARY 1. *Suppose $\mathcal{F}=\mathcal{F}_\Theta$ as defined in (7), and $\Psi(\cdot,z)$ is affine for every $z$, that is, there exist functions $\beta(\cdot),b(\cdot)$ such that*

$$\Psi(w,z)=\beta(z)^\top w+b(z).$$

*Set*

$$\widehat{p}_k:=\sum_{i=1}^{n_k}\widehat{p}_{ki},\qquad \beta_k:=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\beta(\widehat{Z})\mid\widehat{X}=\widehat{x}_k],\qquad b_k:=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[b(\widehat{Z})\mid\widehat{X}=\widehat{x}_k].$$

*Then, the dual problem (6) is equivalent to the following convex programs. When $p=1$, (6) is equivalent to*

$$\inf_{(B,\delta)\in\Theta}\quad \rho^p\cdot\max_{k\in[K]}\|B\beta_k\|_*+\sum_{k=1}^K\widehat{p}_k\left(\beta_k^\top(B^\top\widehat{x}_k+\delta)+b_k\right).$$

*When $p\in(1,+\infty)$, (6) is equivalent to*

$$\inf_{\lambda\geq 0,(B,\delta)\in\Theta}\quad \lambda\rho^p+\sum_{k=1}^K\widehat{p}_k\left(\beta_k^\top(B^\top\widehat{x}_k+\delta)+b_k+\lambda(p-1)\left(\frac{\|B\beta_k\|_*}{\lambda p}\right)^{\frac{p}{p-1}}\right).$$

*Here $\|\cdot\|_*$ is the dual norm of $\|\cdot\|_{\mathcal{X}}$.*

As a special case, we assume further that $\Psi(w,z)$ is bilinear. When $p=2$, the above convex program can be written as a positive semidefinite program.

COROLLARY 2. *Suppose $\mathcal{F}=\mathcal{F}_\Theta$ as defined in (7) and $\Psi(w,z)$ is bilinear:*

$$\Psi(w,z)=w^\top Az+\beta^\top w+\alpha^\top z+b.$$

*Set*

$$\widehat{p}_k=\sum_{i=1}^{n_k}\widehat{p}_{ki},\qquad \bar{z}_k=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z}\mid\widehat{X}=\widehat{x}_k].\tag{8}$$

*Then (D) with $p=2$ is equivalent to the following positive semidefinite program*

$$\begin{array}{ll}\inf\limits_{\substack{(B,\delta)\in\Theta\\ \lambda\geq 0,\{y_k\}_k\subset\mathbb{R}}} & \lambda\rho^2+\sum\limits_{k=1}^K\widehat{p}_k y_k\\ \text{s.t.} & \begin{pmatrix}\lambda I & -\frac{1}{2}BA & -\frac{1}{2}B\beta-\lambda\widehat{x}_k\\ -\frac{1}{2}(BA)^\top & \lambda I & -\frac{1}{2}(A^\top\delta+\alpha)-\lambda\bar{z}_k\\ -\frac{1}{2}(B\beta)^\top-\lambda\widehat{x}_k^\top & -\frac{1}{2}(A^\top\delta+\alpha)^\top-\lambda\bar{z}_k^\top & y_k-\beta^\top\delta-b+\lambda\|\bar{z}_k\|^2+\lambda\|\widehat{x}_k\|^2\end{pmatrix}\succeq O, \qquad k\in[K].\end{array}$$

*Here $O$ stands for the zero matrix and $I$ represents the identity matrix.*

EXAMPLE 7 (REVISITED). We revisit the contextual linear optimization problem in Example 7, where the decision is restricted to a polygon $\mathcal{D}=\{w\in\mathbb{R}^m:\ Cw\le c\}$, and the context $X$ is bounded in an ellipsoid $\mathcal{X}=\left\{x\in\mathbb{R}^d:\ (x-x_0)^\top\Sigma(x-x_0)\le R\right\}$. Here $\Sigma\in\mathbb{R}^{d\times d}$ is symmetric and positive definite, $x_0\in\mathbb{R}^d$, $R>0$, $C\in\mathbb{R}^{L\times m}$, and $c\in\mathbb{R}^L$. $Cw\le c$ means $C_\ell^\top w\le c_\ell$, where $C_\ell^\top$ is the $\ell$-th row of $C$ and $c_\ell$ is the $\ell$-th entry of $c$, for each $\ell\in[L]$. $\Theta$ defined by (4) is convex. Using Corollary 2, (D) with $p=2$ can be reformulated as the following positive semidefinite program

$$\begin{aligned}\inf_{\substack{B\in\mathbb{R}^{d\times m},\delta\in\mathbb{R}^m\\ \lambda\ge0,\{y_k\}_k\in\mathbb{R}^K\\ \{\mu_k\}_k\ge0,\{\nu_\ell\}_\ell\ge0}}\quad & \lambda\rho^2+\sum_{k=1}^K\widehat{p}_k y_k\\ \text{s.t.}\quad & \begin{pmatrix}\lambda I+\mu_k\Sigma & -\frac{1}{2}B & -\lambda\widehat{x}_k-\mu_k\Sigma x_0\\ -\frac{1}{2}B^\top & \lambda I & -\frac{1}{2}\delta-\lambda\bar{z}_k\\ -\lambda\widehat{x}_k^\top-\mu_k(\Sigma x_0)^\top & -\frac{1}{2}\delta^\top-\lambda\bar{z}_k^\top & \begin{array}{r}y_k+\lambda\|\bar{z}_k\|^2+\lambda\|\widehat{x}_k\|^2\\ +\mu_k x_0^\top\Sigma x_0-\mu_k R\end{array}\end{pmatrix}\succeq O\quad\forall k\in[K],\\ & \begin{pmatrix}\nu_\ell\Sigma & -\frac{1}{2}BC_\ell-\nu_\ell\Sigma x_0\\ -\frac{1}{2}C_\ell^\top B^\top-\nu_\ell(\Sigma x_0)^\top & -C_\ell^\top\delta+c_\ell+\nu_\ell x_0^\top\Sigma x_0-\nu_\ell R\end{pmatrix}\succeq O\qquad\forall\ell\in[L].\end{aligned}$$

Recall $\widehat{p}_k$ and $\bar{z}_k$ are defined in (8). Detailed computations are provided in Appendix EC.5. ♣

Our second result shows (D) is tractable when the loss function $\Psi$ is quadratic and $p=2$. We study the Example 4 in detail, and we provide the proof in Appendix EC.4.

COROLLARY 3. *We revisit the conditional mean estimation problem in Example 4. Assume $\mathcal{X}=\mathbb{R}^d$, $\mathcal{Z}=\mathbb{R}^m$, and $\Theta=\mathbb{R}^{d\times m}\times\mathbb{R}^m$. Each space is equipped with the Euclidean norm $\ell^2$. With quadratic loss $\Psi(w,z)=\|w-z\|^2$ and affine policy $f(x)=B^\top x+\delta$, the dual problem (D) can be formulated as the following positive semidefinite program*

$$\begin{aligned}\inf_{\substack{(B,\delta)\in\Theta\\ \lambda\ge1,t\ge0\\ \{y_k\}\in\mathbb{R}^K\\ W\in\mathbb{R}^{d\times d},u\in\mathbb{R}^d}}\quad & \lambda\rho^2+\lambda\sum_k\widehat{p}_k y_k\\ \text{s.t.}\quad & \begin{pmatrix}(\lambda-1)I-W & -(\lambda-1)\widehat{x}_k+B\mu_k-u\\ -(\lambda-1)\widehat{x}_k^\top+\mu_k^\top B^\top-u^\top & (\lambda-1)(y_k+|\widehat{x}_k|^2)-t+2\delta^\top\mu_k-|\mu_k|^2-\sigma_k\end{pmatrix}\succeq O\quad\forall k\in[K],\end{aligned}$$

$$\begin{pmatrix} W & B & u \\ B^\top & I & \delta \\ u^\top & \delta^\top & t \end{pmatrix} \succeq O.$$

*Here* $\mu_k = \mathbb{E}[\widehat{Z} \mid \widehat{X} = \widehat{x}_k]$ *and* $\sigma_k = \mathbb{V}\mathrm{ar}[\widehat{Z} \mid \widehat{X} = \widehat{x}_k]$ *represent the conditional mean and the conditional variance.* ♣

### 4.2. Optimizing over RKHS Decision Rules

Restricting decision rules to affine functions of the contexts, as in Section 4.1, can be suboptimal. To address this limit, in this subsection, we consider the decision rule space $\mathcal{F}$ to be a vector-valued reproducing kernel Hilbert space (Carmeli et al., 2010). This class of policies can be interpreted as affine functions applied to nonlinear transformations of the contexts, where the nonlinear transformations are specified implicitly through a kernel function. The kernel function measures similarity between contexts and is often more interpretable than explicitly specifying the nonlinear transformation itself. Moreover, when the kernel function is universal, the corresponding RKHS policies are asymptotically optimal.

Specifically, let $\mathcal{F} = \mathbb{H} \subset \{\mathbb{h} : \mathcal{X} \to \mathbb{R}^m\}$ be a vector-valued reproducing kernel Hilbert space (RKHS) equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and a kernel $\boldsymbol{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$, such that the following reproducing property holds: the $j$-th component of $\mathbb{h}$ can be expressed by

$$\mathbb{h}_j(x) = \langle \mathbb{h}, \boldsymbol{K}(\cdot, x) e_j \rangle_{\mathbb{H}}, \qquad \forall j \in [m],\ x \in \mathcal{X},$$

where $e_j$ is the $j$-th unit vector of $\mathbb{R}^m$. Define a map $\varphi : \mathcal{X} \to \mathbb{H}^m$ via $\varphi_j(x) := \boldsymbol{K}(\cdot, x) e_j$, $j \in [m]$, which lifts the state space $\mathcal{X}$ to the feature space $\mathbb{H}^m$. Under this definition, the above expression can be written as

$$\mathbb{h}(x) = \langle \mathbb{h}, \varphi(x) \rangle_{\mathbb{H}} := \{\langle \mathbb{h}, \varphi_j(x) \rangle_{\mathbb{H}}\}_{j \in [m]},$$

and the true loss of a policy $\mathbb{h}$ can be rewritten as

$$\begin{aligned} \mathbb{E}_{(X,Z) \sim \mathbb{P}_{\mathsf{true}}}[\Psi(\mathbb{h}(X), Z)] &= \mathbb{E}_{(X,Z) \sim \mathbb{P}_{\mathsf{true}}}[\Psi(\langle \mathbb{h}, \varphi(X) \rangle_{\mathbb{H}}, Z)] \\ &= \mathbb{E}_{(\Phi,Z) \sim \mathbb{Q}_{\mathsf{true}}}[\Psi(\langle \mathbb{h}, \Phi \rangle_{\mathbb{H}}, Z)], \end{aligned}$$

where $\mathbb{Q}_{\mathsf{true}} = (\varphi \oplus \mathrm{id}_{\mathcal{Z}})_{\#}\mathbb{P}_{\mathsf{true}} \in \mathcal{P}(\mathbb{H}^m \times \mathcal{Z})$ is the true joint distribution between the lifted feature $\Phi$ and the random parameter $Z$. Recall that $\#$ denotes the push-forward of a measure, and the direct sum $\varphi \oplus \mathrm{id}_{\mathcal{Z}} : \mathcal{X} \times \mathcal{Z} \to \mathbb{H}^m \times \mathcal{Z}$ is defined as $(\varphi \oplus \mathrm{id}_{\mathcal{Z}})(x, z) := (\varphi(x), z)$.

We now define the causal transport robust formulation, following a way similar to the RKHS-based 1-Wasserstein uncertainty sets proposed in Shafieezadeh-Abadeh et al. (2019). Define the nominal distribution as $\widehat{\mathbb{Q}} := (\varphi \oplus \mathrm{id}_{\mathcal{Z}})_{\#}\widehat{\mathbb{P}}$, representing the joint nominal distribution of the feature $\Phi$ and the random parameter $Z$. For $\rho > 0$, define the 1-causal transport uncertainty set on the space of feature distributions:

$$\mathfrak{M} = \left\{\mathbb{Q} \in \mathcal{P}(\mathbb{H}^m \times \mathcal{Z}) : \mathsf{C}_1(\widehat{\mathbb{Q}}, \mathbb{Q}) \le \rho\right\}.$$

We consider the following problem

$$v_{\mathsf{P}} = \inf_{\mathbb{h} \in \mathbb{H}} \max_{\mathbb{Q} \in \mathfrak{M}} \mathbb{E}_{(\Phi, Z) \sim \mathbb{Q}}[\Psi(\langle \mathbb{h}, \Phi \rangle_{\mathbb{H}}, Z)]. \tag{9}$$

Note that both the inf and max problems are infinite-dimensional, as they are optimizing over functions and probability distributions, respectively.

We derive the following finite-dimensional convex reformulation of (9). For simplicity, we equip $\mathcal{Z}$ with infinity metric $d(z, z') = \infty \mathbf{1}\{z \neq z'\}$.

Corollary 4. *Suppose $\Psi(\cdot, z)$ is convex and Lipschitz in the decision variable. Define a constant*

$$L := \sup_{\widehat{x} \in \operatorname{supp} \widehat{\mathbb{P}}_{\widehat{X}}} \left\| \mathbb{E}_{\widehat{Z} \sim \widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[ \Psi(\cdot, \widehat{Z}) \mid \widehat{X} = \widehat{x} \right] \right\|_{\mathrm{Lip}}.$$

*Then $v_{\mathsf{P}}$ equals*

$$v_{\mathsf{D}} = \min_{\{\beta_k\}_k \subset \mathbb{R}^m} \left\{ \mathbb{E}_{(\widehat{X}, \widehat{Z}) \sim \widehat{\mathbb{P}}} \left[ \Psi\left( \sum_{k \in [K]} \boldsymbol{K}(\widehat{X}, \widehat{x}_k)\beta_k, \widehat{Z} \right) \right] + \rho L \left( \sum_{j,k \in [K]} \beta_j^\top \boldsymbol{K}(\widehat{x}_j, \widehat{x}_k)\beta_k \right)^{\frac{1}{2}} \right\}. \tag{10}$$

*Moreover, the optimal $\mathbb{h}$ that solves (9) is given by*

$$\mathbb{h}^\star = \sum_{k \in [K]} \boldsymbol{K}(\cdot, \widehat{x}_k)\beta_k^\star,$$

*where $\beta^\star$ is the minimizer of (10).*

This result demonstrates the tractability of the robust formulation (9). A key computational advantage is that it reduces the infinite-dimensional minimax problem to a minimization over only $mK$ real variables, with the robust loss naturally serving as a norm regularization. The proof is provided in Appendix EC.4. At a high level, the argument builds on the proof of Shafieezadeh-Abadeh et al. (2019, Theorem 28), which applies the Representer Theorem for RKHS, and extends the analysis to vector-valued functions and the causal transport distance.

### 4.3. Optimizing over All (Non-parametric) Decision Rules

In this subsection, we consider $\mathcal{F}$ to be unrestricted and contain all measurable functions $\{f : \mathcal{X} \to \mathcal{D}\}$. In general, this infinite-dimensional problem is hard to solve. Nonetheless, below, we provide a tractable way to find the optimal decision rule for this problem in certain settings.

Recall that our dual reformulation in Theorem 1 states that

$$v_{\mathsf{D}} = \min_{f:\mathcal{X}\to\mathcal{D}} \min_{\lambda\geq 0} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \sup_{x\in\mathcal{X}} \left\{ \varphi(f(x);\lambda,\widehat{X}) - \lambda\|x-\widehat{X}\| \right\} \right] \right\}, \tag{11}$$

where $\varphi(w;\lambda,\widehat{x}) := \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[ \sup_{z\in\mathcal{Z}} \left\{ \Psi(w,z) - \lambda\|z-\widehat{Z}\| \right\} \mid \widehat{X} = \widehat{x} \right]$. By replacing $\mathcal{X}$ with $\operatorname{supp}\widehat{\mathbb{P}}$, we define the in-sample dual problem as

$$v_{\widehat{\mathsf{D}}} := \min_{\substack{f:\mathcal{X}\to\mathcal{D} \\ \lambda\geq 0}} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \max_{1\leq k\leq K} \left\{ \varphi(f(\widehat{x}_k);\lambda,\widehat{X}) - \lambda\|\widehat{x}_k-\widehat{X}\| \right\} \right] \right\} \tag{12}$$

$$= \min_{\substack{\widehat{f}\in\widehat{\mathcal{F}} \\ \lambda\geq 0}} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \max_{1\leq k\leq K} \left\{ \varphi(\widehat{f}(\widehat{x}_k);\lambda,\widehat{X}) - \lambda\|\widehat{x}_k-\widehat{X}\| \right\} \right] \right\}, \tag{13}$$

where the second equality holds because the objective value in (12) depends only on the value of $f$ on $\operatorname{supp}\widehat{\mathbb{P}}$. Note that (13) is a finite-dimensional convex optimization problem with $K+1$ decision variables in the outer minimization.

THEOREM 3. *Suppose $p=1$, $\mathcal{D}\subset\mathbb{R}$ is convex, and $\Psi(w,z)$ is convex in $w$. Let $(\lambda^*,\widehat{f}^*)$ be a minimizer to the in-sample dual problem (13). Denote $\varphi_k(w) := \varphi(w;\lambda^*,\widehat{x}_k)$, $w_k := \widehat{f}^*(\widehat{x}_k)$, and $\phi_k := \max_j\{\varphi_k(w_j) - \lambda^*\|\widehat{x}_k-\widehat{x}_j\|\}$. For $x\in\mathcal{X}$, define*

$$I_k(x) := \{w\in\mathcal{D} :\ \varphi_k(w) \leq \lambda^*\|x-\widehat{x}_k\| + \phi_k\}.$$

*Then the intersection of $I_k(x)$'s is nonempty, and every decision rule $f^*\in\mathcal{F}$ satisfying $f^*(x)\in\cap_k I_k(x)$ for all $x\in\mathcal{X}$ is a minimizer to (11). Moreover, let $(\lambda^*,f^*)$ be a minimizer*

*to the dual problem* (D)*, then* $(\lambda^*, \widehat{f}^*)$ *is a minimizer to* (13)*, and* $f^*(x) \in \cap_k I_k(x)$ *defined above.*

Theorem 3 shows that problems (11) and (13) share the same optimal dual variable $\lambda^*$, and to solve the infinite-dimensional optimization over decision rules (11), it suffices first to solve a finite-dimensional robust in-sample optimization (13) and then extend the robust optimal in-sample decision rule to $\mathcal{X} \setminus \operatorname{supp} \widehat{\mathbb{P}}$ such that it is optimal to the original problem. Note that once the in-sample problem (13) is solved, the values $w_k$, $\phi_k$ are immediately available, and the set $I_k$ is defined precisely. There may be more than one way to extend the in-sample robust optimal decision rule $\widehat{f}$ to the entire space, as long as it belongs to the range of $\cap_k I_k(x)$.

*Proof Sketch.* The idea behind proving Theorem 3 is as follows. To show the optimality of the decision rules that lie within the intersection $\cap_k I_k$, the key step is to show $v_{\mathsf{D}} = v_{\widehat{\mathsf{D}}}$. Observe that $v_{\mathsf{D}} \geq v_{\widehat{\mathsf{D}}}$, since the inner supremum in (11) is taken with respect to a larger set compared with the maximization in (12). To see the other direction, the main step is to show that $I_k(x)$ has a nonempty intersection. Once this is shown, it is easy to verify by simple algebra that $f^*(x) \in \cap_k I_k(x)$ attains the value $v_{\widehat{\mathsf{D}}}$, thereby $v_{\mathsf{D}}$ is dominated by the objective value of $f^*$ which equals $v_{\widehat{\mathsf{D}}}$. Thus we have $v_{\mathsf{D}} = v_{\widehat{\mathsf{D}}}$. To show $I_k(x)$ has a nonempty intersection, since they are one-dimensional intervals, it suffices to show they pairwise intersect. This can be established using the convexity of $\varphi$. The necessity of the above interval condition, i.e., for any optimal policy $f^*$, $f^*(x) \in \cap_k I_k(x)$, could be justified by contradiction. The detailed proof can be found in Appendix EC.4. □

Remark 2. The one-dimensional assumption of the decision space $\mathcal{D}$ is crucial. The following example shows that it is impossible to use the form in Theorem 3 to construct an optimal policy as soon as the decision dimension becomes two. Specifically, let us consider a two-dimensional decision space $\mathcal{D}$ and a linear loss function $\Psi(w, z) = w^\top z$, where $\mathcal{D} = B_1$ is the unit disk in $\mathbb{R}^2$, and $\mathcal{X} = \mathcal{Z} = \mathbb{R}^2$. Suppose $\widehat{\mathbb{P}}$ satisfies

$$\widehat{\mathbb{P}}_{\widehat{X}} = \frac{1}{3}\sum_{k=1}^{3} \delta_{\widehat{x}_k}, \qquad \widehat{x}_k = \left(\cos\frac{2k\pi}{3}, \sin\frac{2k\pi}{3}\right), \qquad \mathbb{E}\left[\widehat{Z} \mid \widehat{X}\right] = \frac{2\sqrt{3}}{3}\widehat{X},$$

where $\widehat{X} = \{\widehat{x}_1, \widehat{x}_2, \widehat{x}_3\}$ consists of three points on the unit circle that form an equilateral triangle (see Figure 5). We can show that the intersection of $I_k(x)$ defined in Theorem 3

is empty, although the in-sample dual problem is explicitly solvable. Detailed calculations are available in Appendix EC.4.

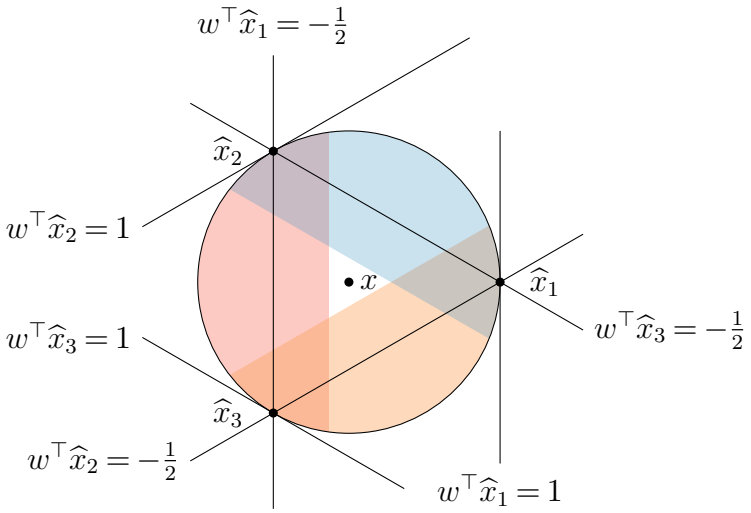


**Figure 5** **Intersection of $I_k(x)$ when $x = 0$ is empty. $I_1(x)$ is the red region, $I_2(x)$ is the orange region, and $I_3(x)$ is the blue region.**

Remark 3 (Comparison with the Shapley Policy in Zhang et al. (2024)). In Zhang et al. (2024), the authors study (3) with Wasserstein uncertainty sets, focusing on the newsvendor cost. They show that when optimization over all decision rules, the optimal decision rule, called Shapley policy, can be found by first solving for the in-sample Wasserstein robust optimal decision rule $\widehat{f}_{\mathsf{W}}$, then extending to the entire space by solving

$$f_{\mathsf{W}}(x) \in \arg\min_{w\in\mathbb{R}} \max_k \frac{|w - \widehat{f}_{\mathsf{W}}(\widehat{x}_k)|}{\|x - \widehat{x}_k\|}, \tag{$\mathsf{W}_{\mathsf{Lip}}$}$$

which minimizes the maximal slope. Using the same idea, if we define

$$f_\infty(x) \in \arg\min_{w\in\mathbb{R}} \max_k \frac{|w - \widehat{f}^*(\widehat{x}_k)|}{\|x - \widehat{x}_k\|}, \tag{$\mathsf{C}_{\mathsf{Lip}}$}$$

where $\widehat{f}^*(\widehat{x}_k)$'s are defined in Theorem 3, then it can be verified that $f_\infty(x) \in \cap_k I_k(x)$. Therefore, this shows that $f_\infty(x)$ is a robust optimal decision rule for (11). Note that we use the subscript $\infty$ to indicate the $\infty$-norm (maximum) of the slope function $k \mapsto \frac{|w-\widehat{f}^*(\widehat{x}_k)|}{\|x-\widehat{x}_k\|}$.

Differently, we can define another decision rule that minimizes the 1-norm of the slope function,

$$f_1(x) \in \arg\min_{w\in\mathbb{R}} \sum_k \frac{|w - \widehat{f}^*(\widehat{x}_k)|}{\|x - \widehat{x}_k\|}. \tag{$\mathsf{C}_{\mathsf{TV}}$}$$

The resulting decision rule may not be optimal, but we can always truncate its values to ensure they fall within $\cap_k I_k(x)$ and thereby make it robustly optimal. Namely, if we use $\overline{I}(\cdot)$ and $\underline{I}(\cdot)$ to represent the upper and lower bound of the region $\cap_k I_k(x)$, then we define

$$\bar{f}_1(x) := \max\Big(\underline{I}(x),\ \min\big(f_1(x), \overline{I}(x)\big)\Big). \qquad (\mathsf{C}_{\mathsf{TV\text{-}trunc}})$$

We denote the truncated decision rule as $\bar{f}_1(x)$.

We illustrate the two robust optimal decision rules defined above using a conditional median estimate problem with $Z = \mu(X) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, $\mu(x) = \sin(2x) + 2\exp(-16x^2)$, as demonstrated in Figure 6. Based on the blue training data, one seeks to learn the conditional median expressed by the dashed line. After solving the in-sample causal problem, $\mathsf{C}_{\mathsf{Lip}}$ and $\mathsf{C}_{\mathsf{TV}}$ extend the same in-sample policy to different decision rules $f_\infty$ and $f_1$. $f_\infty$ is always in the optimal region, but $f_1$ needs to be clipped to fit in the optimal region.

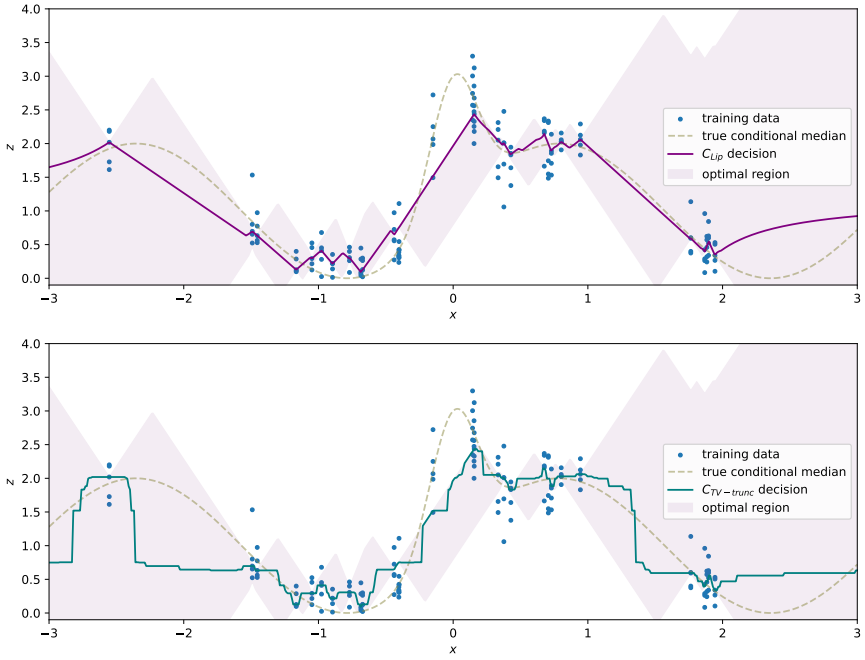


**Figure 6** **Two robust optimal decision rules $f_\infty$ and $\bar{f}_1$ of a median estimation problem. The horizontal axis represents $x$, and the vertical axis represents $z$. The dashed line represents the true conditional median $z = \mu(x)$.**

EXAMPLE 5 (REVISITED). Consider the feature-based newsvendor problem in Example 5. When $h = b = 1$, this is equivalent to conditional median estimation. As detailed in

Appendix EC.5, the in-sample dual problem (13) can be transformed into a linear programming problem

$$\begin{aligned}\inf_{\substack{\{w_k\}_k,\{y_k\}_k\subset\mathbb{R}\\ \{c_{kji}\}_{kji}\subset\mathbb{R},\lambda\geq 1}} \quad & \lambda\rho+\sum_{k=1}^{K} y_k \\ \text{s.t.} \quad & y_j \geq \sum_{i=1}^{n_j} \widehat{p}_{ki}(c_{kji}-\lambda\|\widehat{x}_k-\widehat{x}_j\|) && \forall j,k\in[K],\\ & c_{kji}\geq w_k-\widehat{z}_{ji} && \forall k,j\in[K], i\in[n_j],\\ & c_{kji}\geq \widehat{z}_{ji}-w_k && \forall k,j\in[K], i\in[n_j].\end{aligned}$$

This is a linear programming with $K(n+2)+1$ variables and $K(2n+K)+1$ constraints. ♣

Example 6 (revisited). Consider the personalized pricing problem in Example 6. By Theorem 1, its strong dual problem can be written as

$$\inf_{\substack{f:\mathcal{X}\to\mathbb{R}\\ \lambda\geq 0}}\left\{\lambda\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{-f(x)z^\top\begin{pmatrix}f(x)\\1\end{pmatrix}-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\lambda\|x-\widehat{X}\|^p\right\}\right]\right\}.$$

In the case of $p=1$, we notice that $f$ is real-valued and $\Psi$ is convex in $w$, so we may use Theorem 3 to reformulate the problem as

$$\inf_{\substack{\widehat{f}:\widehat{\mathcal{X}}\to\mathbb{R}\\ \lambda\geq 0}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq K}\left\{\varphi(\widehat{f}(\widehat{x}_k);\lambda,\widehat{X})-\lambda\|\widehat{x}_k-\widehat{X}\|\right\}\right]\right\}.$$

where

$$\varphi(w;\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{-wz^\top\begin{pmatrix}w\\1\end{pmatrix}-\lambda\|z-\widehat{Z}\|\right\}\mid\widehat{X}=\widehat{x}\right].$$

In particular, it can be reformulated as the following

$$\begin{aligned}\inf_{\{w_k\}_k,\{c_k\}_k,\lambda\geq 0} \quad & \lambda\rho+\sum_{k\in[K]}\widehat{p}_k c_k \\ \text{s.t.} \quad & c_j+\begin{pmatrix}w_k^2 & w_k\end{pmatrix}\bar{z}_k+\lambda\|\widehat{x}_k-\widehat{x}_j\|\geq 0 && \forall j,k\in[K],\\ & \|(w_k^2\ \ w_k)\|_*\leq\lambda && \forall k\in[K].\end{aligned} \tag{14}$$

where $\widehat{p}_k=\sum_{i=1}^{n_k}\widehat{p}_{ki}$ and $\bar{z}_k=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z}\mid\widehat{X}=\widehat{x}_k]$. Here $\|\cdot\|_*$ is the dual norm of $\|\cdot\|_{\mathcal{Z}}$. When $\mathcal{Z}=\mathbb{R}^2$ is equipped with $\ell^1$ or $\ell^\infty$ norm, (14) can be reduced to a quadratic program, whereas when the $\ell^2$ norm is chosen, (14) can be written as a second-order conic program. A detailed calculation can be found in Appendix EC.5. ♣

# 5. Numerical Experiments

## 5.1. Feature-based Newsvendor

**5.1.1. Synthetic Data** In this section, we illustrate our proposed approach in the context of a feature-based newsvendor. We consider a similar setup as in Zhang et al. (2024), where the demand $Z$ depends on $X$ in a nonlinear way:

$$Z = f(\beta^\top X) + \varepsilon, \qquad f(\lambda) := c[\sin(2\lambda) + 2\exp(-16\lambda^2) + 1],$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is a standard Gaussian variable independent from $\beta$ and $X$. Let the coefficient vector $\beta \in \mathbb{R}^{100}$, with each component independently sampled from a uniform distribution $\mathcal{U}([-0.1, 0.1])$. The covariate $X$ is sampled from a 100-dimensional multivariate normal distribution $\mathcal{N}(0, (\sigma_{ij})_{ij})$, with mean zero and covariate matrix defined by $\sigma_{ij} = 0.5^{|i-j|}$ with $i, j = 1, \ldots, 100$. The constant $c = 1.7$ is chosen such that the signal-to-noise ratio is approximately 3:1. Since the demand should be positive, we reject all samples with $Z < 0$.

We experiment with different unit overage cost $h \in \{0.2, 0.5, 0.8, 1\}$ while fixing the unit underage cost $b = 1$. To understand the effect of the sample size, we choose $K \in \{10, 30, 100, 300\}$ and $n_k \in \{1, 3, 10, 30, 100\}$. The testing data size is 10000. The hyper-parameters are tuned based on 5-fold cross-validation. We set $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_2$ and $\|\cdot\|_{\mathcal{Z}} = \infty \cdot \mathbf{1}\{z \neq \widehat{z}\}$. To generate the boxplots, we run 20 repeated experiments (except for $K = 10$, we run 50 experiments to get a more accurate depiction). All experiments are performed in Ubuntu 18.04 using Python 3.6.9 with a convex optimization solver Gurobi 9.1.1, on a Dell Precision 5820 Tower Workstation with Intel® Xeon® W-2125 CPU (32 cores) and 32GB RAM (DDR4 2666MHz). Due to constraints associated with the solver's capabilities, the experiments with $n_k = 30$, $K = 300$ and $n_k = 100$, $K = 100, 300$ are not included in the comparison.

In our first set of experiments, we delve into the effects of adopting different distributional uncertainty sets of the inner worst-case expectation, namely, the Wasserstein DRO with Shapley extension ($\mathsf{W}_{\mathsf{Lip}}$) in Zhang et al. (2024) versus causal transport DRO with Shapley extension ($\mathsf{C}_{\mathsf{Lip}}$). Both approaches incorporate the Shapley extension to extend the in-sample optimal policy. Figure 7 shows the relative difference in the out-of-sample expected cost between $\mathsf{C}_{\mathsf{Lip}}$ and $\mathsf{W}_{\mathsf{Lip}}$ with the same training and testing data set—a negative number indicates that $\mathsf{C}_{\mathsf{Lip}}$ outperforms $\mathsf{W}_{\mathsf{Lip}}$.

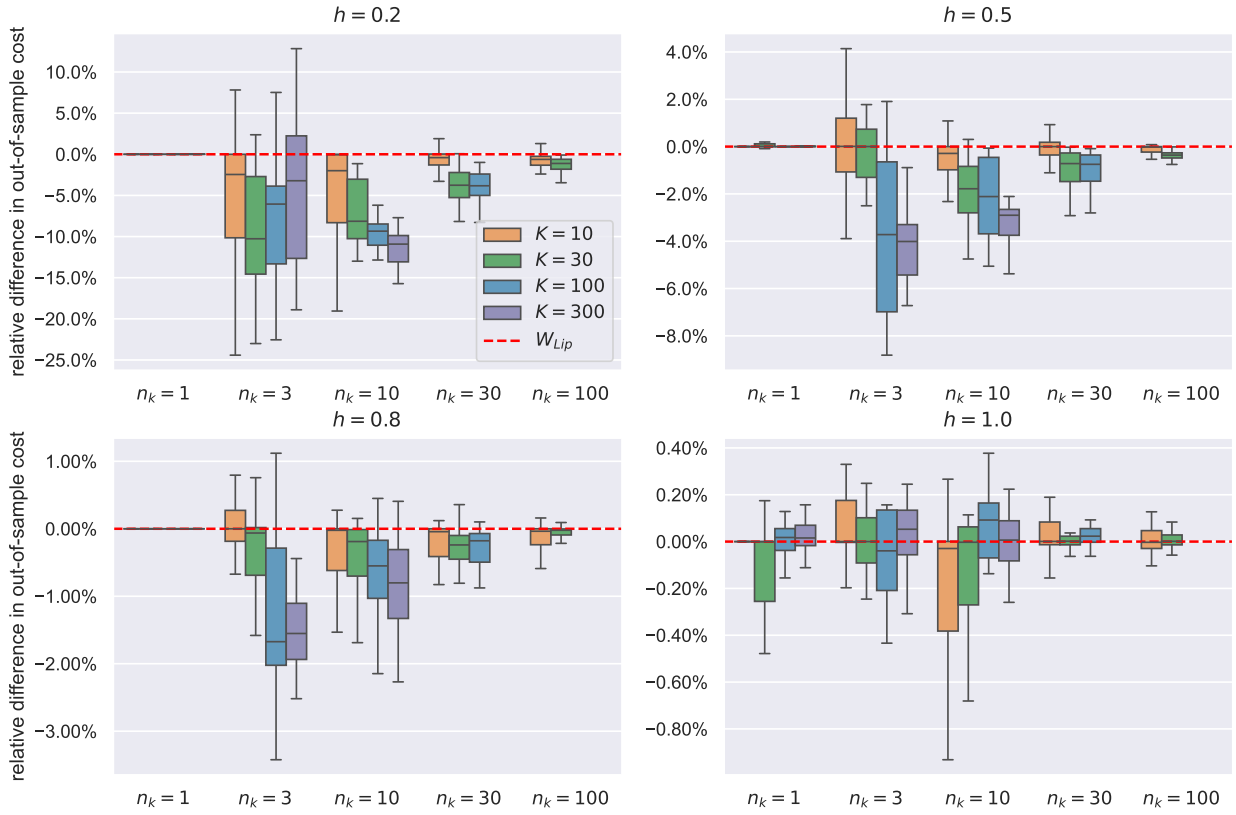


**Figure 7** **Boxplots of the relative differences in the out-of-sample performance between $\mathsf{W}_{\mathsf{Lip}}$ (baseline) and $\mathsf{C}_{\mathsf{Lip}}$**

We have the following observations.

(I) When each covariate group contains only a single sample ($n_k = 1$), $\mathsf{C}_{\mathsf{Lip}}$ and $\mathsf{W}_{\mathsf{Lip}}$ have the same performance because the two formulations are equivalent (Remark 1).

(II) As the sample size per covariate group increases beyond a single sample, $\mathsf{C}_{\mathsf{Lip}}$ begins to exhibit a performance advantage over $\mathsf{W}_{\mathsf{Lip}}$, particularly when dealing with skewed loss functions ($h = 0.2, 0.5, 0.8$). This edge is most pronounced at lower sample sizes $n_k = 3, 10$, which shows the value of (even a little) conditional information. The marginal benefit provided by $\mathsf{C}_{\mathsf{Lip}}$ tends to diminish with larger sample sizes per covariate group ($n_k = 30, 100$). One explanation is that the worst-case distribution of $\mathsf{W}_{\mathsf{Lip}}$ does not significantly degrade the conditional information structure when there are many samples at the same covariate value.

(III) The comparative advantage of $\mathsf{C}_{\mathsf{Lip}}$ over $\mathsf{W}_{\mathsf{Lip}}$ generally amplifies with the increase in the number of covariate groups $K$. An explanation is that when $K$ is large, $\mathsf{C}_{\mathsf{Lip}}$ can fully take advantage of the conditional information to extrapolate other conditional distributions.

In addition to our main comparisons above, we further investigate two supplementary scenarios. The first one compares the Shapley extension versus non-Shapley extensions in the nonparametric setting, as discussed in Theorem 3 and Remark 3. The second compares truncated versus non-truncated causal transport DRO solutions, as identified in Theorem 3. We refer the two additional experiments in Appendix EC.6.

**5.1.2. Real Data** We use a real-world dataset from a meal delivery company (Asuncion et al., 2007; Qi et al., 2024), which records 145 weeks of historical weekly demand. The company fulfills customer meal orders using perishable ingredients, making its weekly procurement planning naturally fit into a newsvendor framework. In addition to demand data, the dataset includes contextual features such as checkout price, promotion email indicators, and whether the meal was featured on the company's homepage. We compare the empirical performance of our proposed $\mathsf{C}_{\mathsf{Lip}}$ against several benchmarks, including: empirical risk minimization using affine policy with $\ell^1$ and $\ell^2$ regularization ($\mathsf{ERM2}\,(\ell^1/\ell^2)$) and kernel-weights optimization ($\mathsf{KO}$) in Ban and Rudin (2019); conditional stochastic optimization using random forests ($\mathsf{RandForest}$) and $k$-nearest neighbors ($k\mathsf{NN}$) in Bertsimas and Kallus (2020); and stochastic optimization forests with different splitting criteria ($\mathsf{StochOptForest}\,(\mathsf{apx\text{-}soln}/\mathsf{apx\text{-}risk})$) in Kallus and Mao (2023).

We train all models on the most recent one year (52 weeks) of data using a rolling window, then test them out-of-sample on subsequent periods with planning horizons of 1, 4, and 12 weeks—corresponding to weekly, monthly, and quarterly re-optimization. In Figure 8, we compare differences in testing cost across competing methods under three levels of $h = 0.1, 0.25, 0.5$—which are chosen to reflect the practical scenarios—and three planning horizons. The y-axis represents the cost difference relative to our $\mathsf{C}_{\mathsf{Lip}}$ (red dashed line at zero), with positive values indicating higher cost (worse performance) compared to $\mathsf{C}_{\mathsf{Lip}}$.

From Figure 8, we see that $\mathsf{C}_{\mathsf{Lip}}$ achieves consistently strong performance across all nine configurations (three $h$ values $\times$ three planning horizons), both in terms of mean performance (white circle) and median performance (black line). For small $h$, $\mathsf{ERM2}\,(\ell^1/\ell^2)$ sometimes attains better performance, though without consistency. In the weekly planning horizon, forest-based methods ($\mathsf{StochOptForest}\,(\mathsf{apx\text{-}soln}/\mathsf{apx\text{-}risk})$, $\mathsf{RandForest}$) sometimes obtain lower median costs than $\mathsf{C}_{\mathsf{Lip}}$ at $h = 0.1$ and 0.25, but this comes with much higher

variability and worse mean performance. Overall, the results indicate that $\mathsf{C}_{\mathsf{Lip}}$ consistently delivers strong, stable out-of-sample performance.

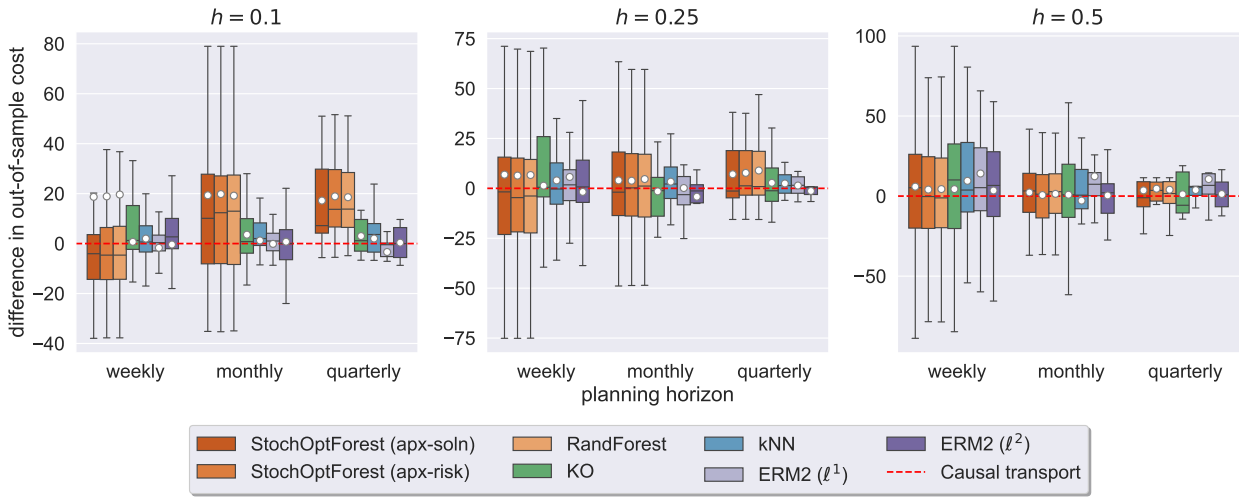


**Figure 8** **Boxplots of the relative differences in the out-of-sample performance between $\mathsf{C}_{\mathsf{Lip}}$ and other benchmarks**

### 5.2. Contextual Linear Optimization

This section presents numerical experiments on the contextual linear optimization problem from Example 7, examined in the setting of portfolio optimization. The goal is to determine the optimal weights $w \in \mathbb{R}^m$ across $m$ assets in order to maximize the portfolio return $w^\top Z$, where $Z \in \mathbb{R}^m$ denotes the random return. To align with the setting in (1), we recast it as a minimization problem by defining the loss function $\Psi(w,z) = -w^\top z$, i.e., the negative of the portfolio return. The allocation weights are subject to three constraints. First, the budget constraint requires $w^\top \mathbf{1} = 1$. Second, short-selling is permitted but limited, in the sense that no individual position may be shorter than 30%. Finally, the $m = 10$ assets are grouped into three sectors– $\{1,2,3\}$, $\{4,5,6,7\}$, $\{8,9,10\}$–and the average weight within each group must be nonnegative, preventing systematic shorting of an entire sector.

The contexts are uniformly sampled from the ellipsoid $\mathcal{X} = \{x \in \mathbb{R}^d : x^\top \Sigma x \le 1\}$, with $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = 0.5^{|i-j|}$ with $i,j = 1,\ldots,d$, where $d = 5$. Given a context $X$, the random variable $Z \in \mathbb{R}^m$ is generated according to the following nonlinear model inspired by Feng et al. (2024):

$$Z_i = \sum_{j=1}^{d} f^{(\beta_{ij})}(X_j) + \sigma\left(\sum_{j=1}^{d} f^{(4)}(X_j) + 1\right)\cdot \varepsilon_i, \qquad i \in [m].$$

Here, the nonlinear functions are defined as

$$f^{(1)}(x) = \sin(2\pi x), \quad f^{(2)}(x) = 0.5 \exp x, \quad f^{(3)}(x) = 1.5|(x-0.4)(x-0.6)|, \quad f^{(4)}(x) = x\cos(2\pi x).$$

The indices $\beta_{ij}$ are independently sampled from $\{1,2,3\}$ at the beginning of each experiment and remain unchanged within that experiment. The noise scaling factor $\sigma \in \{1,3,10\}$ controls the signal-to-noise ratio. The noise $\varepsilon_i$ follows a $t$-distribution with 3 degrees of freedom. We generate $K$ context samples of $X$. For each sample $\widehat{x}_k$, we then generate $n_k$ samples of $Z$, resulting $n = \sum_{k\in[K]} n_k$ data pairs $(\widehat{x}_k, \widehat{z}_{ki})$.

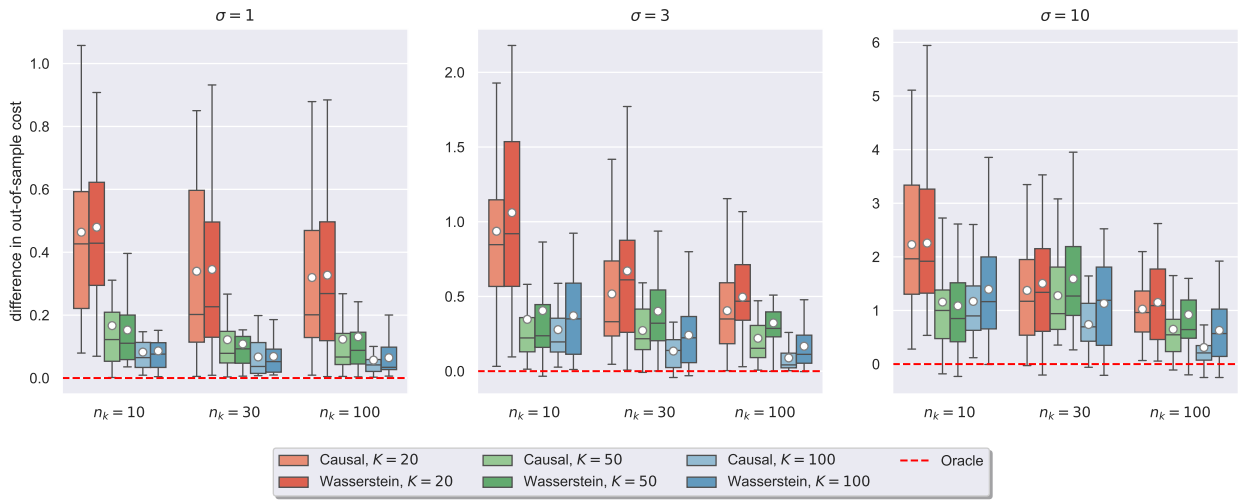


**Figure 9** **Experiment with fixed number $n_k$ of $z$ per $x$ chosen in $\{10,30,100\}$, with noise level $\sigma \in \{1,3,10\}$ and number of contexts $K \in \{20,50,100\}$. They are compared with a baseline scenario which applies** ERM1 **algorithm on a much larger training data set.**

To further investigate the comparison between Wasserstein and causal transport uncertainty sets with affine policy class, we vary the number of context samples $K \in \{20,50,100\}$, the number of outcome samples per context $n_k \in \{10,30,100\}$, and the noise level $\sigma \in \{1,3,10\}$. Figure 9 reports boxplots of the out-of-sample cost of W-DRO and C-DRO, compared with an oracle model using ERM1 algorithm introduced in Ban and Rudin (2019) on a training data set of $10^7$ samples. Across nearly all settings, C-DRO outperforms W-DRO in terms of median cost, with the performance gap widening as the noise level increases. At a low noise level, the gap narrows as the number of contexts increases. Consistent with the newsvendor experiments in Section 5.1.1, when a moderate amount of conditional information is available, C-DRO achieves a clear advantage over W-DRO. In contrast, when

conditional information is either scarce or overly abundant, the difference between the two methods diminishes. This is likely because, in the former case, C-DRO cannot exploit the limited conditional information, while in the latter case, the worst-case distribution of W-DRO does not significantly distort the underlying conditional structure.

We also investigate the impact of the distribution shift. Fixing $\sigma = 1$, $n_k = 30$, we use the same data above to train the model, but the testing data distribution has a different conditional distribution $Z \mid X$ (known as concept shift):

$$Z_i^{\text{test}} = \sum_{j=1}^{d} f^{(\beta_{ij})}(X_j) + \sigma \left( \sum_{j=1}^{d} f^{(4)}(X_j) + 1 \right) \cdot \varepsilon_i + \sigma_{\text{shift}} \cdot \varepsilon_i^{\text{shift}}, \qquad i \in [m].$$

where $\sigma_{\text{shift}} \in \{1, 3, 10\}$ is a scaling factor of the distributional shift, $\varepsilon_i^{\text{shift}} \sim \mathcal{N}(\mu_i^{\text{shift}}, \sigma_i^{\text{shift}})$ are independent normal random variables, with context-dependent mean and variance given by

$$\mu_i^{\text{shift}} = \sum_{j=1}^{d} f^{(\beta_{ij}^{\text{mean}})}(X_j), \qquad \sigma_i^{\text{shift}} = \sum_{j=1}^{d} f^{(\beta_{ij}^{\text{var}})}(X_j),$$

where $\beta_{ij}^{\text{mean}}$ and $\beta_{ij}^{\text{var}}$ are uniformly sampled from $\{1, 2, 3\}$, independently from $\beta_{ij}$ and each other, but are fixed within each experiment.

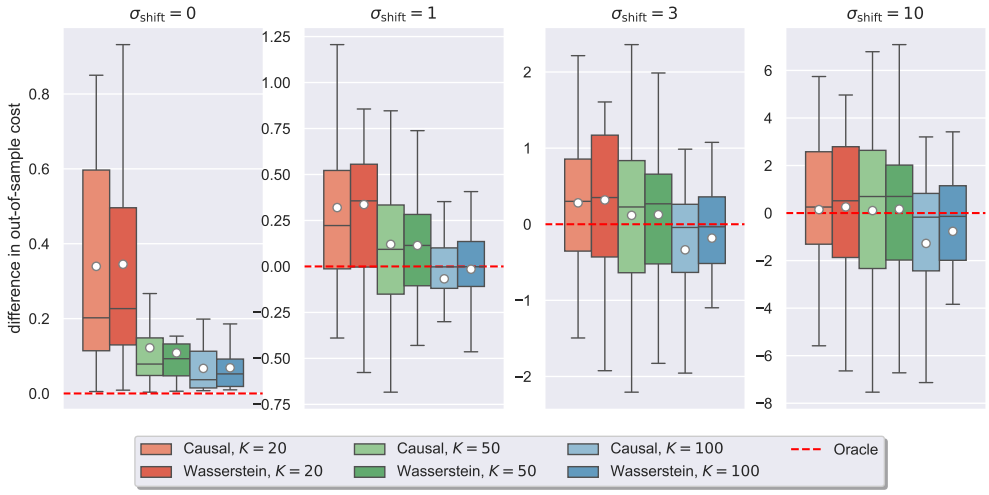


**Figure 10** **Experiment with concept shift, with shift level $\sigma_{\text{shift}} \in \{0, 1, 3, 10\}$ and number of contexts $K \in \{20, 50, 100\}$.**

Figure 10 reports boxplots of the out-of-sample cost of W-DRO and C-DRO, compared again to the baseline using ERM1 algorithm with $10^7$ samples. We observe that C-DRO

consistently outperforms W-DRO, especially with a large concept shift $\sigma_{\text{shift}}$ and a large number of contexts $K$. This demonstrates the improved robustness of C-DRO in hedging against shifts in the conditional distributions.

## 6. Concluding Remarks

In this paper, we propose a new distributionally robust decision rule optimization for decision-making with side information based on the causal transport distance. These results open up new research directions for distributionally robust optimization and adjustable robust optimization. For future work, it would be interesting to investigate the performance guarantees of the proposed framework.

## References

Acciaio B, Backhoff-Veraguas J, Carmona R (2019) Extended mean field control problems: stochastic maximum principle and transport perspective. *SIAM journal on Control and Optimization* 57(6):3666–3693.

Acciaio B, Backhoff-Veraguas J, Zalashko A (2020) Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *Stochastic Processes and their Applications* 130(5):2918–2953.

Analui B, Pflug GC (2014) On distributionally robust multiperiod stochastic optimization. *Computational Management Science* 11(3):197–220.

Asuncion A, Newman D, et al. (2007) UCI machine learning repository.

Backhoff J, Beiglbock M, Lin Y, Zalashko A (2017) Causal transport in discrete time and applications. *SIAM Journal on Optimization* 27(4):2528–2562.

Ban GY, Gallien J, Mersereau AJ (2019) Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing & Service Operations Management* 21(4):798–815.

Ban GY, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

Bartl D, Wiesel J (2023) Sensitivity of multiperiod optimization problems with respect to the adapted Wasserstein distance. *SIAM Journal on Financial Mathematics* 14(2):704–720.

Basciftci B, Ahmed S, Shen S (2021) Distributionally robust facility location problem under decision-dependent stochastic demand. *European Journal of Operational Research* 292(2):548–561.

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *INFORMS TutORials in Operations Research*, 1–19 (INFORMS).

Bazier-Matte T, Delage E (2020) Generalization bounds for regularized portfolio selection with market side information. *INFOR: Information Systems and Operational Research* 58(2):374–401.

Bertsimas D, Georghiou A (2015) Design of near optimal decision rules in multistage adaptive mixed-integer optimization. *Operations Research* 63(3):610–627.

Bertsimas D, Goyal V (2012) On the power and limitations of affine policies in two-stage adaptive optimization. *Mathematical programming* 134(2):491–531.

Bertsimas D, Iancu DA, Parrilo PA (2010) Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research* 35(2):363–394.

Bertsimas D, Iancu DA, Parrilo PA (2011) A hierarchy of near-optimal policies for multistage adaptive optimization. *IEEE Transactions on Automatic Control* 56(12):2809–2824.

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Science* 66(3):1025–1044.

Bertsimas D, Koduri N (2022) Data-driven optimization: A reproducing kernel hilbert space approach. *Operations Research* 70(1):454–471.

Bertsimas D, McCord C (2019) From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637* .

Bertsimas D, McCord C, Sturt B (2022) Dynamic optimization with side information. *European Journal of Operational Research* .

Bertsimas D, Van Parys B (2021) Bootstrap robust prescriptive analytics. *Mathematical Programming* 1–40.

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

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

Brandt MW, Santa-Clara P, Valkanov R (2009) Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies* 22(9):3411–3447.

Cao J, Gao R (2021) Contextual decision-making under parametric uncertainty and data-driven optimistic optimization. *Available at Optimization Online* .

Carmeli C, De Vito E, Toigo A, Umanitá V (2010) Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications* 8(01):19–61.

Chen X, Sim M, Sun P, Zhang J (2008) A linear decision-based approximation approach to stochastic programming. *Operations Research* 56(2):344–357.

Chenreddy AR, Bandi N, Delage E (2022) Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems* 35:9525–9537.

El Balghiti O, Elmachtoub AN, Grigas P, Tewari A (2019) Generalization bounds in the predict-then-optimize framework. *Advances in Neural Information Processing Systems* 32:14412–14421.

El Housni O, Goyal V (2021) On the optimality of affine policies for budgeted uncertainty sets. *Mathematics of Operations Research* 46(2):674–711.

Elmachtoub A, Liang JCN, McNellis R (2020) Decision trees for decision-making under the predict-then-optimize framework. *International Conference on Machine Learning*, 2858–2867 (PMLR).

Elmachtoub AN, Grigas P (2022) Smart "predict, then optimize". *Management Science* 68(1):9–26.

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

Esteban-Pérez A, Morales JM (2022) Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming* 195(1-2):1069–1105.

Estes A (2021) Slow rates of convergence in optimization with side information. *Available at SSRN 3803427* .

Feng X, He X, Jiao Y, Kang L, Wang C (2024) Deep nonparametric quantile regression under covariate shift. *Journal of Machine Learning Research* 25(385):1–50.

Gao R (2023) Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research* 71(6):2291–2306.

Gao R, Arora R, Huang Y (2024a) Data-driven multistage distributionally robust linear optimization with nested distance. *arXiv preprint arXiv:2407.16346* .

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

Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.

Georghiou A, Tsoukalas A, Wiesemann W (2025) On the optimality of affine decision rules in distributionally robust optimization. *Management Science* .

Hanasusanto GA, Kuhn D (2013) Robust data-driven dynamic programming. *Advances in Neural Information Processing Systems* 26.

Hanasusanto GA, Kuhn D, Wiesemann W (2015) K-adaptability in two-stage robust binary programming. *Operations Research* 63(4):877–891.

Hanasusanto GA, Kuhn D, Wiesemann W (2016) K-adaptability in two-stage distributionally robust binary programming. *Operations Research Letters* 44(1):6–11.

Hannah L, Powell W, Blei D (2010) Nonparametric density estimation for stochastic optimization with an observable state variable. *Advances in Neural Information Processing Systems* 23:820–828.

Ho-Nguyen N, Kılınç-Karzan F (2022) Risk guarantees for end-to-end prediction and optimization processes. *Management Science* 68(12):8680–8698.

Hu Y, Kallus N, Mao X (2022) Fast rates for contextual linear optimization. *Management Science* 68(6):4236–4245.

Hu Y, Wang J, Xie Y, Krause A, Kuhn D (2024) Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems* 36.

Iancu DA, Sharma M, Sviridenko M (2013) Supermodularity and affine policies in dynamic robust optimization. *Operations Research* 61(4):941–956.

Jean J (1980) Weak and strong solutions of stochastic differential equations. *Stochastics* 3(1-4):171–191.

Jiang Y (2024) Duality of causal distributionally robust optimization: the discrete-time case. *arXiv preprint arXiv:2401.16556* .

Kallus N, Mao X (2023) Stochastic optimization forests. *Management Science* 69(4):1975–1994.

Kannan R, Bayraksan G, Luedtke JR (2023) Residuals-based distributionally robust optimization with covariate information. *Mathematical Programming* 1–57.

Kannan R, Bayraksan G, Luedtke JR (2025) Data-driven sample average approximation with covariate information. *Operations Research* .

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*, 130–166 (INFORMS).

Kurtz T (2014) Weak and strong solutions of general stochastic models. *Electronic Communications in Probability* 19:1–16.

Lassalle R (2013) Causal transference plans and their monge-kantorovich problems. *arXiv preprint arXiv:1303.6925* .

Liu M, Qi M, Shen ZJM (2021) End-to-end deep learning for inventory management with fixed ordering cost and its theoretical analysis. *Available at SSRN 3888897* .

Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Operations Research Letters* 33(4):341–348.

Loke GG, Tang Q, Xiao Y (2022) Decision-driven regularization: A blended model for predict-then-optimize. *Available at SSRN 3623006* .

Muñoz MA, Pineda S, Morales JM (2022) A bilevel framework for decision-making under uncertainty with contextual information. *Omega* 108:102575.

Nguyen VA, Zhang F, Wang S, Blanchet J, Delage E, Ye Y (2025) Robustifying conditional portfolio decisions via optimal transport. *Operations Research* 73(5):2801–2829.

Oroojlooyjadid A, Snyder LV, Takáč M (2020) Applying deep learning to the newsvendor problem. *IISE Transactions* 52(4):444–463.

Perakis G, Sim M, Tang Q, Xiong P (2023) Robust pricing and production with information partitioning and adaptation. *Management Science* 69(3):1398–1419.

Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quantitative Finance* 7(4):435–442.

Pflug GC (2010) Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization* 20(3):1406–1420.

Pflug GC, Pichler A (2012) A distance for multistage stochastic optimization models. *SIAM Journal on Optimization* 22(1):1–23.

Pflug GC, Pichler A (2014) *Multistage stochastic optimization* (Springer).

Pflug GC, Pichler A (2015) Dynamic generation of scenario trees. *Computational Optimization and Applications* 62(3):641–668.

Pflug GC, Pichler A (2016) From empirical observations to tree models for stochastic optimization: convergence properties. *SIAM Journal on Optimization* 26(3):1715–1740.

Pichler A, Shapiro A (2021) Mathematical foundations of distributionally robust multistage optimization. *SIAM Journal on Optimization* 31(4):3044–3067.

Pólik I, Terlaky T (2007) A survey of the s-lemma. *SIAM review* 49(3):371–418.

Postek K, Hertog Dd (2016) Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. *INFORMS Journal on Computing* 28(3):553–574.

Qi M, Grigas P, Shen ZJ (2025) Integrated conditional estimation-optimization. *Operations Research* .

Qi M, Shen ZJ (2022) Integrating prediction/estimation and optimization with applications in operations management. *Tutorials in operations research: emerging and impactful topics in operations*, 36–58 (INFORMS).

Qi M, Shen ZJ, Zheng Z (2024) Learning newsvendor problems with intertemporal dependence and moderate non-stationarities. *Production and Operations Management* 33(5):1196–1213.

Qi M, Shi Y, Qi Y, Ma C, Yuan R, Wu D, Shen ZJ (2023) A practical end-to-end inventory management model with deep learning. *Management Science* 69(2):759–773.

Rahimian H, Bayraksan G, Homem-de Mello T (2019) Controlling risk and demand ambiguity in newsvendor models. *European Journal of Operational Research* 279(3):854–868.

Rüschendorf L (1985) The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields* 70(1):117–129.

Rychener Y, Kuhn D, Sutter T (2023) End-to-end learning for stochastic optimization: A bayesian perspective. *International Conference on Machine Learning*, 29455–29472 (PMLR).

Sadana U, Chenreddy A, Delage E, Forel A, Frejinger E, Vidal T (2024) A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research* .

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

Shapiro A, Dentcheva D, Ruszczyński A (2014) *Lectures on stochastic programming: modeling and theory* (SIAM).

Shen Y, Xu P, Zavlanos M (2024) Wasserstein distributionally robust policy evaluation and learning for contextual bandits. *Transactions on Machine Learning Research* ISSN 2835-8856.

Srivastava PR, Wang Y, Hanasusanto GA, Ho CP (2021) On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach. *arXiv preprint arXiv:2110.04855* .

Sturt B (2023) A nonparametric algorithm for optimal stopping based on robust optimization. *Operations Research* 71(5):1530–1557.

Subramanyam A, Gounaris CE, Wiesemann W (2019) K-adaptability in two-stage mixed-integer robust optimization. *Mathematical Programming Computation* 1–32.

Toktay LB, Wein LM (2001) Analysis of a forecasting-production-inventory system with stationary demand. *Management Science* 47(9):1268–1281.

Tulabandhula T, Rudin C (2013) Machine learning with operational costs. *Journal of Machine Learning Research* 14:1989–2028.

Van Parys B, Bennouna MA (2022) Robust two-stage optimization with covariate data. *Available on Optimization Online* .

Van Parys BP, Esfahani PM, Kuhn D (2021) From data to decisions: Distributionally robust optimization is optimal. *Management Science* 67(6):3387–3402.

Vayanos P, Georghiou A, Yu H (2025) Robust optimization with decision-dependent information discovery. *Management Science* .

Wozabal D (2012) A framework for optimization under ambiguity. *Annals of Operations Research* 193(1):21–47.

Xu T, Wenliang LK, Munn M, Acciaio B (2020) Cot-gan: Generating sequential data via causal optimal transport. *Advances in neural information processing systems* 33:8798–8809.

Yakubovich VA (1977) The s-procedure in non-linear control theory. *Vestnik Leninggradskogo Universiteta, Ser. Matematika* 4:73–93.

Yamada T, Watanabe S (1971) On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University* 11(1):155–167.

Yu X, Shen S (2022) Multistage distributionally robust mixed-integer programming with decision-dependent moment-based ambiguity sets. *Mathematical Programming* 196(1):1025–1064.

Zhang L, Yang J, Gao R (2024) Optimal robust policy for feature-based newsvendor. *Management Science* 70(4):2315–2329.

Zhang L, Yang J, Gao R (2025) A short and general duality proof for Wasserstein distributionally robust optimization. *Operations Research* 73(4):2146–2155.

Zhu K, Thonemann UW (2004) An adaptive forecasting algorithm and inventory policy for products with short life cycles. *Naval Research Logistics (NRL)* 51(5):633–653.

Zhu T, Xie J, Sim M (2022) Joint estimation and robustness optimization. *Management Science* 68(3):1659–1677.

# Proofs of Statements

## EC.1. Causal Transport Distance

LEMMA EC.1 **(Equivalent Definition)**. *Let $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$ be a transport plan. Then the following are equivalent.*

(I) $\gamma \in \Gamma_c(\widehat{\mathbb{P}}, \mathbb{P})$.

(II) *For $\widehat{\mathbb{P}}$-almost every $(\widehat{X}, \widehat{Z}) \in \mathcal{X} \times \mathcal{Z}$,*

$$\gamma_{X|(\widehat{X},\widehat{Z})} = \gamma_{X|\widehat{X}}.$$

(III) *Let $\mathrm{Proj}_X : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be the projection into $X$ coordinate. For $\widehat{\mathbb{P}}$-almost every $(\widehat{x}, \widehat{z}_1), (\widehat{x}, \widehat{z}_2) \in \mathcal{X} \times \mathcal{Z}$,*

$$(\mathrm{Proj}_X)_{\#}\gamma(\mathrm{d}x|\widehat{x}, \widehat{z}_1) = (\mathrm{Proj}_X)_{\#}\gamma(\mathrm{d}x|\widehat{x}, \widehat{z}_2).$$

(IV) *For $\widehat{\mathbb{P}}_{\widehat{X}}$-almost every $\widehat{X}$ and $\mathbb{P}_X$-almost every $X$,*

$$\gamma_{\widehat{Z}|(\widehat{X},X)} = \gamma_{\widehat{Z}|\widehat{X}} = \widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}.$$

(V) *Let $\mathrm{Proj}_{\widehat{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathcal{Z}$ be the projection into $\widehat{Z}$ coordinate: $\mathrm{Proj}_{\widehat{Z}}(\widehat{z}, z) = \widehat{z}$. For $\widehat{\mathbb{P}}_{\widehat{X}}$-almost every $\widehat{x} \in \mathcal{X}$ and $\mathbb{P}_X$-almost every $x_1, x_2 \in \mathcal{X}$,*

$$(\mathrm{Proj}_{\widehat{Z}})_{\#}\gamma(\mathrm{d}\widehat{z}|\widehat{x}, x_1) = (\mathrm{Proj}_{\widehat{Z}})_{\#}\gamma(\mathrm{d}\widehat{z}|\widehat{x}, x_2).$$

*Moreover, $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$ plus any one from the above is equivalent to $\gamma \in \mathcal{P}((\mathcal{X} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Z}))$, satisfying*

(VI) *$\gamma$ has a decomposition into successive regular kernels*

$$\gamma(\mathrm{d}\widehat{x}\,\mathrm{d}\widehat{z}\,\mathrm{d}x\,\mathrm{d}z) = \gamma_1(\mathrm{d}\widehat{x}\,\mathrm{d}x)\gamma_2(\mathrm{d}\widehat{z}\,\mathrm{d}z|\widehat{x}, x)$$

*satisfying*

$$\begin{aligned} \gamma_1 &\in \Gamma(\widehat{\mathbb{P}}_{\widehat{X}}, \mathbb{P}_X), \\ (\mathrm{Proj}_{\widehat{Z}})_{\#}\gamma_2(\mathrm{d}\widehat{z}|\widehat{x}, x) &= \widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}(\mathrm{d}\widehat{z}|\widehat{x}) \qquad \textit{for } \gamma_1\textit{-almost every } (\widehat{x}, x), \\ (\mathrm{Proj}_{(X,Z)})_{\#}\gamma_{Z|X}(\mathrm{d}z|x) &= \mathbb{P}_{Z|X}(\mathrm{d}z|x) \qquad \textit{for } \mathbb{P}_X\textit{-almost every } x. \end{aligned}$$

*That is,*

$$\gamma_1 \in \Gamma(\widehat{\mathbb{P}}_{\widehat{X}}, \mathbb{P}_X), \qquad \gamma_2 \in \Gamma(\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}, \mathbb{Q}^{(\widehat{X})}) \text{ where } \mathbb{E}_{\widehat{X}\sim(\gamma_1)_{\widehat{X}|X}}[\mathbb{Q}^{(\widehat{X})}|X] = \mathbb{P}_{Z|X}.$$

*Proof.* The equivalence of (I), (II), and (IV) follows from the definition. It is also easy to check from the definition that (II) is equivalent to (III), and (IV) is equivalent to (V).

Suppose (VI) holds, then projecting $\gamma$ onto $(X, \widehat{X}, \widehat{Z})$ coordinate, we have

$$(\mathrm{Proj}_{(X,\widehat{X},\widehat{Z})})_{\#}\gamma(\mathrm{d}x\,\mathrm{d}\widehat{x}\,\mathrm{d}\widehat{z}) = \gamma_1(\mathrm{d}\widehat{x}\,\mathrm{d}x)\cdot(\mathrm{Proj}_{\widehat{Z}})_{\#}\gamma_2(\mathrm{d}\widehat{z}|\widehat{x},x) = \gamma_1(\mathrm{d}\widehat{x}\,\mathrm{d}x)\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}(\mathrm{d}\widehat{z}|\widehat{x}).$$

Projecting onto $(\widehat{X}, \widehat{Z})$ yields

$$(\mathrm{Proj}_{(\widehat{X},\widehat{Z})})_{\#}\gamma(\mathrm{d}\widehat{x}\,\mathrm{d}\widehat{z}) = (\mathrm{Proj}_{\widehat{X}})_{\#}\gamma_1(\mathrm{d}\widehat{x})\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}(\mathrm{d}\widehat{z}|\widehat{x}) = \widehat{\mathbb{P}}_{\widehat{X}}(\mathrm{d}\widehat{x})\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}(\mathrm{d}\widehat{z}|\widehat{x}) = \widehat{\mathbb{P}}(\widehat{x},\widehat{z}).$$

As for the other marginal,

$$(\mathrm{Proj}_{(X,Z)})_{\#}\gamma(\mathrm{d}x\,\mathrm{d}z) = (\mathrm{Proj}_{X})_{\#}\gamma_1(\mathrm{d}x)\cdot(\mathrm{Proj}_{(X,Z)})_{\#}\gamma_{Z|X}(\mathrm{d}z|x) = \mathbb{P}_X(\mathrm{d}x)\mathbb{P}_{Z|X}(\mathrm{d}z|x) = \mathbb{P}(\mathrm{d}x\,\mathrm{d}z).$$

So indeed we have $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$. □

*Proof of Lemma 1.* Since $\gamma^{(q)}$ are transport plans starting from $\widehat{\mathbb{P}}$,

$$\gamma^{(q)}_{(\widehat{X},\widehat{Z})} = \widehat{\mathbb{P}}, \qquad \gamma^{(q)}_{\widehat{X}} = \widehat{\mathbb{P}}_{\widehat{X}}, \qquad \forall q \in [0,1].$$

Together with

$$\gamma^{(q)}_{(X,\widehat{X},\widehat{Z})} = (1-q)\gamma^{(0)}_{(X,\widehat{X},\widehat{Z})} + q\gamma^{(1)}_{(X,\widehat{X},\widehat{Z})}, \qquad \gamma^{(q)}_{(X,\widehat{X})} = (1-q)\gamma^{(0)}_{(X,\widehat{X})} + q\gamma^{(1)}_{(X,\widehat{X})},$$

we know that

$$\gamma^{(q)}_{X|(\widehat{X},\widehat{Z})} = (1-q)\gamma^{(0)}_{X|(\widehat{X},\widehat{Z})} + q\gamma^{(1)}_{X|(\widehat{X},\widehat{Z})}, \qquad \gamma^{(q)}_{X|\widehat{X}} = (1-q)\gamma^{(0)}_{X|\widehat{X}} + q\gamma^{(1)}_{X|\widehat{X}}.$$

Because $\gamma^{(0)}$ and $\gamma^{(1)}$ are causal, by equivalent definition (II), for $\widehat{\mathbb{P}}$-almost every $(\widehat{X}, \widehat{Z}) \in \mathcal{X} \times \mathcal{Z}$,

$$\gamma^{(0)}_{X|(\widehat{X},\widehat{Z})} = \gamma^{(0)}_{X|\widehat{X}}, \qquad \gamma^{(1)}_{X|(\widehat{X},\widehat{Z})} = \gamma^{(1)}_{X|\widehat{X}}.$$

Therefore

$$\gamma^{(q)}_{X|(\widehat{X},\widehat{Z})} = \gamma^{(q)}_{X|\widehat{X}},$$

so $\gamma^{(q)}$ is also causal.

*Proof.* With probability one, each $\widehat{x}$ in the support of $\widehat{\mathbb{P}}$ corresponds to only one $\widehat{z}$, so that

$$\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}=\widehat{x}_k} = \delta_{\widehat{z}_k}.$$

Now let $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$. Because

$$\mathbb{E}_{X|\widehat{X}}[\gamma_{\widehat{Z}|(\widehat{X},X)}] = \gamma_{\widehat{Z}|\widehat{X}} = \delta_{\widehat{Z}},$$

the only choice is $\gamma_{\widehat{Z}|(\widehat{X},X)} = \delta_{\widehat{X}}$, for $(\gamma_1)_{X|\widehat{X}}$-a.e. $X$. Therefore, $\gamma$ is causal. □

## EC.2. Supremum of Convex Functions

In this subsection, we provide several auxiliary results on the properties of the supremum of a family of convex functions. Analysis in this subsection will be used in the proof of Theorem 1.

LEMMA EC.2 **(Dual Objective Function)**. *Recall the dual objective function $h(\lambda)$ defined in (5). Let $\mathcal{I}=\{\lambda\in[0,\infty): h(\lambda)<\infty\}$. Then it has the following properties:*

(I) *There exists $\kappa\geq 0$, such that either $\mathcal{I}=(\kappa,\infty)$ or $\mathcal{I}=[\kappa,\infty)$.*

(II) *$h$ is convex and continuous in $\mathcal{I}$.*

(III) *$h(\lambda)\to\infty$ as $\lambda\to\infty$.*

(IV) *$h$ has a minimizer $\lambda^*\in[\kappa,\infty)$.*

*Proof.* (I) $h(\lambda)-\lambda\rho^p$ is monotonically decreasing in $\lambda$, therefore we can find $\kappa$ such that $h$ is infinite for smaller $\lambda$, and finite for greater $\lambda$.

(II) $h$ is a combination of supremums and expectations of convex functions, and therefore $h$ is convex. Since $h<\infty$ in $\mathcal{I}$, $h$ is continuous in $\mathcal{I}$ with only a possible exception at $\kappa\in\mathcal{I}$. Now we verify its lower semi-continuity at $\kappa$.

$$\begin{aligned}
\liminf_{\lambda\downarrow\kappa} h(\lambda) &= \kappa\rho^p+\liminf_{\lambda\downarrow\kappa}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\lambda\|x-\widehat{X}\|^p\right\}\right]\\
&\geq \kappa\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\liminf_{\lambda\downarrow\kappa}\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\kappa\|x-\widehat{X}\|^p\right\}\right]\\
&\geq \kappa\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\liminf_{\lambda\downarrow\kappa}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\kappa\|x-\widehat{X}\|^p\right\}\right]\\
&\geq \kappa\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\liminf_{\lambda\downarrow\kappa}\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\kappa\|x-\widehat{X}\|^p\right\}\right]\\
&\geq \kappa\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\liminf_{\lambda\downarrow\kappa}\Psi(f(x),z)-\lambda\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\kappa\|x-\widehat{X}\|^p\right\}\right]\\
&= \kappa\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(f(x),z)-\kappa\|z-\widehat{Z}\|^p\right\}\mid\widehat{X}\right]-\kappa\|x-\widehat{X}\|^p\right\}\right]=h(\kappa).
\end{aligned}$$

We repeatedly used the Fatou's lemma and that minimax is greater than maximin. By convexity, $h$ is continuous in $\mathcal{I}$.

(III) This is simply because we can pick $x=\widehat{X}$, $z=\widehat{Z}$ so

$$\begin{aligned}
h(\lambda) &\geq \lambda\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(f(\widehat{X}),\widehat{Z})-\lambda\|\widehat{Z}-\widehat{Z}\|^p\mid\widehat{X}\right]-\lambda\|\widehat{X}-\widehat{X}\|^p\right]\\
&= \lambda\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\mid\widehat{X}\right]\right]=\lambda\rho^p+\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right]\to+\infty
\end{aligned}$$

as $\lambda\to+\infty$.

(IV) It follows from (I)-(III). □

LEMMA EC.3 **(Exchange Sup and Derivative for Convex Functions)**. *Let $\Lambda$ be an index set. Let $\{F_\alpha\}_{\alpha\in\Lambda}$ be a family of real-valued convex functions defined on an interval $\mathcal{I}$. Suppose its sup is pointwise bounded, $\Phi(\lambda)=\sup_{\alpha\in\Lambda}F_\alpha(\lambda)<\infty$. Denote $f_\alpha(\lambda)=F'_\alpha(\lambda)$, and $\phi(\lambda)=\Phi'(\lambda)$. For any function $f$, we denote $f^*$ [resp. $f_*$] to be the upper [resp. lower] semicontinuous envelope of $f$. For every $\varepsilon>0$, define the $\varepsilon$-argmax set $\Omega_\varepsilon$ and $\overline{D},\underline{D}$ by*

$$\begin{aligned}
&\Omega_\varepsilon(\lambda):=\{\alpha\in\Lambda:\ F_\alpha(\lambda)\geq\Phi(\lambda)-\varepsilon\},\\
&\overline{D}_\varepsilon(\lambda):=\sup_{\alpha\in\Omega_\varepsilon(\lambda)}f^*_\alpha(\lambda),\qquad \overline{D}(\lambda)=\lim_{\varepsilon\to0}\overline{D}_\varepsilon(\lambda),\\
&\underline{D}_\varepsilon(\lambda):=\inf_{\alpha\in\Omega_\varepsilon(\lambda)}f_{\alpha*}(\lambda),\qquad \underline{D}(\lambda)=\lim_{\varepsilon\to0}\underline{D}_\varepsilon(\lambda).
\end{aligned}$$

*Then*

(I) *For every $\lambda\in\mathcal{I}$, $\underline{D}(\lambda)\leq\overline{D}(\lambda)$.*

(II) *For every $\lambda,\mu\in\mathcal{I}$ with $\lambda<\mu$, $\overline{D}(\lambda)\leq\phi^*(\lambda)\leq\phi_*(\mu)\leq\underline{D}(\mu)$.*

(III) *Fix $\lambda\in\mathcal{I}$, $\delta>0$, $\epsilon>0$. If $\lambda_1\in\mathcal{I}$ such that $\lambda_1<\lambda$ is sufficiently close to $\lambda$, then we can find $\alpha\in\Lambda$ such that*

$$f^*_\alpha(\lambda_1)\leq\phi_*(\lambda)+\delta,\qquad F_\alpha(\lambda_1)\geq\Phi(\lambda)-\epsilon.$$

*If $\lambda_2\in\mathcal{I}$ such that $\lambda_2>\lambda$ is sufficiently close to $\lambda$, we can find $\beta\in\Lambda$ such that*

$$f_{\beta*}(\lambda_2)\geq\phi^*(\lambda)-\delta,\qquad F_\beta(\lambda_2)\geq\Phi(\lambda)-\epsilon.$$

*Proof.* $\Phi$ is the sup of a family of convex functions, so $\Phi$ is convex. Since $\Phi$ and $F_\alpha$ are convex and finite in $\mathcal{I}$, they have locally Lipschitz, monotonously increasing derivatives $\phi$ and $f_\alpha$. Monotonicity implies $f^*_\alpha$ and $\phi^*$ [resp. $f_{\alpha*}$ and $\phi_*$] are right [resp. left] continuous, and thus convexity implies for $\lambda<\mu$,

$$f_\alpha{}^*(\lambda)\leq\frac{F_\alpha(\mu)-F_\alpha(\lambda)}{\mu-\lambda}\leq f_{\alpha*}(\mu),\qquad \phi^*(\lambda)\leq\frac{\Phi(\mu)-\Phi(\lambda)}{\mu-\lambda}\leq\phi_*(\mu). \tag{EC.1}$$

(I) $\varepsilon$-argmax set $\Omega_\varepsilon$ is never empty by definition. Therefore, $\underline{D}_\varepsilon(\lambda)\leq\overline{D}_\varepsilon(\lambda)$ holds for all $\varepsilon$. As $\varepsilon\to0$, $\Omega_\varepsilon(\lambda)$ shrinks, so $\overline{D}_\varepsilon(\lambda)\downarrow\overline{D}(\lambda)$, $\underline{D}_\varepsilon(\lambda)\uparrow\underline{D}(\lambda)$, we have $\underline{D}(\lambda)\leq\overline{D}(\lambda)$.

(II) Fix any $\varepsilon>0$, and $\lambda<\mu$. For any $\alpha\in\Omega_\varepsilon(\lambda)$, $\beta\in\Omega_\varepsilon(\mu)$, using (EC.1) we have

$$\begin{aligned}
&F_\alpha(\mu)-\varepsilon\leq\Phi(\mu)-\varepsilon\leq F_\beta(\mu)\leq F_\beta(\lambda)+(\mu-\lambda)f_{\beta*}(\mu)\leq\Phi(\lambda)+(\mu-\lambda)f_{\beta*}(\mu),\\
&F_\beta(\lambda)-\varepsilon\leq\Phi(\lambda)-\varepsilon\leq F_\alpha(\lambda)\leq F_\alpha(\mu)-(\mu-\lambda)f_\alpha{}^*(\lambda)\leq\Phi(\mu)-(\mu-\lambda)f_\alpha{}^*(\lambda).
\end{aligned}$$

By these two inequalities, we conclude

$$\begin{aligned}&-\varepsilon+(\mu-\lambda)f_\alpha^*(\lambda)\le\Phi(\mu)-\Phi(\lambda)\le\varepsilon+(\mu-\lambda)f_{\beta_*}(\mu),\\ \Rightarrow&-\frac{\varepsilon}{\mu-\lambda}+f_\alpha^*(\lambda)\le\frac{\Phi(\mu)-\Phi(\lambda)}{\mu-\lambda}\le\frac{\varepsilon}{\mu-\lambda}+f_{\beta_*}(\mu).\end{aligned}$$

By taking the sup over $\alpha\in\Omega_\varepsilon(\lambda)$, taking the inf over $\beta\in\Omega_\varepsilon(\mu)$, we have

$$-\frac{\varepsilon}{\mu-\lambda}+\overline{D}_\varepsilon(\lambda)\le\frac{\Phi(\mu)-\Phi(\lambda)}{\mu-\lambda}\le\frac{\varepsilon}{\mu-\lambda}+\underline{D}_\varepsilon(\mu).$$

Let $\varepsilon\to 0$,

$$\overline{D}(\lambda)\le\frac{\Phi(\mu)-\Phi(\lambda)}{\mu-\lambda}\le\underline{D}(\mu). \tag{EC.2}$$

We now combine (EC.1) with (EC.2) to show that $\phi^*(\lambda)\le\underline{D}(\mu)$, $\overline{D}(\lambda)\le\phi_*(\mu)$. To finish the proof of (II), we use the monotonicity $\phi^*(\lambda)\le\phi_*(\mu)$, and

$$\phi^*(\lambda)=\lim_{\mu\downarrow\lambda}\phi(\mu)\ge\lim_{\mu\downarrow\lambda}\phi_*(\mu)\ge\overline{D}(\lambda),\qquad\phi_*(\mu)=\lim_{\lambda\uparrow\mu}\phi(\lambda)\le\lim_{\lambda\uparrow\mu}\phi^*(\lambda)\le\underline{D}(\mu).$$

(III) Since $\Phi$ is continuous in the interior of $\mathcal{I}$, we can let $\lambda_1$ and $\lambda_2$ be close enough to $\lambda$ such that

$$\Phi(\lambda_1),\Phi(\lambda_2)\ge\Phi(\lambda)-\frac{\epsilon}{2}.$$

Let $\varepsilon<\frac{\epsilon}{2}$ be small enough such that $\overline{D}_\varepsilon(\lambda_1)<\overline{D}(\lambda_1)+\delta$, $\underline{D}_\varepsilon(\lambda_2)>\underline{D}(\lambda_2)-\delta$. Pick any $\alpha\in\Omega_\varepsilon(\lambda_1),\beta\in\Omega_\varepsilon(\lambda_2)$, then

$$\begin{aligned}f_\alpha^*(\lambda_1)\le\overline{D}_\varepsilon(\lambda_1)<\overline{D}(\lambda_1)+\delta\le\phi_*(\lambda)+\delta,\\ f_{\beta_*}(\lambda_2)\ge\underline{D}_\varepsilon(\lambda_2)>\underline{D}(\lambda_2)-\delta\ge\phi^*(\lambda)-\delta.\end{aligned}$$

Moreover, by the definition of $\Omega_\varepsilon(\lambda)$,

$$\begin{aligned}F_\alpha(\lambda_1)\ge\Phi(\lambda_1)-\varepsilon\ge\Phi(\lambda_1)-\frac{\epsilon}{2}\ge\Phi(\lambda)-\epsilon,\\ F_\beta(\lambda_2)\ge\Phi(\lambda_2)-\varepsilon\ge\Phi(\lambda_2)-\frac{\epsilon}{2}\ge\Phi(\lambda)-\epsilon.\end{aligned}$$

□

Lemma EC.4. *With the same notations as the previous lemma, let $\Lambda$ be an Euclidean space. Suppose for each $\lambda\in\mathrm{Int}(\mathcal{I})$, $F_\alpha(\lambda)$ is upper semicontinuous in $\alpha$, and $|f_\alpha(\lambda)|\to\infty$ as $|\alpha|\to\infty$. Then*

(I) $\Omega_0(\lambda)$ *is nonempty.*

(II) *There exists* $\alpha, \beta \in \Omega_0(\lambda)$*, such that*

$$f_{\alpha *}(\lambda) = \phi_*(\lambda), \qquad f^*_\beta(\lambda) = \phi^*(\lambda), \qquad F_\alpha(\lambda) = F_\beta(\lambda) = \Phi(\lambda).$$

(III) $\overline{D}_0(\lambda) = \overline{D}(\lambda) = \phi^*(\lambda)$*, and* $\underline{D}_0(\lambda) = \underline{D}(\lambda) = \phi_*(\lambda)$*.*

*Proof.* Let $\lambda_0 \in \mathrm{Int}(\mathcal{I})$. Then we can find $\kappa < \lambda_0 < \mu$ all inside $\mathrm{Int}(\mathcal{I})$. For some small $\delta$, $\kappa' = \kappa - \delta$ and $\mu' = \mu + \delta$ are also inside $\mathrm{Int}(\mathcal{I})$.

(I) By Lemma EC.3 (II), $\phi_*(\lambda) \le \underline{D}(\lambda) \le \overline{D}(\lambda) \le \phi^*(\lambda)$, and since $\lambda$ is in the interior of $\mathcal{I}$, $\Phi$ is locally Lipschitz, $\underline{D}(\lambda), \overline{D}(\lambda)$ are finite. Thus for some small $\varepsilon$, $\underline{D}_\varepsilon(\lambda)$ and $\overline{D}_\varepsilon(\lambda)$ are finite. This implies that $\Omega_\varepsilon$ is bounded, otherwise $|f_\alpha(\lambda)| \to \infty$ as $\alpha \to \infty$. Because $F_\alpha$ is upper semicontinuous, $\Omega_\varepsilon$ is also closed, so it is compact, thus

$$\Phi(\lambda) = \sup_{\alpha \in \Lambda} F_\alpha(\lambda) = \sup_{\alpha \in \Omega_\varepsilon(\lambda)} F_\alpha(\lambda)$$

is attainable, i.e.,

$$\Omega_0(\lambda) = \arg\max_{\alpha \in \Lambda} F_\alpha(\lambda)$$

is nonempty.

(II) For every $\lambda$, since $\Omega_0(\lambda) \subset \Omega_\varepsilon(\lambda)$ for any $\varepsilon$, we know that $\overline{D}_\varepsilon(\lambda) \ge \overline{D}_0(\lambda)$, $\underline{D}_\varepsilon(\lambda) \le \underline{D}_0(\lambda)$. Let $\varepsilon \to 0$ we have $\overline{D}(\lambda) \ge \overline{D}_0(\lambda)$, $\underline{D}(\lambda) \le \underline{D}_0(\lambda)$. So for every $\alpha \in \Omega_0(\lambda)$,

$$\phi_*(\lambda) \le \underline{D}(\lambda) \le \underline{D}_0(\lambda) \le f_{\alpha *}(\lambda) \le f^*_\alpha(\lambda) \le \overline{D}_0(\lambda) \le \overline{D}(\lambda) \le \phi^*(\lambda). \tag{EC.3}$$

Let $\lambda_n \uparrow \lambda_0$ be an increasing sequence inside $[\kappa, \mu]$. For each $\lambda_n$, $\Omega_0(\lambda_n)$ is nonempty, so we can find $\alpha_n$ such that

$$F_{\alpha_n}(\lambda_n) = \Phi(\lambda_n), \qquad \phi_*(\lambda_n) \le f_{\alpha_n *}(\lambda_n) \le f^*_{\alpha_n}(\lambda_n) \le \phi^*(\lambda_n).$$

First, we claim that $F_{\alpha_n}$ are uniformly bounded in $[\kappa, \mu]$. The upper bound $F_{\alpha_n} \le \Phi$ is clear. As for the lower bound, we first use the convexity of $\Phi$, for all $\lambda \in [\kappa, \mu]$,

$$\Phi(\lambda) \ge \Phi(\kappa) + \phi^*(\kappa)(\lambda - \kappa), \qquad \Phi(\lambda) \ge \Phi(\mu) - \phi_*(\mu)(\mu - \lambda).$$

then we use the convexity of $F_{\alpha_n}$, for $\lambda \in [\lambda_n, \mu]$,

$$\begin{aligned} F_{\alpha_n}(\lambda) &\geq F_{\alpha_n}(\lambda_n) + f^*_{\alpha_n}(\lambda_n)(\lambda - \lambda_n) \\ &\geq \Phi(\lambda_n) + \phi_*(\lambda_n)(\lambda - \lambda_n) \\ &\geq \Phi(\kappa) + \phi^*(\kappa)(\lambda_n - \kappa) + \phi^*(\kappa)(\lambda - \lambda_n) \\ &= \Phi(\kappa) + \phi^*(\kappa)(\lambda - \kappa). \end{aligned}$$

For $\lambda \in [\kappa, \lambda_n]$,

$$\begin{aligned} F_{\alpha_n}(\lambda) &\geq F_{\alpha_n}(\lambda_n) - f_{\alpha_n *}(\lambda_n)(\lambda_n - \lambda) \\ &\geq \Phi(\lambda_n) - \phi^*(\lambda_n)(\lambda_n - \lambda) \\ &\geq \Phi(\mu) - \phi_*(\mu)(\mu - \lambda_n) - \phi_*(\mu)(\lambda_n - \lambda) \\ &= \Phi(\mu) - \phi_*(\mu)(\mu - \lambda). \end{aligned} \tag{EC.4}$$

Therefore, for all $\lambda \in [\kappa, \mu]$,

$$F_{\alpha_n}(\lambda) \geq \min\left\{\Phi(\kappa) + \phi^*(\kappa)(\lambda - \kappa), \Phi(\mu) - \phi_*(\mu)(\mu - \lambda)\right\}.$$

Next, we claim that $F_{\alpha_n}$ are equicontinuous in $[\kappa, \mu]$. Since

$$F_{\alpha_n}(\kappa) \geq \min\left\{\Phi(\kappa), \Phi(\mu) - \phi_*(\mu)(\mu - \kappa)\right\} = \Phi(\mu) - \phi_*(\mu)(\mu - \kappa),$$

by convexity of $F_{\alpha_n}$ we have

$$f_{\alpha_n *}(\kappa) \geq \frac{F_{\alpha_n}(\kappa) - F_{\alpha_n}(\kappa')}{\kappa - \kappa'} \geq \frac{\Phi(\mu) - \phi_*(\mu)(\mu - \kappa) - \Phi(\kappa')}{\delta}.$$

Similarly, we have

$$f^*_{\alpha_n}(\mu) \leq \frac{F_{\alpha_n}(\mu') - F_{\alpha_n}(\mu)}{\mu' - \mu} \leq \frac{\Phi(\mu') - \Phi(\kappa) - \phi^*(\kappa)(\mu - \kappa)}{\delta}.$$

$f_{\alpha_n}$ are increasing between $\kappa$ and $\mu$, so they are uniformly bounded, thus $F_{\alpha_n}$ are uniformly Lipschitz.

Since $f_{\alpha_n}$ are uniformly bounded, we know that $\{\alpha_n\}_{n\in\mathbb{N}}$ is bounded by the assumption of the lemma. Up to a subsequence, we may assume $\alpha_n \to \alpha$. Since $F_{\alpha_n}$ are uniformly bounded and equicontinuous in $[\kappa, \mu]$, by Arzelà–Ascoli Lemma it admits a

subsequence uniformly converging to some $F_\infty$, and since $F_\alpha$ is upper semicontinuous in $\alpha$, we know that $F_\alpha \geq \lim_{n\to\infty} F_{\alpha_n} = F_\infty$. Therefore, up to a subsequence,

$$\Phi(\lambda_0) \geq F_\alpha(\lambda_0) \geq F_\infty(\lambda_0) = \lim_{n\to\infty} F_{\alpha_n}(\lambda_n) = \lim_{n\to\infty} \Phi(\lambda_n) = \Phi(\lambda_0).$$

Thus $\alpha \in \Omega_0(\lambda_0)$. Moreover, by taking $n \to \infty$ in (EC.4), for any $\lambda \in [\kappa, \lambda_0)$ we have

$$\begin{aligned}\Phi(\lambda) \geq F_\alpha(\lambda) \geq F_\infty(\lambda) &= \lim_{n\to\infty} F_{\alpha_n}(\lambda_n) - f_{\alpha_n *}(\lambda_n)(\lambda_n - \lambda) \\ &\geq \lim_{n\to\infty} F_{\alpha_n}(\lambda_n) - \phi_*(\lambda_n)(\lambda_n - \lambda) = \Phi(\lambda_0) - \phi_*(\lambda_0)(\lambda_0 - \lambda),\end{aligned}$$

and they all equal at $\lambda = \lambda_0$. So the left derivative at $\lambda_0$

$$\phi_*(\lambda_0) \geq f_{\alpha *}(\lambda_0) \geq \phi_*(\lambda_0)$$

are equal. This shows that $f_{\alpha *}(\lambda_0) = \phi_*(\lambda_0)$. The proof for the $\beta$ part is exactly symmetric to the $\alpha$, so we omit it here.

(III) This is the consequence of part (II) and (EC.3). □

The proofs in this appendix were inspired by the construction in Gao and Kleywegt (2023). However, because of the two-layer structure of the causal transport distance, we need Lemma EC.3 and EC.4, which cover more general scenarios than the ones considered in Gao and Kleywegt (2023).

## EC.3. Proofs for Section 3.1

*Proof of Theorem 1.* It suffices to prove the direction $v_{\mathsf{P}}^f \geq v_{\mathsf{D}}^f$. We may assume

$$\mathbb{E}_{\widehat{\mathbb{P}}}[\Psi(f(\widehat{X}), \widehat{Z})] < \infty, \tag{EC.5}$$

otherwise $v_{\mathsf{P}}^f = \infty$ because $\widehat{\mathbb{P}}$ is feasible, and the strong duality holds automatically.

For each $x \in \mathcal{X}$, $\widehat{z} \in \mathcal{Z}$ we denote

$$G_{(z)}(\lambda; x, \widehat{z}) := \Psi(f(x), z) - \lambda \|z - \widehat{z}\|^p.$$

It is a linearly decreasing function of $\lambda$. Thus, the supremum over $z$

$$\Upsilon(\lambda; x, \widehat{z}) := \sup_{z\in\mathcal{Z}} \left\{ G_{(z)}(\lambda; x, \widehat{z}) \right\} \tag{EC.6}$$

is a decreasing convex function of $\lambda$. Because the expectation of decreasing convex functions are decreasing and convex, we have for each $\widehat{x} \in \mathcal{X}$,

$$F_{(x)}(\lambda; \widehat{x}) := \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[ \Upsilon(\lambda; x, \widehat{Z}) \mid \widehat{X} = \widehat{x} \right] - \lambda \|x - \widehat{x}\|^p$$

is a family of decreasing convex functions of $\lambda$. Their supremum

$$\Phi(\lambda; \widehat{x}) := \sup_{x \in \mathcal{X}} \left\{ F_{(x)}(\lambda; \widehat{x}) \right\} \tag{EC.7}$$

is again convex and decreasing. Finally, the dual objective function

$$h(\lambda) = \lambda \rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \Phi(\lambda; \widehat{X}) \right]$$

is also convex. By Lemma EC.2, there exists $\kappa \in [0, \infty]$ such that $h$ is finite in $(\kappa, \infty)$ and infinite in $[0, \kappa)$. Moreover, in the case $\kappa < \infty$, $h$ attains its global minimum at $\lambda^* \geq \kappa$. Therefore, we can separate the following cases.

***Case 1:*** $\kappa = \infty$

This means $h(\lambda) = \infty$ for any $\lambda \geq 0$, therefore $v_{\mathsf{D}}^f = \infty$. Now fix $\lambda > 0$, then $h(\lambda) = 0$ implies

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \Phi(\lambda; \widehat{X}) \right] = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \sup_{x \in \mathcal{X}} F_{(x)}(\lambda; \widehat{X}) \right] = \infty.$$

Recall that $\widehat{\mathbb{P}}_{\widehat{X}}$ is a discrete measure. Therefore, for some $\widehat{x}_*$ in the support of $\widehat{\mathbb{P}}_{\widehat{X}}$, $\Phi(\lambda; \widehat{x}_*) = \infty$. For some large number $N > 0$ to be determined, by the definition of $\Phi(\lambda; \widehat{x})$, one can find $x_* \in \mathcal{X}$ such that $F_{(x_*)}(\lambda; \widehat{x}_*) > N$. Define $T_1 : \mathcal{X} \to \mathcal{X}$ by

$$T_1(x) = \begin{cases} x & x \neq \widehat{x}_* \\ x_* & x = \widehat{x}_* \end{cases}.$$

Then

$$\begin{aligned} \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ F_{(T_1(\widehat{X}))}(\lambda; \widehat{X}) \right] &= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ F_{(\widehat{X})}(\lambda; \widehat{X}) \mathbf{1}\{\widehat{X} \neq \widehat{x}_*\} \right] + \widehat{\mathbb{P}}_{\widehat{X}}(\{\widehat{x}_*\}) \cdot F_{(x_*)}(\lambda; \widehat{x}_*) \\ &\geq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ F_{(\widehat{X})}(\lambda; \widehat{X}) \mathbf{1}\{\widehat{X} \neq \widehat{x}_*\} \right] + \widehat{\mathbb{P}}_{\widehat{X}}(\{\widehat{x}_*\}) \cdot N \end{aligned}$$

can be made arbitrarily large by selecting sufficiently large $N$. In particular, we can choose $N$ large enough to ensure

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ F_{T_1(\widehat{X})}(\lambda; \widehat{X}) \right] \geq \mathbb{E}_{\widehat{\mathbb{P}}} \left[ \Psi(f(\widehat{X}), \widehat{Z}) \right] + 2\lambda \rho^p.$$

Here we have used the assumption (EC.5). We may abbreviate as $X := T_1(\widehat{X})$, so that

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[F_X(\lambda;\widehat{X})\right] \geq \mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] + 2\lambda\rho^p.$$

By the definition of $F_{(x)}$, we have

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;X,\widehat{Z}) \mid \widehat{X}\right] - \lambda\|X-\widehat{X}\|^p\right] \geq \mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] + 2\lambda\rho^p.$$

Rearranging the terms, we conclude

$$\begin{aligned} 2\lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\lambda\|X-\widehat{X}\|^p\right] &\leq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;X,\widehat{Z}) - \Psi(f(\widehat{X}),\widehat{Z}) \mid \widehat{X}\right]\right] \\ &= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}} G_{(z)}(\lambda;X,\widehat{Z}) - \Psi(f(\widehat{X}),\widehat{Z}) \mid \widehat{X}\right]\right] \end{aligned} \tag{EC.8}$$

We would like to construct $T_2 : \mathcal{X}\times\mathcal{Z}\to\mathcal{Z}$ in two different scenarios. Fix $\widehat{x}_k$ in the support of $\widehat{\mathbb{P}}_{\widehat{X}}$ and $x_k = T_1(\widehat{x}_k)$. The first scenario would be if $\Upsilon(\lambda;x_k,\widehat{z}_*) = \infty$ for some $\widehat{z}_*$ within the support of $\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}=\widehat{x}_k}$. Similar to the construction of $T_1$, one can find some one-point transport map $T_2(\widehat{x}_k,\cdot)$ such that

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[G_{(T_2(\widehat{x}_k,\widehat{Z}))}(\lambda;x_k,\widehat{Z}) \mid \widehat{X}=\widehat{x}_k\right] \geq N$$

for any large $N$ to be prescribed. The second scenario would be if $\Upsilon(\lambda;x_k,\widehat{z}_*) < \infty$ for every $\widehat{z}_*$ within the support of $\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}$. For each $(\widehat{x}_k,\widehat{z}_{ki})$ pair, we can find $z_{ki}\in\mathcal{Z}$ such that

$$G_{(z_{ki})}(\lambda;x_k,\widehat{z}_{ki}) \geq \Upsilon(\lambda;x_k,\widehat{z}_{ki}) - \lambda\rho^p.$$

Define $T_2(x_k,\widehat{z}_{ki}) = z_{ki}$, then

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[G_{(T_2(\widehat{x}_k,\widehat{Z}))}(\lambda;x_k,\widehat{Z}) \mid \widehat{X}=\widehat{x}_k\right] \geq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;x_k,\widehat{Z}) \mid \widehat{X}=\widehat{x}_k\right] - \lambda\rho^p.$$

Combine two scenarios and denote $Z = T_2(\widehat{X},\widehat{Z})$, then we have

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[G_{(Z)}(\lambda;X,\widehat{Z}) \mid \widehat{X}\right]\right] \geq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;X,\widehat{Z}) \mid \widehat{X}\right]\right] - \lambda\rho^p.$$

Indeed, if the first scenario happens for any $\widehat{x}_k$, then the right-hand side can be made arbitrarily large. Together with (EC.8), we have

$$\begin{aligned} \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\lambda\|X-\widehat{X}\|^p\right] &\leq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[G_{(Z)}(\lambda;X,\widehat{Z}) - \Psi(f(\widehat{X}),\widehat{Z}) \mid \widehat{X}\right]\right] \\ &= \mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(X),Z) - \Psi(f(\widehat{X}),\widehat{Z}) - \lambda\|Z-\widehat{Z}\|^p\right]. \end{aligned}$$

Assemble $T_1$ and $T_2$ yields a causal transport map $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{X} \times \mathcal{Z}$ defined by $T(\widehat{x}, \widehat{z}) = (T_1(\widehat{x}), T_2(\widehat{x}, \widehat{z}))$. It induces a causal transport plan $\gamma_1 = (T \otimes \mathrm{id}_{\mathcal{X}\times\mathcal{Z}})_\sharp \widehat{\mathbb{P}}$, with $\sharp$ denotes push-forward of a measure. Then $((X, Z), (\widehat{X}, \widehat{Z})) \sim \gamma_1$. Denote the transportation cost between $(\widehat{X}, \widehat{Z})$ and $(X, Z)$ via $\gamma_1$ by

$$D = \mathbb{E}_{\gamma_1}\left[\|X - \widehat{X}\|^p + \|Z - \widehat{Z}\|^p\right],$$

then

$$\mathbb{E}_{\gamma_1}\left[\Psi(f(X), Z) - \Psi(f(\widehat{X}), \widehat{Z})\right] \geq \lambda\rho^p + \lambda D.$$

Let $\gamma_0 = (\mathrm{id}_{\mathcal{X}\times\mathcal{Z}} \otimes \mathrm{id}_{\mathcal{X}\times\mathcal{Z}})_\sharp \widehat{\mathbb{P}}$ denote the joint distribution induced by identity transport map. Let $\gamma_\theta = \theta\gamma_1 + (1-\theta)\gamma_0$ be the transport plan which perturbs $\gamma_0$ by moving $\theta := \min\{1, \frac{\rho^p}{D}\}$ portion of mass from $(\widehat{X}, \widehat{Z})$ to $(X, Z)$. By the convexity lemma 1, this transport plan is causal. Denote $\mathbb{P}_\theta = (\gamma_\theta)_{(X,Z)}$ to be the marginal of $\gamma_\theta$. Then

$$\mathsf{C}_p(\widehat{\mathbb{P}}, \mathbb{P})^p \leq \mathbb{E}_{\gamma_\theta}\left[\|X - \widehat{X}\|^p + \|Z - \widehat{Z}\|^p\right] = \theta D \leq \rho^p,$$

So $\mathbb{P}_\theta$ is primal feasible, and

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\theta}[\Psi(f(X), Z)] - \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi(f(\widehat{X}), \widehat{Z})] &= \mathbb{E}_{\gamma_\theta}\left[\Psi(f(X), Z) - \Psi(f(\widehat{X}), \widehat{Z})\right] \\
&= \theta\mathbb{E}_{\gamma_1}\left[\Psi(f(x), Z) - \Psi(f(\widehat{x}), \widehat{Z})\right] \\
&\geq \theta\left(\lambda\rho^p + \lambda D\right) \\
&\geq \lambda\rho^p.
\end{aligned}$$

Therefore

$$v_{\mathsf{P}}^f \geq \mathbb{E}_{\mathbb{P}_\theta}[\Psi(f(X), Z)] \geq \mathbb{E}_{\widehat{\mathbb{P}}}[\Psi(f(\widehat{X}), \widehat{Z})] + \lambda\rho^p,$$

and since $\lambda$ can be arbitrarily large, we have

$$v_{\mathsf{P}}^f = \infty = v_{\mathsf{D}}^f.$$

***Case 2:*** $\kappa<\infty, \lambda^*>\kappa$

Fix some small $\delta>0$, $\varepsilon>0$. Applying Lemma EC.3 on (EC.7), for $\widehat{x}\in\mathcal{X}$ we can find $\overline{x},\underline{x}\in\mathcal{X}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda^+}F_{(\underline{x})}(\lambda_1;\widehat{x})\le\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Phi(\lambda^*;\widehat{x})+\delta, \qquad \frac{\mathrm{d}}{\mathrm{d}\lambda^-}F_{(\overline{x})}(\lambda_2;\widehat{x})\ge\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Phi(\lambda^*;\widehat{x})-\delta,$$
$$F_{(\underline{x})}(\lambda_1,\widehat{x})\ge\Phi(\lambda^*,\widehat{x})-\varepsilon, \qquad F_{(\overline{x})}(\lambda_2,\widehat{x})\ge\Phi(\lambda^*,\widehat{x})-\varepsilon$$

for $\kappa<\lambda_1<\lambda^*<\lambda_2$ and $\lambda_1,\lambda_2$ sufficiently close to $\lambda^*$. Fix $x\in\mathcal{X}$. Apply Lemma EC.3 on (EC.6), for $\widehat{z}\in\mathcal{Z}$ we can find $\overline{z},\underline{z}\in\mathcal{Z}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda^+}G_{(\underline{z})}(\lambda_3;x,\widehat{z})\le\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Upsilon(\lambda_1;x,\widehat{z})+\delta, \qquad \frac{\mathrm{d}}{\mathrm{d}\lambda^-}G_{(\overline{z})}(\lambda_4;x,\widehat{z})\ge\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Upsilon(\lambda_2;x,\widehat{z})-\delta,$$
$$G_{(\underline{z})}(\lambda_3;x,\widehat{z})\ge\Upsilon(\lambda_1,x,\widehat{z})-\varepsilon, \qquad G_{(\overline{z})}(\lambda_4;x,\widehat{z})\ge\Upsilon(\lambda_2,x,\widehat{z})-\varepsilon$$

for $\kappa<\lambda_3<\lambda_1<\lambda^*<\lambda_2<\lambda_4$ and $\lambda_3,\lambda_4$ sufficiently close to $\lambda_1,\lambda_2$. Now suppose $\widehat{\mathbb{P}}$ is supported over a finite set of $\{(\widehat{x}_k,\widehat{z}_{ki})\}_{ki}$, we know that for $\lambda_1,\lambda_2,\lambda_3,\lambda_4$ sufficiently close to $\lambda^*$ we can find $\overline{x}_k,\underline{x}_k,\overline{z}_{ki},\underline{z}_{ki}$ such that the above are satisfied simultaneously. We denote the transport map by $\overline{x}_k=\overline{T}_1(\widehat{x}_i)$, $\overline{z}_{ki}=\overline{T}_2(\widehat{x}_k,\widehat{z}_{ki})$, and $\overline{T}(\widehat{x}_k,\widehat{z}_{ki})=(\overline{x}_k,\overline{z}_{ki})$. We define $\underline{T}$ similarly, so we can construct $(\overline{X},\overline{Z})=\overline{T}(\widehat{X},\widehat{Z})$, $(\underline{X},\underline{Z})=\underline{T}(\widehat{X},\widehat{Z})$. We denote the law of $((\overline{X},\overline{Z}),(\underline{X},\underline{Z}))$ by $\overline{\gamma}=(\overline{T}\otimes\mathrm{id}_{\mathcal{X}\times\mathcal{Z}})_\sharp\widehat{\mathbb{P}}$, and the law of $(\overline{X},\overline{Z})$ is $\overline{\mathbb{P}}=\overline{\gamma}_{(X,Z)}$ the marginal. Similarly we define $\underline{\gamma}$ and $\underline{\mathbb{P}}$. We also define $\widehat{\gamma}=(\mathrm{id}_{\mathcal{X}\times\mathcal{Z}}\otimes\mathrm{id}_{\mathcal{X}\times\mathcal{Z}})_\sharp\widehat{\mathbb{P}}$ to be the identity transport plan. For convenience, denote the law of $(\overline{X},\widehat{X})$ to be $\overline{\gamma}_1=\overline{\gamma}_{(X,\widehat{X})}$, and the law of $(\underline{X},\widehat{X})$ to be $\underline{\gamma}_1=\underline{\gamma}_{(X,\widehat{X})}$. Similarly define $\overline{\gamma}_2=\overline{\gamma}_{(Z,\widehat{Z})|(X,\widehat{X})}$ and $\underline{\gamma}_2=\underline{\gamma}_{(Z,\widehat{Z})|(X,\widehat{X})}$ to be the conditional law of $(\overline{Z},\widehat{Z})$ and $(\underline{Z},\widehat{Z})$ given $(\overline{X},\widehat{X})$ and $(\underline{X},\widehat{X})$, respectively.

We know that $h(\lambda)$ attains its minimum $v_{\mathsf{D}}^f$ at some $\lambda^*\in\mathcal{I}$, so $h'(\lambda^*+)\ge0$ and $h'(\lambda^*-)\le0$ (if $\lambda^*>\kappa$), so

$$\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\bigg|_{\lambda=\lambda^*}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda,\widehat{X})\right]\le-\rho^p\le\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\bigg|_{\lambda=\lambda^*}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda,\widehat{X})\right]$$

where

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\bigg|_{\lambda=\lambda^*}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda,\widehat{X})\right]&=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\bigg|_{\lambda=\lambda^*}\Phi(\lambda,\widehat{X})\right]\\&\ge\mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\bigg|_{\lambda=\lambda_1}F_{(\underline{X})}(\lambda;\widehat{X})\right]-\delta\\&=\mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\bigg|_{\lambda=\lambda_1}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;\underline{X},\widehat{Z})\mid(\underline{X},\widehat{X})\right]-\lambda\|\underline{X}-\widehat{X}\|^p\right\}\right]-\delta\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Big|_{\lambda=\lambda_1}\Upsilon(\lambda;\underline{X},\widehat{Z})\mid(\underline{X},\widehat{X})\right]-\|\underline{X}-\widehat{X}\|^p\right]-\delta\\
&\geq \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Big|_{\lambda=\lambda_3}G_{(\underline{Z})}(\lambda;\underline{X},\widehat{Z})\mid(\underline{X},\widehat{X})\right]-\|\underline{X}-\widehat{X}\|^p\right]-2\delta\\
&\geq \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[-\|\underline{Z}-\widehat{Z}\|^p\mid(\underline{X},\widehat{X})\right]-\|\underline{X}-\widehat{X}\|^p\right]-2\delta\\
&= -\mathbb{E}_{((\underline{X},\underline{Z}),(\widehat{X},\widehat{Z}))\sim\underline{\gamma}}\left[\|\underline{X}-\widehat{X}\|^p+\|\underline{Z}-\widehat{Z}\|^p\right]-2\delta,
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Big|_{\lambda=\lambda^*}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda,\widehat{X})\right] &= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Big|_{\lambda=\lambda^*}\Phi(\lambda,\widehat{X})\right]\\
&\leq \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Big|_{\lambda=\lambda_2}F_{(\overline{X})}(\lambda;\widehat{X})\right]+\delta\\
&= \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Big|_{\lambda=\lambda_2}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;\overline{X},\widehat{Z})\mid(\overline{X},\widehat{X})\right]-\lambda\|\overline{X}-\widehat{X}\|^p\right\}\right]+\delta\\
&= \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Big|_{\lambda=\lambda_2}\Upsilon(\lambda;\overline{X},\widehat{Z})\mid(\overline{X},\widehat{X})\right]-\|\overline{X}-\widehat{X}\|^p\right]+\delta\\
&\leq \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\mathbb{E}_{(\overline{Z},\widehat{Z})\sim\overline{\gamma}_2}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Big|_{\lambda=\lambda_4}G_{(\overline{Z})}(\lambda;\overline{X},\widehat{Z})\mid(\overline{X},\widehat{X})\right]-\|\overline{X}-\widehat{X}\|^p\right]+2\delta\\
&\leq \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\mathbb{E}_{(\overline{Z},\widehat{Z})\sim\overline{\gamma}_2}\left[-\|\overline{Z}-\widehat{Z}\|^p\mid(\overline{X},\widehat{X})\right]-\|\overline{X}-\widehat{X}\|^p\right]+2\delta\\
&= -\mathbb{E}_{((\overline{X},\overline{Z}),(\widehat{X},\widehat{Z}))\sim\overline{\gamma}}\left[\|\underline{X}-\widehat{X}\|^p+\|\underline{Z}-\widehat{Z}\|^p\right]+2\delta,
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\overline{d} &:= \mathbb{E}_{((\overline{X},\overline{Z}),(\widehat{X},\widehat{Z}))\sim\overline{\gamma}}\left[\|\underline{X}-\widehat{X}\|^p+\|\underline{Z}-\widehat{Z}\|^p\right]\leq\rho^p+2\delta,\\
\underline{d} &:= \mathbb{E}_{((\underline{X},\underline{Z}),(\widehat{X},\widehat{Z}))\sim\underline{\gamma}}\left[\|\underline{X}-\widehat{X}\|^p+\|\underline{Z}-\widehat{Z}\|^p\right]\geq\rho^p-2\delta.
\end{aligned}
$$

Based on these, we construct a feasible primal solution. There exists $q_\delta^\varepsilon\in[0,1]$ depending on $\lambda_1,\lambda_2,\lambda_3,\lambda_4$, such that

$$
\begin{aligned}
\rho^p &= (1-q_\delta^\varepsilon)\left(\overline{d}-2\delta\right)+q_\delta^\varepsilon\left(\underline{d}+2\delta\right),\\
\rho^p+2(1-2q_\delta^\varepsilon)\delta &= (1-q_\delta^\varepsilon)\overline{d}+q_\delta^\varepsilon\underline{d}.
\end{aligned}
$$

Let $q^\delta := \frac{\rho^p}{\rho^p+2(1-2q_\delta^\varepsilon)_+\delta}\leq 1$. Define a transport plan $\gamma_\delta^\varepsilon$ by

$$
\gamma_\delta^\varepsilon := q^\delta\left[(1-q_\delta^\varepsilon)\overline{\gamma}+q_\delta^\varepsilon\underline{\gamma}\right]+(1-q^\delta)\widehat{\gamma}.
$$

Its marginal distribution $\mathbb{P}_\delta^\varepsilon = (\gamma_\delta^\varepsilon)_{(X,Z)}$ is given by

$$\mathbb{P}_\delta^\varepsilon = q^\delta \left[(1-q_\delta^\varepsilon)\overline{\mathbb{P}} + q_\delta^\varepsilon \underline{\mathbb{P}}\right] + (1-q^\delta)\widehat{\mathbb{P}}.$$

Then $\mathbb{P}_\delta^\varepsilon$ is primal feasible because

$$\begin{aligned}\mathsf{C}_p(\mathbb{P}_\delta^\varepsilon, \widehat{\mathbb{P}})^p &\le \mathbb{E}_{((X,Z),(\widehat{X},\widehat{Z}))\sim\gamma_\delta^\varepsilon}\left[\|X-\widehat{X}\|^p + \|Z-\widehat{Z}\|^p\right] \\ &\le q^\delta\left[(1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon \underline{d}\right] \le \rho^p.\end{aligned}$$

In the meantime,

$$\begin{aligned}v_{\mathsf{D}}^f - \lambda^*\rho^p &= h(\lambda^*) - \lambda^*\rho^p \\ &= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda^*, \widehat{X})\right] \\ &\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[F_{(\underline{X})}(\lambda_1;\widehat{X})\right] + \varepsilon \\ &= \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda_1;\underline{X},\widehat{Z}) \mid \widehat{X}\right] - \lambda_1\|\underline{X}-\widehat{X}\|^p\right] + \varepsilon \\ &\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[G_{(\underline{Z})}(\lambda_3;\underline{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right] - \lambda_1\|\underline{X}-\widehat{X}\|^p\right] + 2\varepsilon \\ &\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[\Psi(f(\underline{X}),\underline{Z}) - \lambda_3\|\underline{Z}-\widehat{Z}\|^p \mid (\underline{X},\widehat{X})\right] - \lambda_1\|\underline{X}-\widehat{X}\|^p\right] + 2\varepsilon \\ &\le \mathbb{E}_{((\underline{X},\underline{Z}),(\widehat{X},\widehat{Z}))\sim\underline{\gamma}}\left[\Psi(f(\underline{X}),\underline{Z}) - \lambda_3\|\underline{Z}-\widehat{Z}\|^p - \lambda_1\|\underline{X}-\widehat{X}\|^p\right] + 2\varepsilon \\ &\le \mathbb{E}_{\underline{\mathbb{P}}}\left[\Psi(f(\underline{X}),\underline{Z})\right] - \lambda_3\underline{d} + 2\varepsilon.\end{aligned}$$

Similarly

$$\begin{aligned}v_{\mathsf{D}}^f - \lambda^*\rho^p &= \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda^*, \widehat{X})\right] \\ &\le \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[F_{(\overline{X})}(\lambda_2;\widehat{X})\right] + \varepsilon \\ &= \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda_2;\overline{X},\widehat{Z}) \mid (\overline{X},\widehat{X})\right] - \lambda_2\|\overline{X}-\widehat{X}\|^p\right] + \varepsilon \\ &\le \mathbb{E}_{(\overline{X},\widehat{X})\sim\overline{\gamma}_1}\left[\mathbb{E}_{(\overline{Z},\widehat{Z})\sim\overline{\gamma}_2}\left[G_{(\overline{Z})}(\lambda_4;\overline{X},\widehat{Z}) \mid (\overline{X},\widehat{X})\right] - \lambda_2\|\overline{X}-\widehat{X}\|^p\right] + 2\varepsilon \\ &\le \mathbb{E}_{((\overline{X},\overline{Z}),(\widehat{X},\widehat{Z}))\sim\overline{\gamma}}\left[\Psi(f(\overline{X}),\overline{Z}) - \lambda_4\|\overline{Z}-\widehat{Z}\|^p - \lambda_2\|\overline{X}-\widehat{X}\|^p\right] + 2\varepsilon \\ &\le \mathbb{E}_{\overline{\mathbb{P}}}\left[\Psi(f(\overline{X}),\overline{Z})\right] - \lambda_2\overline{d} + 2\varepsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}v_{\mathsf{P}}^f &\ge \mathbb{E}_{(X,Z)\sim\mathbb{P}_\delta^\varepsilon}\left[\Psi(f(X),Z)\right] \\ &= q^\delta\left((1-q_\delta^\varepsilon)\mathbb{E}_{\overline{\mathbb{P}}}\left[\Psi(f(\overline{X}),\overline{Z})\right] + q_\delta^\varepsilon\mathbb{E}_{\underline{\mathbb{P}}}\left[\Psi(f(\underline{X}),\underline{Z})\right]\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right]\end{aligned}$$

$$
\begin{aligned}
&\geq q^\delta \left((1-q_\delta^\varepsilon)\left(v_{\mathsf{D}}^f - \lambda^*\rho^p + \lambda_2\overline{d} - 2\varepsilon\right) + q_\delta^\varepsilon\left(v_{\mathsf{D}}^f - \lambda^*\rho^p + \lambda_3\underline{d} - 2\varepsilon\right)\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] \\
&\geq q^\delta \left(v_{\mathsf{D}}^f - \lambda^*\rho^p + \lambda_3((1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon\underline{d}) - 2\varepsilon\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] \\
&\geq q^\delta \left(v_{\mathsf{D}}^f - \lambda^*\rho^p + \lambda_3(\rho^p + 2(1-2q_\delta^\varepsilon)\delta) - 2\varepsilon\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] \\
&= q^\delta \left(v_{\mathsf{D}}^f - (\lambda^*-\lambda_3)\rho^p + 2\lambda_3(1-2q_\delta^\varepsilon)\delta - 2\varepsilon\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right].
\end{aligned}
$$

As $\delta \to 0$, $q^\delta \to 1$. Thus take the limit as $\lambda_3 \to \lambda^*$ and $\delta \to 0$, it follows that

$$v_{\mathsf{P}}^f \geq v_{\mathsf{D}}^f - 2\varepsilon.$$

Since $\varepsilon$ can be taken arbitrarily small, $v_{\mathsf{P}}^f \geq v_{\mathsf{D}}^f$.

***Case 3:*** $\lambda^* = \kappa < \infty$

In this case, we can still choose $\overline{x}, \overline{z}$, and we still have

$$F_{(\overline{x})}(\lambda_2, \widehat{x}) > \Phi(\lambda^*, \widehat{x}) - \varepsilon, \qquad G_{(\overline{z})}(\lambda_4; x, \widehat{z}) > \Upsilon(\lambda_2, x, \widehat{z}) - \varepsilon.$$

and

$$\overline{d} = \mathbb{E}_{\overline{\gamma}}\left[\|\overline{X} - \widehat{X}\|^p + \|\overline{Z} - \widehat{Z}\|^p\right] \leq \rho^p + 2\delta.$$

We separate the cases $\kappa = 0$ and $\kappa > 0$.

***Case 3.1:*** $\lambda^* = \kappa = 0$

Let $q^\delta := \frac{\rho^p}{\rho^p + 2\delta} \leq 1$. Define $\gamma_\delta^\varepsilon := q^\delta\overline{\gamma} + (1-q^\delta)\widehat{\gamma}$, then its marginal is a distribution $\mathbb{P}_\delta^\varepsilon$ given by

$$\mathbb{P}_\delta^\varepsilon := q^\delta\overline{\mathbb{P}} + (1-q^\delta)\widehat{\mathbb{P}}.$$

Then it is primal feasible because

$$\mathsf{C}_p(\mathbb{P}_\delta^\varepsilon, \widehat{\mathbb{P}})^p \leq \mathbb{E}_{\gamma_\delta^\varepsilon}\left[\|\overline{X} - \widehat{X}\|^p + \|\overline{Z} - \widehat{Z}\|^p\right] \leq q^\delta\overline{d} \leq \rho^p,$$

thus

$$
\begin{aligned}
v_{\mathsf{P}}^f &\geq \mathbb{E}_{(X,Z)\sim\mathbb{P}_\delta^\varepsilon}\left[\Psi(f(X),Z)\right] \\
&= q^\delta\mathbb{E}_{(\overline{X},\overline{Z})\sim\overline{\mathbb{P}}}\left[\Psi(f(\overline{X}),\overline{Z})\right] + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}[\Psi(f(\widehat{X}),\widehat{Z})] \\
&\geq q^\delta\left(v_{\mathsf{D}}^f - \lambda^*\rho^p + \lambda_2\overline{d} - 2\varepsilon\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}[\Psi(f(\widehat{X}),\widehat{Z})] \\
&\geq q^\delta\left(v_{\mathsf{D}}^f - 2\varepsilon\right) + (1-q^\delta)\mathbb{E}_{\widehat{\mathbb{P}}}[\Psi(f(\widehat{X}),\widehat{Z})]
\end{aligned}
$$

using $\lambda^* = 0$. Let $\delta \to 0$, $q^\delta \to 1$, we have $v_{\mathsf{P}}^f \geq v_{\mathsf{D}}^f - 2\varepsilon$, and by taking $\varepsilon \to 0$ we have $v_{\mathsf{P}}^f \geq v_{\mathsf{D}}^f$.

***Case 3.2:*** $\lambda^* = \kappa > 0$

Fix any $0 < \kappa' < \kappa$. We have

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\kappa';\widehat{X}) - \Phi(\kappa;\widehat{X})\right] = h(\kappa') - h(\kappa) = \infty. \tag{EC.9}$$

We denote

$$\mathcal{X}^*(\lambda;\widehat{x}) := \left\{x \in \mathcal{X} :\ F_{(x)}(\lambda;\widehat{x}) \geq F_{(\widehat{x})}(\lambda;\widehat{x})\right\}.$$

Then $\mathcal{X}^*(\lambda;\widehat{x})$ is nonempty because $\widehat{x} \in \mathcal{X}^*(\lambda;\widehat{x})$. Since

$$\Phi(\kappa';\widehat{x}) = \sup_{x\in\mathcal{X}} F_{(x)}(\kappa';\widehat{x}) = \sup_{x\in\mathcal{X}^*(\kappa';\widehat{x})} F_{(x)}(\kappa';\widehat{x}),$$

we can rewrite (EC.9) as

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}^*(\kappa';\widehat{X})} F_{(x)}(\kappa';\widehat{X}) - \Phi(\kappa;\widehat{X})\right] = \infty.$$

Thus for any fixed $R > 0$, we can pick $\underline{X} = \underline{T}_1(\widehat{X}) \in \mathcal{X}^*(\kappa';\widehat{X})$, which induces $\underline{\gamma}_1$, such that

$$\begin{aligned} R &< \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[F_{(\underline{X})}(\kappa';\widehat{X}) - \Phi(\kappa;\widehat{X})\right] \\ &\leq \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[F_{(\underline{X})}(\kappa';\widehat{X}) - F_{(\underline{X})}(\kappa;\widehat{X})\right] \\ &= \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa;\underline{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right] + (\kappa-\kappa')\|\underline{X}-\widehat{X}\|^p\right]. \end{aligned} \tag{EC.10}$$

Moreover, because $\underline{X} \in \mathcal{X}^*(\kappa';\widehat{X})$, we have

$$\begin{aligned} F_{(\widehat{X})}(\kappa';\widehat{X}) &\leq F_{(\underline{X})}(\kappa';\widehat{X}), \\ \kappa'\|\underline{X}-\widehat{X}\|^p &\leq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa';\widehat{X},\widehat{Z}) \mid \widehat{X}\right], \\ \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\kappa'\|\underline{X}-\widehat{X}\|^p\right] &\leq \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa';\widehat{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right]\right]. \end{aligned} \tag{EC.11}$$

We denote

$$\mathcal{Z}^*(\lambda;x,\widehat{z}) := \left\{z \in \mathcal{Z} :\ G_{(z)}(\lambda;x,\widehat{z}) \geq G_{(\widehat{z})}(\lambda;x,\widehat{z})\right\}.$$

Then $\mathcal{Z}^*(\lambda;x,\widehat{z})$ is nonempty because $\widehat{z} \in \mathcal{Z}^*(\lambda;x,\widehat{z})$. Since

$$\Upsilon(\kappa';x,\widehat{z}) = \sup_{z\in\mathcal{Z}} G_{(z)}(\kappa';x,\widehat{z}) = \sup_{z\in\mathcal{Z}^*(\kappa';x,\widehat{z})} G_{(z)}(\kappa';x,\widehat{z}),$$

we can rewrite (EC.10) and (EC.11) as

$$R < \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}^*(\kappa';\underline{X},\widehat{Z})} G_{(z)}(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa;\underline{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right] + (\kappa-\kappa')\|\underline{X}-\widehat{X}\|^p\right],$$

$$\mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\kappa'\|\underline{X}-\widehat{X}\|^p\right] \le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}^*(\kappa';\underline{X},\widehat{Z})} G_{(z)}(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa';\widehat{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right]\right].$$

Thus we can pick $\underline{Z} = \underline{T}_2(\widehat{X},\widehat{Z}) \in \mathcal{Z}^*(\kappa';\underline{X},\widehat{Z})$, which induces $\gamma_2$, such that

$$\begin{aligned}
R-\varepsilon &< \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[G_{(\underline{Z})}(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa;\underline{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right] + (\kappa-\kappa')\|\underline{X}-\widehat{X}\|^p\right] \\
&\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[G_{(\underline{Z})}(\kappa';\underline{X},\widehat{Z}) - G_{(\underline{Z})}(\kappa;\underline{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right] + (\kappa-\kappa')\|\underline{X}-\widehat{X}\|^p\right] \\
&= \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[(\kappa-\kappa')\|\underline{Z}-\widehat{Z}\|^p \mid (\underline{X},\widehat{X})\right] + (\kappa-\kappa')\|\underline{X}-\widehat{X}\|^p\right] \\
&= (\kappa-\kappa')\underline{d},
\end{aligned}$$

and simultaneously ensure

$$\begin{aligned}
\mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\kappa'\|\underline{X}-\widehat{X}\|^p\right] - \delta &\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[G_{(\underline{Z})}(\kappa';\underline{X},\widehat{Z}) - \Upsilon(\kappa';\widehat{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right]\right] \\
&\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{(\underline{Z},\widehat{Z})\sim\underline{\gamma}_2}\left[G_{(\underline{Z})}(\kappa';\underline{X},\widehat{Z}) - G_{(\widehat{Z})}(\kappa';\widehat{X},\widehat{Z}) \mid (\underline{X},\widehat{X})\right]\right] \\
&= \mathbb{E}_{((\underline{X},\underline{Z}),(\widehat{X},\widehat{Z}))\sim\underline{\gamma}}\left[\Psi(f(\underline{X}),\underline{Z}) - \kappa'\|\underline{Z}-\widehat{Z}\|^p - \Psi(f(\widehat{X}),\widehat{Z})\right], \\
\kappa'\underline{d} &\le \mathbb{E}_{((\underline{X},\underline{Z}),(\widehat{X},\widehat{Z}))\sim\underline{\gamma}}\left[\Psi(f(\underline{X}),\underline{Z}) - \Psi(f(\widehat{X}),\widehat{Z})\right] \\
&\le \mathbb{E}_{\underline{\mathbb{P}}}[\Psi(f(\underline{X}),\underline{Z})] - \mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right].
\end{aligned}$$

In conclusion, we have

$$\frac{R-\varepsilon}{\kappa-\kappa'} < \underline{d} \le \frac{\mathbb{E}_{\underline{\mathbb{P}}}[\Psi(f(\underline{X}),\underline{Z})] - \mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right]}{\kappa'}.$$

We can choose $R = \varepsilon + (\kappa-\kappa')N\rho^p$ for some $N >> 1$ to be specified later. Because

$$\overline{d} - 2\delta \le \rho^p \le \frac{d}{N} \le \underline{d} + 2\delta,$$

there exists $q_\delta^\varepsilon \in [0,1]$ depending on $\lambda_2, \lambda_4, \kappa'$, such that

$$\begin{aligned}
\rho^p &= (1-q_\delta^\varepsilon)\left[\overline{d} - 2\delta\right] + q_\delta^\varepsilon\left[\underline{d} + 2\delta\right], \\
&= (1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon\underline{d} - 2(1-2q_\delta^\varepsilon)\delta, \\
\rho^p + 2(1-2q_\delta^\varepsilon)\delta &= (1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon\underline{d}.
\end{aligned}$$

Let $q^\delta := \frac{\rho^p}{\rho^p + 2(1-2q_\delta^\varepsilon)_+\delta} \le 1$. Define a distribution $\mathbb{P}_\delta^\varepsilon$ by

$$\mathbb{P}_\delta^\varepsilon := q^\delta \left[(1-q_\delta^\varepsilon)\overline{\mathbb{P}} + q_\delta^\varepsilon \underline{\mathbb{P}}\right] + \left(1-q^\delta\right)\widehat{\mathbb{P}}.$$

Then $\mathbb{P}_\delta^\varepsilon$ is primal feasible, because

$$\begin{aligned}
\mathsf{C}_p(\mathbb{P}_\delta^\varepsilon, \widehat{\mathbb{P}})^p &\le q^\delta(1-q_\delta^\varepsilon)\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\|\overline{Z}-\widehat{Z}\|^p \mid \widehat{X}\right] + \|\overline{X}-\widehat{X}\|^p\right] \\
&\quad + q^\delta q_\delta^\varepsilon \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\|\underline{Z}-\widehat{Z}\|^p \mid \widehat{X}\right] + \|\underline{X}-\widehat{X}\|^p\right] \\
&\le q^\delta\left[(1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon \underline{d}\right] \le \rho^p.
\end{aligned}$$

Therefore

$$\begin{aligned}
v_{\mathsf{P}}^f &\ge \mathbb{E}_{(X,Z)\sim\mathbb{P}_\delta^\varepsilon}\left[\Psi(f(X),Z)\right] \\
&= \mathbb{E}_{\overline{\mathbb{P}}}\left[q^\delta(1-q_\delta^\varepsilon)\Psi(f(\overline{X}),\overline{Z})\right] + \mathbb{E}_{\underline{\mathbb{P}}}\left[q^\delta q_\delta^\varepsilon \Psi(f(\underline{X}),\underline{Z})\right] + \mathbb{E}_{\widehat{\mathbb{P}}}\left[(1-q^\delta)\Psi(f(\widehat{X}),\widehat{Z})\right] \\
&\ge q^\delta(1-q_\delta^\varepsilon)\left(v_{\mathsf{D}}^f - \kappa\rho^p + \lambda_2\overline{d} - 2\varepsilon\right) + q^\delta q_\delta^\varepsilon \kappa' \underline{d} \\
&\qquad + \left(1-q^\delta+q^\delta q_\delta^\varepsilon\right)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] \\
&\ge q^\delta\kappa'\left((1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon\underline{d}\right) + q^\delta(1-q_\delta^\varepsilon)(v_{\mathsf{D}}^f - \kappa\rho^p - 2\varepsilon) \\
&\qquad + \left(1-q^\delta+q^\delta q_\delta^\varepsilon\right)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right] \\
&\ge q^\delta\kappa'(\rho^p + 2(1-2q_\delta^\varepsilon)\delta) + q^\delta(1-q_\delta^\varepsilon)(v_{\mathsf{D}}^f - \kappa\rho^p - 2\varepsilon) + \left(1-q^\delta+q^\delta q_\delta^\varepsilon\right)\mathbb{E}_{\widehat{\mathbb{P}}}\left[\Psi(f(\widehat{X}),\widehat{Z})\right].
\end{aligned}$$

As $\delta \to 0$, we have $q^\delta \to 1$. Moreover, because

$$\rho^p + 2\delta \ge (1-q_\delta^\varepsilon)\overline{d} + q_\delta^\varepsilon\underline{d} \ge q_\delta^\varepsilon\underline{d} \ge q_\delta^\varepsilon N\rho^p,$$

we know that $q_\delta^\varepsilon \le \frac{1+2\delta\rho^{-p}}{N} \to 0$ as $N\to\infty$ and $\delta\to 0$. Therefore, by taking these limits, we have

$$v_{\mathsf{P}}^f \ge \kappa'\rho^p + v_{\mathsf{D}}^f - \kappa\rho^p - 2\varepsilon = v_{\mathsf{D}}^f - 2\varepsilon - (\kappa-\kappa')\rho^p.$$

Since this is true for any $\kappa' < \kappa$ and $\varepsilon > 0$, we may take $\kappa' \to \kappa$ and $\varepsilon \to 0$ so $v_{\mathsf{P}}^f \ge v_{\mathsf{D}}^f$. □

*Proof of Theorem 2.* Since $\Psi(f(\cdot),\cdot)$ is upper semicontinuous, we know that for each fixed $x\in\mathcal{X}$, $\widehat{z}\in\mathcal{Z}$, $\lambda > \kappa$, $G_{(z)}(\lambda; x, \widehat{z}) = \Psi(f(x),z) - \lambda\|z-\widehat{z}\|^p$ is upper semicontinuous in $z$. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}G_{(z)}(\lambda;x,\widehat{z}) = -\|z-\widehat{z}\|^p \to -\infty \qquad \text{as } |z|\to\infty,$$

By Lemma EC.4 (II), we can find $\overline{z}$, $\underline{z}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Upsilon(\lambda;x,\widehat{z})=-\|\overline{z}-\widehat{z}\|^p,\qquad \frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Upsilon(\lambda;x,\widehat{z})=-\|\underline{z}-\widehat{z}\|^p,\qquad \Upsilon(\lambda;x,\widehat{z})=G_{(\overline{z})}(\lambda;x,\widehat{z})=G_{(\underline{z})}(\lambda;x,\widehat{z}).$$

Now we claim that for each fixed $\widehat{z}\in\mathcal{Z}$, $\lambda>\kappa$, $\Upsilon(\lambda;x,\widehat{z})$ is upper semicontinuous in $x$. We prove it by contradiction. Assume otherwise, then we can find $x_k\to x$, such that

$$\Upsilon(\lambda;x_k,\widehat{z})>\Upsilon(\lambda;x,\widehat{z})+\varepsilon$$

for all $k$. We can find $\underline{z}_k$ such that

$$\Upsilon(\lambda;x_k,\widehat{z})=G_{(\underline{z}_k)}(\lambda;x_k,\widehat{z}),\qquad \frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Upsilon(\lambda;x,\widehat{z})=-\|\underline{z}_k-\widehat{z}\|^p.$$

If $\underline{z}_k$ is bounded, then up to a subsequence it converges to $\underline{z}_\infty$, and since $G$ is upper semicontinuous,

$$\limsup_{k\to\infty}\Upsilon(\lambda;x_k,\widehat{z})=\limsup_{k\to\infty}G_{(\underline{z}_k)}(\lambda;x_k,\widehat{z})\le G_{(\underline{z}_\infty)}(\lambda;x,\widehat{z})\le\Upsilon(\lambda;x,\widehat{z})$$

which is a contradiction. If $\underline{z}_k$ is unbounded, then up to a subsequence, for $\lambda'\in(\kappa,\lambda)$,

$$\begin{aligned}\Upsilon(\lambda';x_k,\widehat{z})&\ge\Upsilon(\lambda;x_k,\widehat{z})-(\lambda-\lambda')\frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Upsilon(\lambda;x_k,\widehat{z})\\&\ge\Upsilon(\lambda;x,\widehat{z})+\varepsilon+(\lambda-\lambda')\|\underline{z}_k-\widehat{z}\|^p\to\infty\end{aligned}$$

as $k\to\infty$. Therefore

$$\begin{aligned}\lim_{k\to\infty}F_{(x_k)}(\lambda',\widehat{x})&=\lim_{k\to\infty}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;x_k,\widehat{Z})\mid\widehat{X}=\widehat{x}\right]-\lambda'\|x_k-\widehat{x}\|^p\\&=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\lim_{k\to\infty}\Upsilon(\lambda;x_k,\widehat{Z})\mid\widehat{X}=\widehat{x}\right]-\lambda'\|x-\widehat{x}\|^p=\infty.\end{aligned}$$

This contradicts with $\Phi(\lambda',\widehat{x})<\infty$.

We can thus construct $\overline{Z},\underline{Z}$ which depends on $\lambda$, $\widehat{Z}$ and $x$. Now we have

$$F_{(x)}(\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\lambda;x,\widehat{Z})\mid\widehat{X}=\widehat{x}\right]-\lambda\|x-\widehat{x}\|^p.$$

It is upper semicontinuous in $x$ because each $\Upsilon(\lambda;x,\widehat{z})$ is upper semicontinuous in $x$, and the finite sum of upper semicontinuous functions is upper semicontinuous. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda^+}F_{(x)}(\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Upsilon(\lambda;x,\widehat{Z})\mid\widehat{X}=\widehat{x}\right]-\|x-\widehat{x}\|^p=-\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\|\overline{Z}-\widehat{Z}\|^p\mid\widehat{X}=\widehat{x}\right]-\|x-\widehat{x}\|^p\to-\infty$$

as $x \to \infty$. By Lemma EC.4 (II) we can find $\overline{x}$ and $\underline{x}$ such that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Phi(\lambda;\widehat{x}) = -\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\|\overline{Z}-\widehat{Z}\|^p \mid \widehat{X}=\widehat{x}\right] - \|\overline{x}-\widehat{x}\|^p, \qquad \frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Phi(\lambda;\widehat{x}) = -\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\|\underline{Z}-\widehat{Z}\|^p \mid \widehat{X}=\widehat{x}\right] - \|\underline{x}-$$
$$\Phi(\lambda;\widehat{x}) = F_{(\underline{x})}(\lambda;\widehat{x}) = F_{(\overline{x})}(\lambda;\widehat{x}).$$

By constructing these for every $\widehat{x}$ in the support of $\widehat{\mathbb{P}}_{\widehat{X}}$, we have $\overline{X}$, $\underline{X}$, $\overline{Z}$, $\underline{Z}$ such that $((\overline{X},\overline{Z}),(\widehat{X},\widehat{Z})) \sim \overline{\gamma}$, $((\underline{X},\underline{Z}),(\widehat{X},\widehat{Z})) \sim \underline{\gamma}$, where

$$\overline{\gamma} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \widehat{p}_{ki}\delta_{((\overline{x}_k,\overline{z}_{ki}),(\widehat{x}_k,\widehat{z}_{ki}))}, \qquad \underline{\gamma} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \widehat{p}_{ki}\delta_{((\underline{x}_k,\underline{z}_{ki}),(\widehat{x}_k,\widehat{z}_{ki}))}.$$

We use notations $\overline{\gamma}_1, \underline{\gamma}_1, \overline{\gamma}_2, \underline{\gamma}_2$ similar as in the proof of Theorem 1.

Now we have both

$$\begin{aligned}
h(\lambda) &= \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\lambda;\widehat{X})\right] \\
&= \lambda\rho^p + \mathbb{E}_{\overline{\gamma}_1}\left[F_{(\overline{X})}(\lambda;\widehat{X})\right] \\
&= \lambda\rho^p + \mathbb{E}_{\overline{\gamma}_1}\left[\mathbb{E}_{\overline{\gamma}_2}\left[\Upsilon(\lambda;\overline{X},\widehat{Z}) \mid (\overline{X},\widehat{X})\right] - \lambda\|\overline{X}-\widehat{X}\|^p\right] \\
&= \lambda\rho^p + \mathbb{E}_{\overline{\gamma}_1}\left[\mathbb{E}_{\overline{\gamma}_2}\left[G_{(\overline{Z})}(\lambda;\overline{X},\widehat{Z}) \mid (\overline{X},\widehat{X})\right] - \lambda\|\overline{X}-\widehat{X}\|^p\right] \\
&= \lambda\rho^p + \mathbb{E}_{\overline{\gamma}_1}\left[\mathbb{E}_{\overline{\gamma}_2}\left[\Psi(f(\overline{X}),\overline{Z}) - \lambda\|\overline{Z}-\widehat{Z}\|^p \mid (\overline{X},\widehat{X})\right] - \lambda\|\overline{X}-\widehat{X}\|^p\right] \\
&= \lambda\left(\rho^p - \overline{d}\right) + \mathbb{E}_{\overline{\mathbb{P}}}\left[\Psi(f(\overline{X}),\overline{Z})\right], \\
h(\lambda) &= \lambda\left(\rho^p - \underline{d}\right) + \mathbb{E}_{\underline{\mathbb{P}}}\left[\Psi(f(\underline{X}),\underline{Z})\right],
\end{aligned}$$

and

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda^+}h(\lambda) &= \rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Phi(\lambda;\widehat{X})\right] \\
&= \rho^p + \mathbb{E}_{\overline{\gamma}_1}\left[-\mathbb{E}_{\overline{\gamma}_2}\left[\|\overline{Z}-\widehat{Z}\|^p \mid (\overline{X},\widehat{X})\right] - \|\overline{X}-\widehat{X}\|^p\right] \\
&= \rho^p - \overline{d}, \\
\frac{\mathrm{d}}{\mathrm{d}\lambda^-}h(\lambda) &= \rho^p - \underline{d}.
\end{aligned}$$

At $\lambda = \lambda^*$, $h$ is minimized, so $\frac{\mathrm{d}}{\mathrm{d}\lambda^-}h(\lambda^*) \le 0 \le \frac{\mathrm{d}}{\mathrm{d}\lambda^+}h(\lambda^*)$. Therefore there exists $q^* \in [0,1]$, such that

$$q^*\left(\rho^p - \overline{d}\right) + (1-q^*)\left(\rho^p - \underline{d}\right) = 0.$$

Then if we denote $\gamma^* = q^*\overline{\gamma} + (1-q^*)\underline{\gamma}$, then

$$\mathbb{E}_{((X,Z),(\widehat{X},\widehat{Z}))\sim\gamma^*}\left[\|X-\widehat{X}\|^p + \|Z-\widehat{Z}\|^p\right] = q^*\overline{d} + (1-q^*)\underline{d} = \rho^p.$$

Therefore, $\mathbb{P}^* = \gamma^*_{(X,Z)} = q^*\overline{\mathbb{P}} + (1-q^*)\underline{\mathbb{P}}$ is feasible, and

$$\mathbb{E}_{\mathbb{P}^*}[\Psi(f(X),Z)] = q^*\mathbb{E}_{\overline{\mathbb{P}}}[\Psi(f(\overline{X}),\overline{Z})] + (1-q^*)\mathbb{E}_{\underline{\mathbb{P}}}[\Psi(f(\overline{X}),\overline{Z})] = h(\lambda^*) = v^f_{\mathsf{D}} = v^f_{\mathsf{P}}$$

it is optimal.

Note that this optimal solution is

$$\mathbb{P}^* = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}\left(q^*\delta_{(\overline{x}_k,\overline{z}_{ki})} + (1-q^*)\delta_{(\underline{x}_k,\underline{z}_{ki})}\right).$$

Now we first consider the following linear optimization problem,

$$\begin{aligned}
\sup_{\{q_k\}_k\subset[0,1]} \quad & \mathbb{E}_{(X,Z)\sim\mathbb{P}}[\Psi(f(X),Z)] \\
& \text{where } \mathbb{P} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}\left(q_k\delta_{(\overline{x}_k,\overline{z}_{ki})} + (1-q_k)\delta_{(\underline{x}_k,\underline{z}_{ki})}\right), \\
\text{s.t.} \quad & \mathbb{E}_{((X,Z),(\widehat{X},\widehat{Z}))\sim\gamma}\left[\|X-\widehat{X}\|^p + \|Z-\widehat{Z}\|^p\right] \le \rho^p \\
& \text{where } \gamma = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}\left(q_k\delta_{((\overline{x}_k,\overline{z}_{ki}),(\widehat{x}_k,\widehat{z}_{ki}))} + (1-q_k)\delta_{((\underline{x}_k,\underline{z}_{ki}),(\widehat{x}_k,\widehat{z}_{ki}))}\right).
\end{aligned}$$

The feasible domain is not empty because $q_k = q^*$ gives a feasible solution $\mathbb{P}^*$. The constraints and the target function are all linear functions of $q_k$, so the inf can be attained at the vertices of the feasible domain, and thus we can find $k_0$ such that $q_k = 1$ or 0 whenever $k \neq k_0$. So, we have found another optimal solution

$$\mathbb{P} = \sum_{k\neq k_0}\sum_{i=1}^{n_k}\widehat{p}_{ki}\delta_{(x^*_k,z^*_{ki})} + \sum_{i=1}^{n_{k_0}}\widehat{p}_{i_0 j}\left(q\delta_{(\overline{x}_{k_0},\overline{z}_{k_0 i})} + (1-q)\delta_{(\underline{x}_{k_0},\underline{z}_{k_0 i})}\right).$$

where $(x^*_k, z^*_{ki}) = (\overline{x}_k,\overline{z}_{ki})$ or $(\underline{x}_k,\underline{z}_{ki})$ depending only on $k$. Note that the marginal $\mathbb{P}_X$ is supported over at most $I+1$ points. □

## EC.4. Proofs for Section 4

*Proof of Corollary 1.* Since $\Psi(\cdot,z)$ is affine for each $z$, $\Psi$ can be written as

$$\Psi(w,z) = \ell^z(w), \qquad \ell^z(w) = \beta^{z\top}w + b^z.$$

Here $\ell^z$ is an affine function with gradient $\beta^z \in \mathcal{D}^*$ and intercept $b^z \in \mathbb{R}$. Then

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(w,\widehat{Z}) \mid \widehat{X}=\widehat{x}_k\right] = \frac{1}{\sum_{i=1}^{n_k}\widehat{p}_{ki}}\sum_{i=1}^{n_k}\widehat{p}_{ki}\Psi(w,\widehat{z}_{ki}) = \frac{1}{\widehat{p}_k}\sum_{i=1}^{n_k}\widehat{p}_{ki}\ell^{\widehat{z}_{ki}}(w)$$

Denote

$$\beta_k := \frac{1}{\widehat{p}_k}\sum_{i=1}^{n_k}\widehat{p}_{ki}\beta^{\widehat{z}_{ki}}, \qquad\qquad b_k := \frac{1}{\widehat{p}_k}\sum_{i=1}^{n_k}\widehat{p}_{ki}b^{\widehat{z}_{ki}},$$

and

$$\ell_k(w) := \frac{1}{\widehat{p}_k}\sum_{i=1}^{n_k}\widehat{p}_{ki}\ell^{\widehat{z}_{ki}}(w) = \beta_k^\top w + b_k, \tag{EC.12}$$

which is an affine function of $w$. Therefore, $\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(w,\widehat{Z}) \mid \widehat{X}=\widehat{x}_k\right] = \ell_k(w)$ is affine. We have

$$\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(f(x),\widehat{Z}) \mid \widehat{X}=\widehat{x}_k\right] - \lambda\|x-\widehat{x}_k\|^p\right\} = \sup_{x\in\mathcal{X}}\left\{\ell_k(f(x)) - \lambda\|x-\widehat{x}_k\|^p\right\}.$$

Suppose $f:\mathcal{X}\to\mathcal{D}$ is an affine decision rule, then $f(x)=B^\top x+\delta$, and

$$\ell_k(f(x)) - \ell_k(f(\widehat{x}_k)) = \beta_k^\top(f(x)-f(\widehat{x}_k)) = \beta_k^\top B^\top(x-\widehat{x}_k).$$

Thus, the supremum over $x$ can be computed explicitly as

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left\{\ell_k(f(x)) - \lambda\|x-\widehat{x}_k\|^p\right\} &= \ell_k(f(\widehat{x}_k)) + \sup_{x\in\mathcal{X}}\left\{(B\beta_k)^\top(x-\widehat{x}_k) - \lambda\|x-\widehat{x}_k\|^p\right\}\\ &= \ell_k(f(\widehat{x}_k)) + \sup_{t\geq 0}\left\{\|B\beta_k\|_* t - \lambda t^p\right\}.\end{aligned}$$

Define a convex function $R_p:\mathbb{R}_+^2\to\mathbb{R}\cup\{+\infty\}$ by

$$R_p(\lambda,\mu) := \sup_{t\geq 0}\left\{\mu t - \lambda t^p\right\} = \begin{cases}\infty\mathbf{1}\{\lambda<\mu\}, & p=1,\\ \lambda(p-1)\left(\frac{\mu}{\lambda p}\right)^{\frac{p}{p-1}}, & p>1.\end{cases} \tag{EC.13}$$

Then

$$\sup_{x\in\mathcal{X}}\left\{\ell_k(f(x)) - \lambda\|x-\widehat{x}_k\|^p\right\} = \ell_k(f(\widehat{x}_k)) + R_p\left(\lambda, |B\beta_k|\right),$$

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(f(x),\widehat{Z}) \mid \widehat{X}\right] - \lambda\|x-\widehat{X}\|^p\right\}\right] = \sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\widehat{p}_{ki}\right)\left[\ell_k(f(\widehat{x}_k)) + R_p\left(\lambda, \|B\beta_k\|_*\right)\right].$$

Note that $R_p$ is a convex function in $\lambda$ and $B$, $\ell_k(f(\widehat{x}_k)) = \ell_k(B^\top \widehat{x}_k + \delta)$ is affine in $B$ and $\delta$, so the right-hand side of the last expression is convex in $\lambda$ and $B$ as well. Hence (6) is a convex program:

$$\inf_{\lambda\geq 0,(B,\delta)\in\Theta} \left\{ \lambda\rho^p + \sum_{k=1}^K \widehat{p}_k \left[ \ell_k(B^\top \widehat{x}_k + \delta)) + R_p\left(\lambda, \|B\beta_k\|_*\right)\right] \right\},$$

where $\ell_k$ is an affine function defined by (EC.12). □

*Proof of Corollary 2.* We start with sup over $z$:

$$\begin{aligned}\sup_{z\in\mathcal{Z}} \left\{ \Psi(w,z) - \lambda\|z-\widehat{z}_{ki}\|^2 \right\} &= \Psi(w,\widehat{z}_{ki}) + \sup_{z\in\mathcal{Z}} \left\{ (A^\top w+\alpha)^\top (z-\widehat{z}_{ki}) - \lambda\|z-\widehat{z}_{ki}\|^2 \right\} \\ &= \Psi(w,\widehat{z}_{ki}) + \sup_{\tilde{z}\in\mathcal{Z}} \left\{ (A^\top w+\alpha)^\top \tilde{z} - \lambda\|\tilde{z}\|^2 \right\}.\end{aligned}$$

By the linearity of $\Psi$ in $z$,

$$\begin{aligned}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[ \sup_{z\in\mathcal{Z}} \left\{ \Psi(w,z) - \lambda\|z-\widehat{Z}\|^2 \right\} \mid \widehat{X} = \widehat{x}_k \right] &= \Psi(w, \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z} \mid \widehat{X} = \widehat{x}_k]) + \sup_{\tilde{z}\in\mathcal{Z}} \left\{ (A^\top w+\alpha)^\top \tilde{z} - \lambda\|\tilde{z}\|^2 \right\} \\ &= \Psi(w,\bar{z}_k) + \sup_{\tilde{z}\in\mathcal{Z}} \left\{ (A^\top w+\alpha)^\top \tilde{z} - \lambda\|\tilde{z}\|^2 \right\} \\ &= \sup_{\tilde{z}\in\mathcal{Z}} \left\{ \Psi(w,\bar{z}_k+\tilde{z}) - \lambda\|\tilde{z}\|^2 \right\} \\ &= \sup_{z\in\mathcal{Z}} \left\{ \Psi(w,z) - \lambda\|z-\bar{z}_k\|^2 \right\}.\end{aligned}$$

where we define $\bar{z}_k = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z} \mid \widehat{X} = \widehat{x}_k]$. Next, we take the supremum in $x$ with decision $w = f(x) = B^\top x + \delta$. Note that

$$\Psi(w,z) = \begin{pmatrix} w^\top & 1 \end{pmatrix} \begin{pmatrix} A & \beta \\ \alpha^\top & b \end{pmatrix} \begin{pmatrix} z \\ 1 \end{pmatrix} \qquad w = \begin{pmatrix} B^\top & \delta \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} \quad \Longrightarrow \quad \Psi(f(x),z) = \begin{pmatrix} x^\top & 1 \end{pmatrix} \begin{pmatrix} B & \mathbf{0} \\ \delta^\top & 1 \end{pmatrix} \begin{pmatrix} A & \beta \\ \alpha^\top & b \end{pmatrix}$$

We thus express the supremum in $x$ by

$$\begin{aligned} y_k &:= \sup_{x\in\mathcal{X},z\in\mathcal{Z}} \Psi(w,z) - \lambda\|z-\bar{z}_k\|^2 - \lambda\|x-\widehat{x}_k\|^2 \\ &= \sup_{x\in\mathcal{X},z\in\mathcal{Z}} \begin{pmatrix} x^\top & z^\top & 1 \end{pmatrix} \begin{pmatrix} B & \mathbf{0} \\ O & \mathbf{0} \\ \delta^\top & 1 \end{pmatrix} \begin{pmatrix} O & A & \beta \\ \mathbf{0}^\top & \alpha^\top & b \end{pmatrix} \begin{pmatrix} x \\ z \\ 1 \end{pmatrix} - \lambda \begin{pmatrix} x^\top & z^\top & 1 \end{pmatrix} \begin{pmatrix} I & O & -\widehat{x}_k \\ O & I & -\bar{z}_k \\ -\widehat{x}_k^\top & -\bar{z}_k^\top & \|\widehat{x}_k\|^2 + \|\bar{z}_k\|^2 \end{pmatrix} \begin{pmatrix} x \\ z \\ 1 \end{pmatrix}. \end{aligned}$$

We have transformed (D) into

$$\begin{aligned} \inf_{\substack{(B,\delta)\in\Theta \\ \lambda\geq 0,\{y_k\}_k\subset\mathbb{R}}} \quad & \lambda\rho^2 + \sum_{k=1}^K \widehat{p}_k y_k \\ \text{s.t.} \quad & X_k \succeq O \end{aligned}$$

where

$$X_k := \lambda \begin{pmatrix} I & O & -\widehat{x}_k \\ O & I & -\bar{z}_k \\ -\widehat{x}_k^\top & -\bar{z}_k^\top & \|\bar{z}_k\|^2 + \|\widehat{x}_k\|^2 \end{pmatrix} + y_k \begin{pmatrix} O & O & \mathbf{0} \\ O & O & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{0}^\top & 1 \end{pmatrix} - \frac{1}{2}\left[ \begin{pmatrix} B & \mathbf{0} \\ O & \mathbf{0} \\ \delta^\top & 1 \end{pmatrix} \begin{pmatrix} O & A & \beta \\ \mathbf{0}^\top & \alpha^\top & b \end{pmatrix} + \begin{pmatrix} O & \mathbf{0} \\ A^\top & \alpha \\ \beta^\top & b \end{pmatrix} \begin{pmatrix} B^\top & O & \delta \\ \mathbf{0}^\top & \mathbf{0}^\top & 1 \end{pmatrix} \right]$$

$$= \begin{pmatrix} \lambda I & -\frac{1}{2}BA & -\frac{1}{2}B\beta - \lambda\widehat{x}_k \\ -\frac{1}{2}(BA)^\top & \lambda I & -\frac{1}{2}(A^\top\delta + \alpha) - \lambda\bar{z}_k \\ -\frac{1}{2}(B\beta)^\top - \lambda\widehat{x}_k^\top & -\frac{1}{2}(A^\top\delta+\alpha)^\top - \lambda\bar{z}_k^\top & y_k - \beta^\top\delta - b + \lambda\|\bar{z}_k\|^2 + \lambda\|\widehat{x}_k\|^2 \end{pmatrix} \succeq O. \tag{EC.14}$$

Since $X_k$ is affine in $\lambda, y_k, B, \delta$, this is a semidefinite program. □

*Proof of Corollary 3.* We first study the dual formulation. For $\Psi(w,z) = \|w-z\|^2$, it holds that

$$\sup_{z\in\mathcal{Z}}\{\|w-z\|^2 - \lambda\|z-\widehat{z}\|^2\} = \begin{cases} \frac{\lambda}{\lambda-1}\|w-\widehat{z}\|^2 & \lambda > 1 \\ 0 & \lambda = 1 \text{ and } w = \widehat{z} \\ \infty & \text{otherwise.} \end{cases}$$

Restricting to $\lambda > 1$, the conditional expectation is

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\{\|w-z\|^2 - \lambda\|z-\widehat{Z}\|^2\} \mid \widehat{X}\right] = \frac{\lambda}{\lambda-1}\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\|w-\widehat{Z}\|^2 \mid \widehat{X}\right] = \frac{\lambda}{\lambda-1}\left(\left\|w - \mathbb{E}[\widehat{Z}\mid\widehat{X}]\right\|^2 + \mathbb{V}\mathrm{ar}[\widehat{Z}\mid\widehat{X}]\right)$$

With affine decision rule, $w = f(x) = B^\top x + \delta$. So

$$\begin{aligned} &\sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\{\|B^\top x+\delta-z\|^2 - \lambda\|z-\widehat{Z}\|^2\} \mid \widehat{X}\right] - \lambda\|x-\widehat{X}\|^2\right\} \\ &= \sup_{x\in\mathcal{X}}\left\{\frac{\lambda}{\lambda-1}\left(\left\|B^\top x+\delta-\mathbb{E}[\widehat{Z}\mid\widehat{X}]\right\|^2 + \mathbb{V}\mathrm{ar}[\widehat{Z}\mid\widehat{X}]\right) - \lambda\|x-\widehat{X}\|^2\right\}. \end{aligned}$$

For each $\widehat{x}_k$, define $y_k$ by

$$\begin{aligned} \lambda y_k &= \sup_{x\in\mathcal{X}}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\{\Psi(f(x),z) - \lambda\|z-\widehat{Z}\|^2\} \mid \widehat{X} = \widehat{x}_k\right] - \lambda\|x-\widehat{x}_k\|^2\right\} \\ &= \sup_{x\in\mathcal{X}}\left\{\frac{\lambda}{\lambda-1}\left(\left\|B^\top x+\delta-\mathbb{E}[\widehat{Z}\mid\widehat{X}=\widehat{x}_k]\right\|^2 + \mathbb{V}\mathrm{ar}[\widehat{Z}\mid\widehat{X}=\widehat{x}_k]\right) - \lambda\|x-\widehat{x}_k\|^2\right\}. \end{aligned}$$

Denote $\mu_k = \mathbb{E}[\widehat{Z}\mid\widehat{X}=\widehat{x}_k]$ and $\sigma_k = \mathbb{V}\mathrm{ar}[\widehat{Z}\mid\widehat{X}=\widehat{x}_k]$. The dual problem can thus be written as

$$\begin{aligned} \inf_{\substack{(B,\delta)\in\Theta \\ \lambda\geq 1 \\ \{y_k\}\in\mathbb{R}^K}} \quad & \lambda\rho^2 + \lambda\sum_k \widehat{p}_k y_k \\ \text{s.t.} \quad & (\lambda-1)y_k \geq \left\|B^\top x+\delta-\mu_k\right\|^2 + \sigma_k - (\lambda-1)\|x-\widehat{x}_k\|^2, \qquad \forall x\in\mathbb{R}^d, k\in[K]. \end{aligned}$$

The constraint can be written as a quadratic form of $x$, so it is equivalent to

$$\begin{pmatrix} (\lambda-1)I - BB^\top & -(\lambda-1)\widehat{x}_k + B(\mu_k-\delta) \\ -(\lambda-1)\widehat{x}_k^\top + (\mu_k-\delta)^\top B^\top & (\lambda-1)(y_k+\|\widehat{x}_k\|^2) - \|\delta-\mu_k\|^2 - \sigma_k \end{pmatrix} \succeq O \qquad k\in[K].$$

Using the standard lifting method, we rewrite it as

$$\begin{pmatrix} (\lambda-1)I - W & -(\lambda-1)\widehat{x}_k + B\mu_k - u \\ -(\lambda-1)\widehat{x}_k^\top + \mu_k^\top B^\top - u^\top & (\lambda-1)(y_k+\|\widehat{x}_k\|^2) - t + 2\delta^\top\mu_k - \|\mu_k\|^2 - \sigma_k \end{pmatrix} \succeq O \qquad \forall k\in[K],$$

$$\begin{pmatrix} W - BB^\top & u - B\delta \\ u^\top - (B\delta)^\top & t - \delta^\top\delta \end{pmatrix} \succeq O$$

for some $W\in\mathbb{R}^{d\times d}$, $u\in\mathbb{R}^d$, $t\in\mathbb{R}$. The latter constraint is equivalent to

$$\begin{pmatrix} W & B & u \\ B^\top & I & \delta \\ u^\top & \delta^\top & t \end{pmatrix} \succeq O,$$

hence $t\geq 0$. The proof is thus completed. □

*Proof of Corollary 4.* Using Theorem 1, we can write (9) in the dual form:

$$v_{\mathsf{D}} = \inf_{\lambda\geq 0, \mathbb{h}\in\mathbb{H}} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\Phi}\sim\widehat{\mathbb{Q}}_\Phi} \left[ \sup_{\phi\in\mathbb{H}^m} \left\{ \mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}} \left[ \Psi(\langle \mathbb{h}, \phi\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi} \right] - \lambda\|\phi - \widehat{\Phi}\| \right\} \right] \right\}.$$

Here, we used the fact that the metric on $\mathcal{Z}$ is the infinity metric. Since $\Psi(\cdot, z)$ is Lipschitz and convex, so is $\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}[\Psi(\cdot,\widehat{Z}) \mid \widehat{\Phi}=\widehat{\phi}]$ for $\widehat{\mathbb{Q}}_{\widehat{\Phi}}$-a.e. $\widehat{\phi}$. Moreover, $\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}}$ is a bounded linear map, so $\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}[\Psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi}=\widehat{\phi}]$ is also a convex Lipschitz function for $\widehat{\mathbb{Q}}_{\widehat{\Phi}}$-a.e. $\widehat{\phi}$.. Therefore

$$\begin{aligned} &\sup_{\phi\in\mathbb{H}^m} \left\{ \mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}} \left[ \Psi(\langle \mathbb{h}, \phi\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi} \right] - \lambda\|\phi - \widehat{\Phi}\| \right\} \\ &\qquad = \mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}} \left[ \Psi(\langle \mathbb{h}, \widehat{\Phi}\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi} \right] + \infty \mathbf{1}\left\{ \lambda < \|\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}[\Psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi}]\|_{\mathrm{Lip}} \right\}. \end{aligned}$$

We now compute this Lipschitz norm. We claim for any convex Lipschitz function $\psi:\mathbb{R}^m\to\mathbb{R}$, it holds that

$$\|\psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}})\|_{\mathrm{Lip}(\mathbb{H}^m)} = \|\mathbb{h}\|_{\mathbb{H}}\|\psi\|_{\mathrm{Lip}(\mathbb{R}^m)}.$$

It is trivial to see that the equality holds when $\|\mathbb{h}\|_{\mathbb{H}}=0$, and that the left-hand side is bounded by the right-hand side. To see that this bound is achievable, we first observe that

$\langle \mathbb{h}, a\mathbb{h}\rangle = \|\mathbb{h}\|_{\mathbb{H}}^2 a$ for any $a \in \mathbb{R}^m$, where $a\mathbb{h} = \{a_i \mathbb{h}\}_{i\in[m]}$. We now take any $a, b \in \mathbb{R}^m$ and $a \neq b$, then

$$\frac{|\psi(a)-\psi(b)|}{\|a-b\|} = \frac{\left|\psi\left(\left\langle \mathbb{h}, \frac{a\mathbb{h}}{\|\mathbb{h}\|_{\mathbb{H}}^2}\right\rangle\right) - \psi\left(\left\langle \mathbb{h}, \frac{b\mathbb{h}}{\|\mathbb{h}\|_{\mathbb{H}}^2}\right\rangle\right)\right|}{\|a-b\|} \leq \|\psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}})\|_{\mathrm{Lip}(\mathbb{H}^m)} \frac{\left\|\frac{a\mathbb{h}}{\|\mathbb{h}\|_{\mathbb{H}}^2} - \frac{b\mathbb{h}}{\|\mathbb{h}\|_{\mathbb{H}}^2}\right\|}{\|a-b\|}$$
$$= \|\psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}})\|_{\mathrm{Lip}(\mathbb{H}^m)} \frac{1}{\|\mathbb{h}\|_{\mathbb{H}}}.$$

Taking supremum over $x, y \in \mathbb{R}^m$ yields the desired equality. Combined, we can simplify $v_{\mathsf{D}}$ to

$$\begin{aligned}
v_{\mathsf{D}} &= \inf_{\lambda\geq 0, \mathbb{h}\in\mathbb{H}} \left\{\lambda\rho + \mathbb{E}_{\widehat{\Phi}\sim\widehat{\mathbb{Q}}_\Phi}\left[\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}\left[\Psi(\langle \mathbb{h}, \widehat{\Phi}\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi}\right] + \infty \mathbf{1}\left\{\lambda < \|\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}[\Psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi}]\|_{\mathrm{Lip}}\right\}\right]\right\} \\
&= \inf_{\mathbb{h}\in\mathbb{H}} \left\{\mathbb{E}_{\widehat{\Phi}\sim\widehat{\mathbb{Q}}_\Phi}\left[\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}\left[\Psi(\langle \mathbb{h}, \widehat{\Phi}\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi}\right]\right] + \rho \sup_{\widehat{\phi}\in\operatorname{supp}\widehat{\mathbb{Q}}_{\widehat{\Phi}}} \left\|\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}\left[\Psi(\langle \mathbb{h}, \cdot\rangle_{\mathbb{H}}, \widehat{Z}) \mid \widehat{\Phi} = \widehat{\phi}\right]\right\|_{\mathrm{Lip}}\right\} \\
&= \inf_{\mathbb{h}\in\mathbb{H}} \left\{\mathbb{E}_{(\widehat{\Phi},\widehat{Z})\sim\widehat{\mathbb{Q}}}\left[\Psi(\langle \mathbb{h}, \widehat{\Phi}\rangle_{\mathbb{H}}, \widehat{Z})\right] + \rho\|\mathbb{h}\|_{\mathbb{H}} \sup_{\widehat{\phi}\in\operatorname{supp}\widehat{\mathbb{Q}}_{\widehat{\Phi}}} \left\|\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{Q}}_{\widehat{Z}|\widehat{\Phi}}}\left[\Psi(\cdot, \widehat{Z}) \mid \widehat{\Phi} = \widehat{\phi}\right]\right\|_{\mathrm{Lip}}\right\} \\
&= \inf_{\mathbb{h}\in\mathbb{H}} \left\{\mathbb{E}_{(\widehat{X},\widehat{Z})\sim\widehat{\mathbb{P}}}\left[\Psi(\mathbb{h}(\widehat{X}), \widehat{Z})\right] + \rho\|\mathbb{h}\|_{\mathbb{H}} \sup_{\widehat{x}\in\operatorname{supp}\widehat{\mathbb{P}}_{\widehat{X}}} \left\|\mathbb{E}_{\widehat{Z}\sim\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(\cdot, \widehat{Z}) \mid \widehat{X} = \widehat{x}\right]\right\|_{\mathrm{Lip}}\right\}.
\end{aligned}$$

A representer theorem for vector-valued RKHS can be derived similarly to the scalar-valued version in (Shafieezadeh-Abadeh et al., 2019, theorem 27), thus the optimal $\mathbb{h}$ is given by

$$\mathbb{h}^\star = \sum_k \boldsymbol{K}(\cdot, \widehat{x}_k)\beta_k, \qquad \beta_k \in \mathbb{R}^m.$$

The proof is identical, so we omit the details. It is now a finite-dimensional convex optimization problem over $\{\beta_k\}_{k\in[K]}$. To derive (10), we compute

$$\begin{aligned}
\mathbb{h}(\widehat{x}) &= \sum_k \boldsymbol{K}(\widehat{x}, \widehat{x}_k)\beta_k, \\
\|\mathbb{h}\|_{\mathbb{H}}^2 &= \sum_{j,k\in[K]} \langle \boldsymbol{K}(\cdot, \widehat{x}_k)\beta_k, \boldsymbol{K}(\cdot, \widehat{x}_j)\beta_j\rangle = \sum_{j,k\in[K]} \beta_j^\top \boldsymbol{K}(\widehat{x}_j, \widehat{x}_k)\beta_k.
\end{aligned}$$

We thus conclude (10). □

*Proof of Theorem 3.* First, we show that $\cap_k I_k(x)$ is nonempty. To begin with, each $I_k(x)$ is nonempty, because the definition of $\phi_k$ implies

$$\varphi_k(w_k) \leq \phi_k \leq \lambda^*\|x - x_k\| + \phi_k,$$

so $w_k \in I_k(x)$. Note that each $I_k(x)$ is an interval since it is the sub-level set of a convex function $\varphi_k$. To prove they have a nonempty intersection, it suffices to show they pairwise intersect. For instance, we show here that $I_1(x)$ and $I_2(x)$ intersect by contradiction. Suppose $I_1$ and $I_2$ are disjoint. Since $w_1 \in I_1(x)$, $w_2 \in I_2(x)$, we know that $I_1$ and $I_2$ are disjoint if and only if we can find $w_3$ in between $w_1$ and $w_2$ outside both intervals. This implies that

$$\begin{aligned}
&\varphi_1(w_3) > \lambda^*\|x - x_3\| + \phi_1 \geq \lambda^*\|x - x_1\| + \varphi_1(w_1),\\
&\varphi_1(w_3) > \lambda^*\|x - x_3\| + \phi_1 \geq \lambda^*\|x - x_1\| + \varphi_1(w_2) - \lambda^*\|x_1 - x_2\|,\\
&\varphi_2(w_3) > \lambda^*\|x - x_3\| + \phi_2 \geq \lambda^*\|x - x_2\| + \varphi_2(w_2),\\
&\varphi_2(w_3) > \lambda^*\|x - x_3\| + \phi_2 \geq \lambda^*\|x - x_2\| + \varphi_2(w_1) - \lambda^*\|x_1 - x_2\|.
\end{aligned}$$

Since $w_3$ is between $w_1$ and $w_2$, we can find $\alpha, \beta \in [0,1]$ with $\alpha + \beta = 1$ and $w_3 = \alpha w_1 + \beta w_2$. By multiplying the first/fourth inequality with $\alpha$ and the second/third inequality with $\beta$ then taking the sum, we have

$$\begin{aligned}
(\varphi_1 + \varphi_2)(w_3) &> \lambda^*(\|x - x_1\| + \|x - x_2\|) + \alpha(\varphi_1 + \varphi_2)(w_1) + \beta(\varphi_1 + \varphi_2)(w_2) - \lambda^*\|x_1 - x_2\|\\
&\geq \alpha(\varphi_1 + \varphi_2)(w_1) + \beta(\varphi_1 + \varphi_2)(w_2),
\end{aligned}$$

using the triangle inequality. However, this contradicts the convexity of $\varphi_1 + \varphi_2$.

Next, we prove that any decision rule in the intersection $\cap_k I_k$ is optimal. For every $f \in \mathcal{F}$, let $\widehat{f} = f|_{\widehat{\mathcal{X}}} \in \widehat{\mathcal{F}}$ be the restriction of $f$ on the set $\widehat{X}$, then

$$\begin{aligned}
&\inf_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \sup_{x \in \mathcal{X}} \left\{ \varphi(f(x); \lambda, \widehat{X}) - \lambda\|x - \widehat{X}\| \right\} \right] \right\}\\
&\geq \inf_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \max_{x \in \widehat{\mathcal{X}}} \left\{ \varphi(f(x); \lambda, \widehat{X}) - \lambda\|x - \widehat{X}\| \right\} \right] \right\}\\
&= \inf_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[ \max_{1 \leq k \leq K} \left\{ \varphi(\widehat{f}(x_k); \lambda, \widehat{X}) - \lambda\|x_k - \widehat{X}\| \right\} \right] \right\} \geq v_{\widehat{\mathsf{D}}}.
\end{aligned} \tag{EC.15}$$

By taking the infimum over $f \in \mathcal{F}$, we would have $v_{\mathsf{D}} \geq v_{\widehat{\mathsf{D}}}$. On the other hand, for the minimizer $\lambda^*$ and $\widehat{f}^* \in \widehat{\mathcal{F}}$ of (13), let $f \in \mathcal{F}$ be an extension in $\cap_k I_k(x)$, then for every $x$ we have

$$\varphi_k(f(x)) - \lambda^*\|x - \widehat{x}\| \leq \max_{1 \leq k \leq K} \left\{ \varphi(\widehat{f}(\widehat{x}_k); \lambda^*, \widehat{x}) - \lambda^*\|x_k - \widehat{x}\| \right\}.$$

Therefore,

$$\begin{aligned}&\lambda^*\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\le k\le K}\left\{\varphi(\widehat{f}(x_k);\lambda^*,\widehat{X})-\lambda^*\|x_k-\widehat{X}\|\right\}\right]\\&\ge\lambda^*\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\varphi(f(x);\lambda^*,\widehat{X})-\lambda^*\|x-\widehat{X}\|\right\}\right]\ge v_{\mathsf{D}}.\end{aligned}$$

Thus $v_{\mathsf{D}}=v_{\widehat{\mathsf{D}}}$.

Finally, we show the necessity of the interval condition. Suppose $f^*\in\mathcal{F}$ is an optimal policy to the problem (11) with optimal dual value $\lambda^*$. By (EC.15), $\lambda^*$ and the restriction $\widehat{f}^*=f|_{\widehat{\mathcal{X}}}\in\widehat{\mathcal{F}}$ are also an optimal dual value and an optimal policy to the problem (13). To show that $f^*(x)\in\cap_k I_k(x)$, we prove by contradiction. Suppose for some $x\in\mathcal{X}$ and some $k\in[K]$, $f^*(x)\notin I_k(x)$. This means

$$\varphi(f^*(x);\lambda^*,\widehat{x}_k)=\varphi_k(f^*(x))>\lambda^*\|x-\widehat{x}_k\|+\phi_k=\lambda^*\|x-\widehat{x}_k\|+\max_j\left\{\varphi_k(w_j)-\lambda^*\|\widehat{x}_k-\widehat{x}_j\|\right\}.$$

That is, there exists $k\in[K]$ such that for all $j\in[K]$,

$$\varphi(f^*(x);\lambda^*,\widehat{x}_k)-\lambda^*\|x-\widehat{x}_k\|>\varphi(f^*(\widehat{x}_j);\lambda^*,\widehat{x}_k)-\lambda^*\|\widehat{x}_k-\widehat{x}_j\|.$$

Then

$$\begin{aligned}v_{\mathsf{D}}&=\lambda^*\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\varphi(f^*(x);\lambda^*,\widehat{X})-\lambda^*\|x-\widehat{X}\|\right\}\right]\\&>\lambda^*\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{j\in[K]}\left\{\varphi(f^*(\widehat{x}_j);\lambda^*,\widehat{X})-\lambda^*\|\widehat{x}_j-\widehat{X}\|\right\}\right]\ge v_{\widehat{\mathsf{D}}},\end{aligned}$$

which contradicts with $v_{\mathsf{D}}=v_{\widehat{\mathsf{D}}}$. Therefore, we must have $f^*(x)\in\cap_k I_k(x)$ for all $x\in\mathcal{X}$, which completes the proof of the theorem. □

*Proof of Remark 2.* Recall that we consider the linear loss function $\Psi(w,z)=w^\top z$, where the decision space $\mathcal{D}=B_1$ is the unit disk in $\mathbb{R}^2$, and $\mathcal{X}=\mathcal{Z}=\mathbb{R}^2$. Suppose $\widehat{\mathbb{P}}$ satisfies

$$\widehat{\mathbb{P}}_{\widehat{X}}=\frac{1}{3}\sum_{k=1}^{3}\delta_{\widehat{x}_k},\qquad\widehat{x}_k=\left(\cos\frac{2k\pi}{3},\sin\frac{2k\pi}{3}\right),\qquad\mathbb{E}\left[\widehat{Z}\mid\widehat{X}\right]=\frac{2\sqrt{3}}{3}\widehat{X}.$$

$\widehat{X}=\{\widehat{x}_1,\widehat{x}_2,\widehat{x}_3\}$ consists of three points on the unit circle that form an equilateral triangle (see Figure 5). Let us first solve the in-sample dual problem (13). Observe that

$$\sup_{z\in\mathcal{Z}}\{\Psi(w,z)-\lambda\|z-\widehat{z}\|\}=w^\top\widehat{z}+\infty\mathbf{1}\{\|w\|>\lambda\},$$

so we can find

$$\varphi(w;\lambda,\widehat{x}_j)=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\Psi(w,z)-\lambda\|z-\widehat{Z}\|\right\}\mid\widehat{X}=\widehat{x}_j\right]=w^\top\mathbb{E}\left[\widehat{Z}\mid\widehat{X}=\widehat{x}_j\right]+\infty\mathbf{1}\{\|w\|>\lambda\}$$
$$=\frac{2\sqrt{3}}{3}w^\top\widehat{x}_j+\infty\mathbf{1}\{\|w\|>\lambda\}.$$

(13) becomes

$$\min_{\substack{\widehat{f}\in\widehat{\mathcal{F}}\\ \lambda\geq\|\widehat{f}(\widehat{x}_k)\|}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq 3}\left\{\frac{2\sqrt{3}}{3}\widehat{f}(\widehat{x}_k)^\top\widehat{X}-\lambda\|\widehat{x}_k-\widehat{X}\|\right\}\right]\right\}$$
$$=\min_{\substack{\{w_k\}_k\in\mathcal{D}\\ \lambda\geq\|w_k\|}}\left\{\lambda\rho+\frac{1}{3}\sum_{j=1}^{3}\max_{1\leq k\leq 3}\left\{\frac{2\sqrt{3}}{3}w_k^\top\widehat{x}_j-\lambda\|\widehat{x}_k-\widehat{x}_j\|\right\}\right\}$$
$$\geq\min_{\substack{\{w_k\}_k\in\mathcal{D}\\ \lambda\geq\|w_k\|}}\left\{\lambda\rho+\frac{1}{3}\sum_{j=1}^{3}\frac{2\sqrt{3}}{3}w_k^\top\widehat{x}_j\right\}$$
$$\geq\min_{\lambda\geq 0}\left\{\lambda\rho-\frac{2\sqrt{3}}{3}(\lambda\wedge 1)\right\}.$$

The last inequality used $\|w_k\|\leq\lambda\wedge 1$ and $\|\widehat{x}_j\|=1$. Provided $0<\rho<\frac{2\sqrt{3}}{3}$, the above is minimized at $\lambda=1$, and we obtain a lower bound $\rho-\frac{2\sqrt{3}}{3}$ for (13). Note that $\|\widehat{x}_k-\widehat{x}_j\|=\sqrt{3}\mathbf{1}\{j\neq k\}$, and $\widehat{x}_k^\top\widehat{x}_j=\mathbf{1}\{j=k\}-\frac{1}{2}\mathbf{1}\{j\neq k\}$. One can directly check that this minimal value is achieved by $w_k=-\widehat{x}_k$ and $\lambda^*=1$, thus (13) is solved.

Now we show that the extension in Theorem 3 does not exist. By the definition of $I_k(x)$, we know $w\in I_k(x)$ iff

$$\varphi_k(w)-\lambda^*\|x-\widehat{x}_k\|\leq\phi_k=\max_j\{\varphi_k(w_j)-\lambda^*\|\widehat{x}_k-\widehat{x}_j\|\}.$$

With $\lambda^*=1$ and $w_j=-\widehat{x}_j$, we rewrite the above as

$$\frac{2\sqrt{3}}{3}w^\top\widehat{x}_k-\|x-\widehat{x}_k\|\leq\max_j\left\{\frac{2\sqrt{3}}{3}w_j^\top\widehat{x}_k-\|\widehat{x}_k-\widehat{x}_j\|\right\}=\max_j\left\{-\frac{2\sqrt{3}}{3}\widehat{x}_j^\top\widehat{x}_k-\|\widehat{x}_k-\widehat{x}_j\|\right\}=-\frac{2\sqrt{3}}{3}.$$

When $x=0$, $\|x-\widehat{x}_k\|=1$, so

$$I_k(0)=\left\{w\in\mathcal{D}:\frac{2\sqrt{3}}{3}w^\top\widehat{x}_k-1\leq-\frac{2\sqrt{3}}{3}\right\}=\left\{w\in\mathcal{D}:w^\top\widehat{x}_k\leq\frac{\sqrt{3}}{2}-1\right\}.$$

Since $\frac{\sqrt{3}}{2}-1<0$, no such $w$ exists in $\cap_k I_k(0)$, because $\widehat{x}_1,\widehat{x}_2,\widehat{x}_3$ form an equilateral triangle. One can also see from Figure 5 that the intersection of $I_k(x)$ is empty. □

## EC.5. Proofs for Examples in Section 4

*Proof of Example 5.* Since $f$ is real-valued and $\Psi$ is convex in $w$, we use Theorem 3, so it has the following reformulation

$$\inf_{\substack{\widehat{f}:\widehat{\mathcal{X}}\to\mathbb{R}\\ \lambda\geq 0}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq K}\left\{\varphi(\widehat{f}(\widehat{x}_k);\lambda,\widehat{X})-\lambda\|\widehat{x}_k-\widehat{X}\|\right\}\right]\right\}$$

with

$$\varphi(w;\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{|w-z|-\lambda\|z-\widehat{Z}\|\right\}\mid\widehat{X}=\widehat{x}\right].$$

For any $\lambda<1$, the supremum over $z$ is infinite, hence $\varphi(w;\lambda,\widehat{x})=\infty$. For $\lambda\geq 1$, the supremum is attained at $z=\widehat{Z}$, so

$$\varphi(w;\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[|w-\widehat{Z}|\,\middle|\,\widehat{X}=\widehat{x}\right]+\infty\mathbf{1}\{\lambda<1\}.$$

Thus, we reach the following reformulation,

$$\inf_{\substack{\widehat{f}:\widehat{\mathcal{X}}\to\mathbb{R}\\ \lambda\geq 1}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq K}\left\{\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[|\widehat{f}(\widehat{x}_k)-\widehat{Z}|\,\middle|\,\widehat{X}=\widehat{x}\right]-\lambda\|\widehat{x}_k-\widehat{X}\|\right\}\right]\right\}.$$

This can be transformed into a linear programming problem

$$\begin{aligned}\inf_{\substack{\{w_k\}_k,\{y_k\}_k\subset\mathbb{R}\\ \{c_{kji}\}_{kji}\subset\mathbb{R},\lambda\geq 1}}\quad & \lambda\rho+\sum_{k=1}^{K}y_k & & \\ \text{s.t.}\quad & y_j\geq\sum_{i=1}^{n_j}\widehat{p}_{ki}(c_{kji}-\lambda\|\widehat{x}_k-\widehat{x}_j\|) & & \forall j,k\in[K],\\ & c_{kji}\geq w_k-\widehat{z}_{ji} & & \forall k,j\in[K],i\in[n_j],\\ & c_{kji}\geq\widehat{z}_{ji}-w_k & & \forall k,j\in[K],i\in[n_j].\end{aligned}$$

□

*Proof of Example 6.* Recall that the problem could be reformulated as

$$\inf_{\substack{\widehat{f}:\widehat{\mathcal{X}}\to\mathbb{R}\\ \lambda\geq 0}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq K}\left\{\varphi(\widehat{f}(\widehat{x}_k);\lambda,\widehat{X})-\lambda\|\widehat{x}_k-\widehat{X}\|\right\}\right]\right\},$$

where

$$\varphi(w;\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{-wz^{\top}\begin{pmatrix}w\\1\end{pmatrix}-\lambda\|z-\widehat{Z}\|\right\}\mid\widehat{X}=\widehat{x}\right].$$

When $\mathcal{Z}$ is equipped with the usual $\ell^p$ norm $\|\cdot\|_{\mathcal{Z}}$, the supremum over $z$ in the definition of $\varphi$ is infinite if $\|w \begin{pmatrix} w & 1 \end{pmatrix}\|_* > \lambda$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|_{\mathcal{Z}}$, otherwise the supremum is achieved at $z = \widehat{Z}$. Therefore

$$\varphi(w;\lambda;\widehat{x}_k) = -\begin{pmatrix} w^2 & w \end{pmatrix} \bar{z}_k + \infty \mathbf{1}\{\|\begin{pmatrix} w^2 & w \end{pmatrix}\|_* > \lambda\}.$$

Hence, we obtain the reformulation (14). Recall that the first component of $z$ represents the price sensitivity coefficient, which is negative.

When $p = 1$, this can be written as the following quadratic constraint program:

$$\begin{aligned} \inf_{\{w_k\}_k, \lambda \geq 0} \quad & \lambda\rho + \sum_{j\in[K]} \widehat{p}_j c_j & \\ \text{s.t.} \quad & c_j + \begin{pmatrix} w_k^2 & w_k \end{pmatrix} \bar{z}_k + \lambda \|\widehat{x}_k - \widehat{x}_j\| \geq 0 & \forall j,k \in [K], \\ & w_k \leq \lambda & \forall k \in [K], \\ & w_k^2 \leq \lambda & \forall k \in [K]. \end{aligned}$$

When $p = \infty$, this can also be written as a quadratic constraint program:

$$\begin{aligned} \inf_{\{w_k\}_k, \lambda \geq 0} \quad & \lambda\rho + \sum_{j\in[K]} \widehat{p}_j c_j & \\ \text{s.t.} \quad & c_j + \begin{pmatrix} w_k^2 & w_k \end{pmatrix} \bar{z}_k + \lambda \|\widehat{x}_k - \widehat{x}_j\| \geq 0 & \forall j,k \in [K], \\ & w_k^2 + w_k \leq \lambda & \forall k \in [K]. \end{aligned}$$

When $p = 2$, this is written as

$$\begin{aligned} \inf_{\{w_k\}_k, \lambda \geq 0} \quad & \lambda\rho + \sum_{j\in[K]} \widehat{p}_j c_j & \\ \text{s.t.} \quad & c_j + \begin{pmatrix} w_k^2 & w_k \end{pmatrix} \bar{z}_k + \lambda \|\widehat{x}_k - \widehat{x}_j\| \geq 0 & \forall j,k \in [K], \\ & w_k^4 + w_k^2 \leq \lambda^2 & \forall k \in [K]. \end{aligned}$$

By introducing auxiliary variable $y_k = w_k^2$, this can be represented as a second-order conic programming:

$$\begin{aligned} \inf_{\{w_k\}_k, \{y_k\}_k, \lambda \geq 0} \quad & \lambda\rho + \sum_{j\in[K]} \widehat{p}_j c_j & \\ \text{s.t.} \quad & c_j + \begin{pmatrix} w_k^2 & w_k \end{pmatrix} \bar{z}_k + \lambda \|\widehat{x}_k - \widehat{x}_j\| \geq 0 & \forall j,k \in [K], \\ & y_k \geq w_k^2 & \forall k \in [K], \\ & y_k^2 + w_k^2 \leq \lambda^2 & \forall k \in [K]. \end{aligned}$$

□

*Proof of Example 7.* (D) and (4) are reduced to

$$\begin{aligned}\inf_{\substack{(B,\delta)\in\Theta\\ \lambda\geq 0,\{y_k\}_k\subset\mathbb{R}}} \quad & \lambda\rho^2+\sum_{k=1}^K \widehat{p}_k y_k \\ \text{s.t.} \quad & \begin{pmatrix} x^\top & z^\top & 1\end{pmatrix} X_k \begin{pmatrix} x\\ z\\ 1\end{pmatrix} \geq 0, && \forall k\in[K], x\in\mathcal{X}, z\in\mathcal{Z}\\ & \begin{pmatrix} C_\ell^\top B^\top & C_\ell^\top\delta - c_\ell\end{pmatrix}\begin{pmatrix} x\\ 1\end{pmatrix}\leq 0, && \forall \ell\in[L], x\in\mathcal{X}.\end{aligned}$$

where $X_k$ is a symmetric matrix defined in (EC.14) with $A=I$, $\alpha=\beta=\mathbf{0}$, and $b=0$:

$$X_k=\begin{pmatrix} \lambda I & -\frac{1}{2}B & -\lambda\widehat{x}_k\\ -\frac{1}{2}B^\top & \lambda I & -\frac{1}{2}\delta-\lambda\bar{z}_k\\ -\lambda\widehat{x}_k^\top & -\frac{1}{2}\delta^\top-\lambda\bar{z}_k^\top & y_k+\lambda\|\bar{z}_k\|^2+\lambda\|\widehat{x}_k\|^2\end{pmatrix}.$$

By the S-lemma Yakubovich (1977); Pólik and Terlaky (2007), two set of constraints are equivalent to

$$\begin{aligned}\inf_{\substack{B\in\mathbb{R}^{d\times m},\delta\in\mathbb{R}\\ \lambda\geq 0,\{y_k\}_k\subset\mathbb{R}\\ \{\mu_k\}_k,\{\nu_\ell\}_\ell\subset\mathbb{R}_+}} \quad & \lambda\rho^2+\sum_{k=1}^K \widehat{p}_k y_k \\ \text{s.t.} \quad & \begin{pmatrix} x^\top & z^\top & 1\end{pmatrix} X_k \begin{pmatrix} x\\ z\\ 1\end{pmatrix} +\mu_k((x-x_0)^\top\Sigma(x-x_0)-R)\geq 0, && \forall k\in[K], x\in\mathbb{R}^d, z\in\mathcal{Z}\\ & -\begin{pmatrix} C_\ell^\top B^\top & C_\ell^\top\delta - c_\ell\end{pmatrix}\begin{pmatrix} x\\ 1\end{pmatrix} +\nu_\ell((x-x_0)^\top\Sigma(x-x_0)-R)\geq 0, && \forall \ell\in[L], x\in\mathbb{R}^d.\end{aligned}$$

Constraints can be written as the semidefinite form:

$$X_k+\mu_k\begin{pmatrix}\Sigma & O & -\Sigma x_0\\ O & O & \mathbf{0}\\ -x_0^\top\Sigma & \mathbf{0}^\top & x_0^\top\Sigma x_0-R\end{pmatrix}\succeq O, \qquad -\begin{pmatrix} O & \frac{1}{2}BC_\ell\\ \frac{1}{2}C_\ell^\top B^\top & C_\ell^\top\delta-c_\ell\end{pmatrix}+\nu_\ell\begin{pmatrix}\Sigma & -\Sigma x_0\\ -x_0^\top\Sigma & x_0^\top\Sigma x_0-R\end{pmatrix}\succeq O.$$

We thus completed the proof of this example. □

We also formulate Example 7 with $\mathfrak{M}$ being the Wasserstein ball. Together with strong duality results for Wasserstein distributionally robust optimization (Gao and Kleywegt, 2023; Zhang et al., 2025), (4) can be written as

$$\begin{aligned}\inf_{\substack{(B,\delta)\in\Theta\\ \lambda\geq 0,\{y_{ki}\}_{ki}\subset\mathbb{R}}} \quad & \lambda\rho^2+\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}y_{ki}\\ \text{s.t.} \quad & \begin{pmatrix}x^\top & z^\top & 1\end{pmatrix} X_{ki}\begin{pmatrix}x\\ z\\ 1\end{pmatrix}\geq 0, && \forall k\in[K], i\in[n_k], x\in\mathcal{X}, z\in\mathcal{Z}\\ & \begin{pmatrix}C_\ell^\top B^\top & C_\ell^\top\delta-c_\ell\end{pmatrix}\begin{pmatrix}x\\ 1\end{pmatrix}\leq 0, && \forall \ell\in[L], x\in\mathcal{X}.\end{aligned}$$

where $X_k$ is a symmetric matrix defined in (EC.14) with $A=I$, $\alpha=\beta=\mathbf{0}$, and $b=0$:

$$X_{ki}=\begin{pmatrix}\lambda I & -\frac{1}{2}B & -\lambda\widehat{x}_k\\ -\frac{1}{2}B^\top & \lambda I & -\frac{1}{2}\delta-\lambda\widehat{z}_{ki}\\ -\lambda\widehat{x}_k^\top & -\frac{1}{2}\delta^\top-\lambda\widehat{z}_{ki}^\top & y_{ki}+\lambda\|\widehat{z}_{ki}\|^2+\lambda\|\widehat{x}_k\|^2\end{pmatrix}.$$

Again by the S-lemma (Yakubovich, 1977; Pólik and Terlaky, 2007), two set of constraints are equivalent to

$$\begin{aligned}\inf_{\substack{B\in\mathbb{R}^{d\times m},\delta\in\mathbb{R}\\ \lambda\geq 0,\{y_{ki}\}_{ki}\subset\mathbb{R}\\ \{\mu_{ki}\}_{ki},\{\nu_\ell\}_\ell\subset\mathbb{R}_+}} \quad & \lambda\rho^2+\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{p}_{ki}y_{ki}\\ \text{s.t.} \quad & \begin{pmatrix}x^\top & z^\top & 1\end{pmatrix} X_{ki}\begin{pmatrix}x\\ z\\ 1\end{pmatrix}+\mu_{ki}((x-x_0)^\top\Sigma(x-x_0)-R)\geq 0, \quad \forall k\in[K], i\in[n_k], x\in\mathbb{R}^d, z\in\mathcal{Z}\\ & -\begin{pmatrix}C_\ell^\top B^\top & C_\ell^\top\delta-c_\ell\end{pmatrix}\begin{pmatrix}x\\ 1\end{pmatrix}+\nu_\ell((x-x_0)^\top\Sigma(x-x_0)-R)\geq 0, && \forall \ell\in[L], x\in\mathbb{R}^d.\end{aligned}$$

Constraints can be written in the semidefinite form:

$$X_{ki}+\mu_{ki}\begin{pmatrix}\Sigma & O & -\Sigma x_0\\ O & O & \mathbf{0}\\ -x_0^\top\Sigma & \mathbf{0}^\top & x_0^\top\Sigma x_0-R\end{pmatrix}\succeq O, \qquad -\begin{pmatrix}O & \frac{1}{2}BC_\ell\\ \frac{1}{2}C_\ell^\top B^\top & C_\ell^\top\delta-c_\ell\end{pmatrix}+\nu_\ell\begin{pmatrix}\Sigma & -\Sigma x_0\\ -x_0^\top\Sigma & x_0^\top\Sigma x_0-R\end{pmatrix}\succeq O.$$

This is again a positive semidefinite program.

## EC.6. Additional Numerical Experiments

### EC.6.1. Comparison of Shapley and Non-Shapley Extension

Next, in our second set of experiments, we aim to compare the performance among different extensions of the in-sample optimal policy that are optimal to the DRO with causal transport distance, as discussed in Theorem 3 and Remark 3. We first compare the performance between $\mathsf{C}_{\mathsf{Lip}}$ and $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$. Figure EC.1 shows the relative differences in out-of-sample expected cost between $\mathsf{C}_{\mathsf{Lip}}$—a positive number indicates that $\mathsf{C}_{\mathsf{Lip}}$ outperforms $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$.

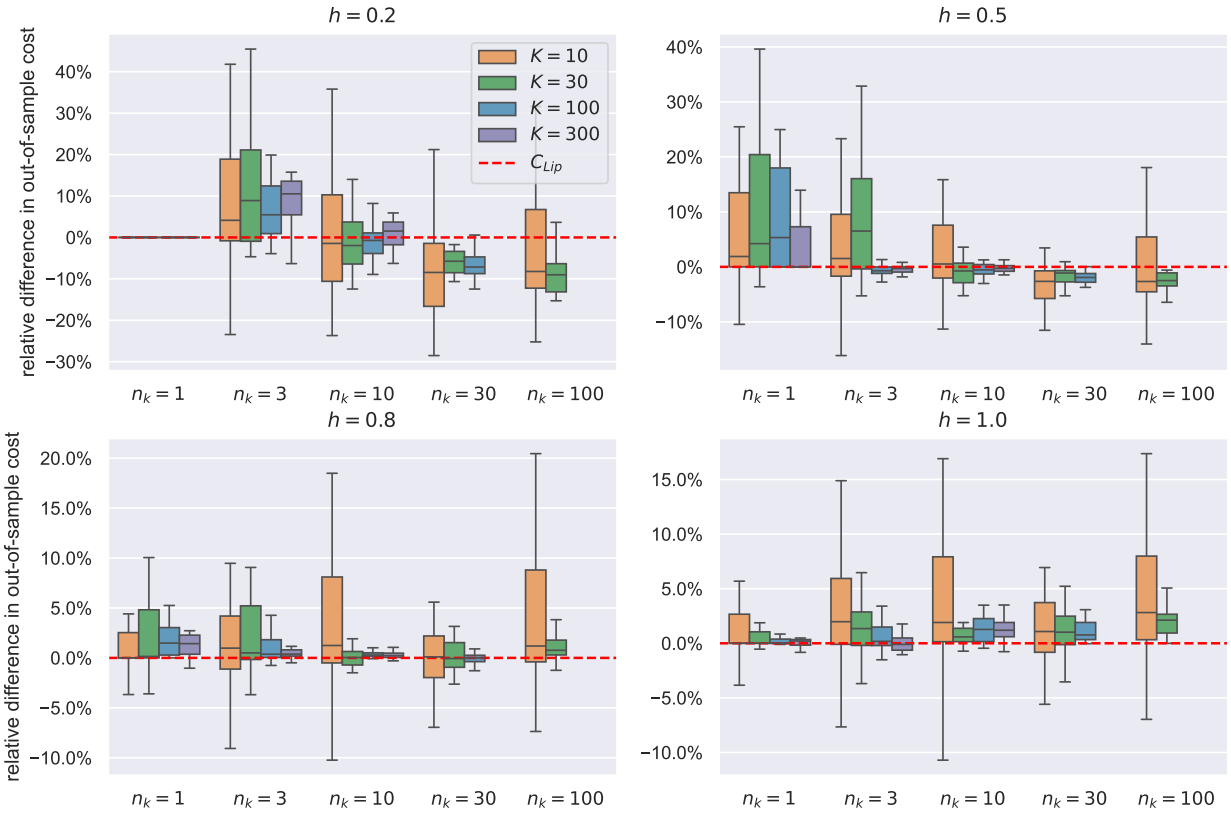


**Figure EC.1** **Boxplots of the relative differences in the out-of-sample performance between $\mathsf{C}_{\mathsf{Lip}}$ (baseline) and $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$**

We observe that both $\mathsf{C}_{\mathsf{Lip}}$ and $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$ have their own competitive advantages. Specifically, $\mathsf{C}_{\mathsf{Lip}}$ demonstrates superior performance compared to $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$ when dealing with a relatively small sample size $K$, given the same $n/K$. On the other hand, for a fixed sample size $K$, $\mathsf{C}_{\mathsf{Lip}}$ outperforms $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$ when the ratio of $n/K$ is low. This can be attributed to $\mathsf{C}_{\mathsf{Lip}}$ minimizing the $\infty$-norm, which leads to a more conservative approach than that of $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$. $\mathsf{C}_{\mathsf{Lip}}$ is more adept at managing situations with sparse data per covariate group. It

minimizes the impact of potential outliers or extreme scenarios, which is helpful when individual covariate groups have fewer observations. However, this conservatism may reduce its effectiveness when the sample size is large.

### EC.6.2. Comparison of Truncated and Non-truncated Policy

Moreover, we want to investigate further the relationship between different policies versus their truncated versions on the optimal region, as identified in Theorem 3. To begin with, we compare the performance of $\mathsf{C}_{\mathsf{TV}}$ with $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$. Figure EC.2 shows the mean of the differences of out-of-sample costs between $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$ and $\mathsf{C}_{\mathsf{TV}}$ with the same training and testing data set. A negative number implies that $\mathsf{C}_{\mathsf{TV}}$ is outperformed by $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$. In general, $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$ has an advantage over $\mathsf{C}_{\mathsf{TV}}$, especially when the overage cost $h$ is small or $n/K$ is small. Their differences are not very large in general, as $\mathsf{C}_{\mathsf{TV}}$ lies within the optimal region under most covariate values.

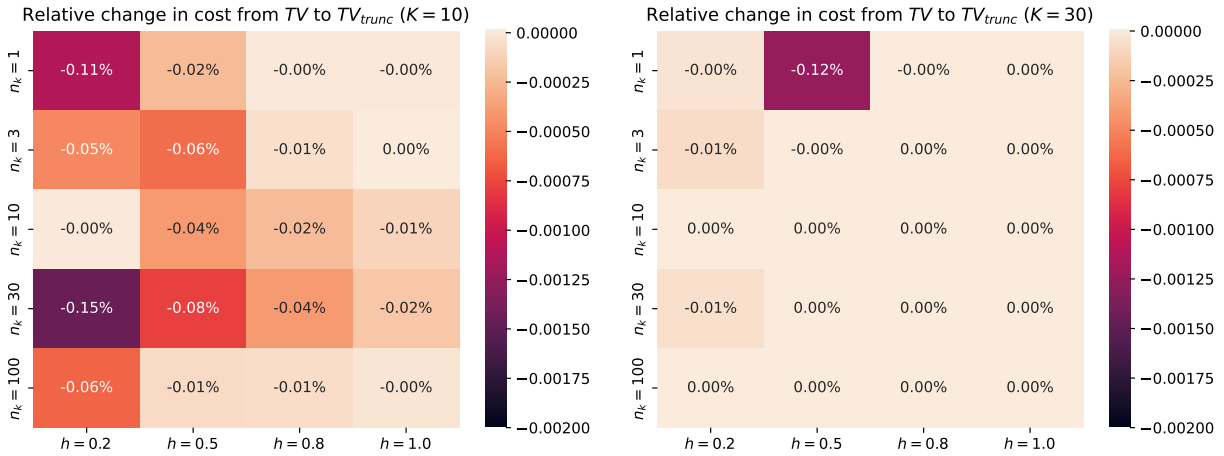


**Figure EC.2** **The differences in the out-of-sample performance between $\mathsf{C}_{\mathsf{TV}}$ and $\mathsf{C}_{\mathsf{TV\text{-}trunc}}$**

To deepen our understanding of the performance differences between policies and their truncated counterparts, we also compare the performance of empirical risk minimization using affine policy with $\ell^1$ and $\ell^2$ regularization ($\mathsf{ERM2}\,(\ell^1/\ell^2)$) in Ban and Rudin (2019) and their truncated versions. The results are shown in Figure EC.3. Setting $\mathsf{C}_{\mathsf{Lip}}$ as the baseline, the enhanced performance of the truncated versions emphasizes the efficacy of the optimal region.

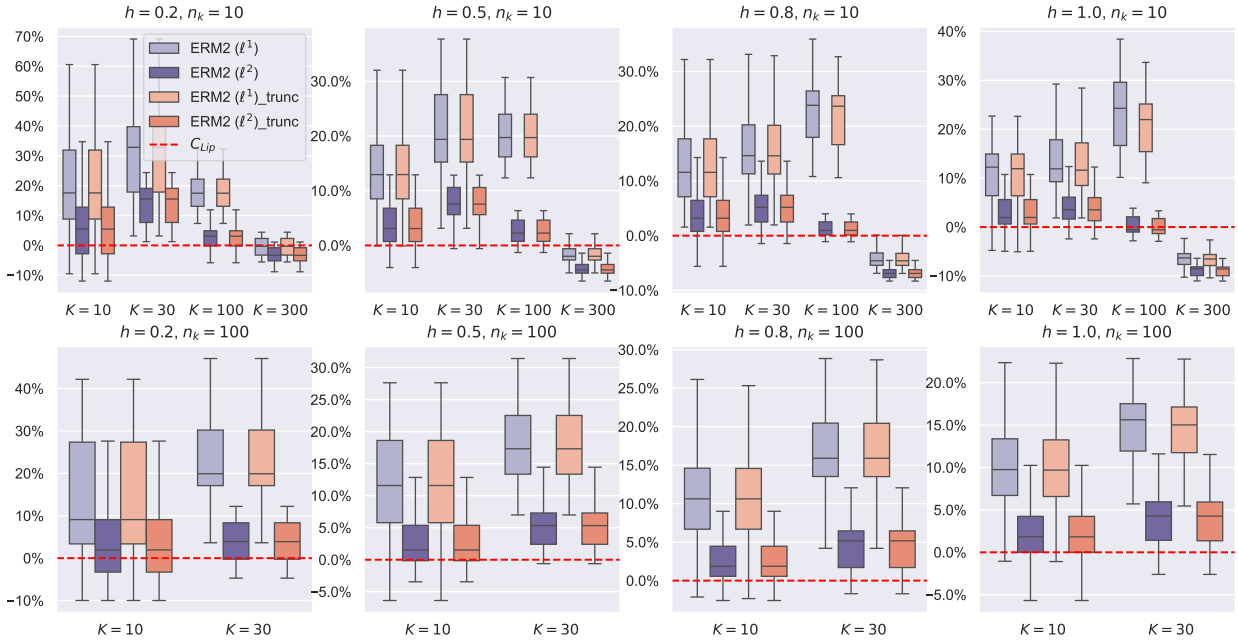


**Figure EC.3** **Boxplots of the relative differences in the out-of-sample performance between $\mathsf{C}_{\mathsf{Lip}}$ (baseline) and ERM2 ($\ell^1/\ell^2$)**