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

Jincheng Yang
Department of Mathematics, The University of Chicago, jincheng@uchicago.edu

Luhao Zhang
Department of Mathematics, The University of Texas at Austin, luhaozhang@utexas.edu

Ningyuan Chen
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. In particular, we propose to consider a distributionally robust formulation based on causal transport distance [48, 49]. Compared with divergence and Wasserstein metric, the causal transport distance is better at capturing the information structure revealed from the conditional distribution of random problem parameters given the covariate values. 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 has a similar conditional information structure as the nominal distribution. When optimizing over affine decision rules, we identify cases where the overall problem can be solved by 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 optimization or conditional stochastic optimization, is concerned with the following problem

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

which finds a decision w from the feasible region $\mathcal{D}$ so as to minimize the conditional expectation of some random cost $\Psi(w,Z)$ dependent on the decision w and a random variable Z, given some side information, represented by a covariate variable X. More informed or personalized decisions can be made with the side information revealed from the covariate data. This problem has received increasing attention nowadays as more side information becomes available to assist the decision making in e-commerce, online platform, etc. Quite often, the decision is made repeatedly for a distribution of covariate values — for example, the manager in an e-commerce company cares about the overall performance across all customer types. By averaging over these covariate values, 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 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 $\mathcal{X}$. The choice of $\mathcal{F}$ can vary

from small parametric classes like affine decision rules, to large non-parametric classes and even all measurable functions.

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 z represents the demand of a product and $h, b \geq 0$ represent the holding cost and the backorder cost respectively, then (2) is known as the big-data newsvendor model [6]. 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 $f^\star(x) = \mathbb{E}[Z|X=x]$ and thus the formulation (2) finds the conditional mean of Z given X. More examples will be given in Section 2.2. We remark that this is not the only formulation for contextual decision-making, and we will discuss other related works in Section 1.3.

Similar to the classical stochastic optimization, the underlying joint distribution $\mathbb{P}_{\mathsf{true}}$ of (X,Z) is often not known exactly, but instead, historical data from the underlying distribution are available. As such, 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. 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, lacks statistical consistency in general.

Two classes of distance-based uncertainty sets have been studied frequently 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 $\frac{1}{K}\sum_{k=1}^K \boldsymbol{\delta}_{(x_k,z_k)}$, 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 can be supported only on in-sample data, but differ from $\widehat{\mathbb{P}}$ 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 interchangeability principle [66]. 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 [7].

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 [46, 31]. Nonetheless, it faces new challenges when additional side information is presented. Let us first 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 being $\widehat{x}$. The angle between the two line segments is ε radian. Notably,

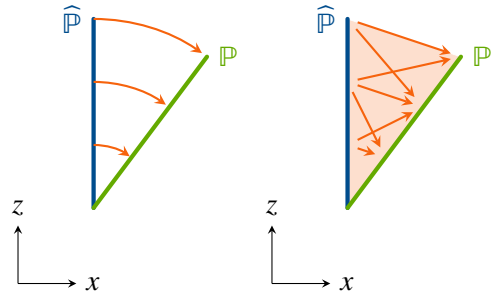

Figure 1 $\widehat{\mathbb{P}}$ and $\mathbb{P}$ has completely different conditional information structure.

the conditional distribution $\mathbb{P}^{\varepsilon}_{Z|X=x}$ is a Dirac measure for $x > \widehat{x}$, which is apparently very different from the conditional distribution $\widehat{\mathbb{P}}_{Z|X=\widehat{x}}$. As will be calculated in Section 2, Wasserstein distance between $\widehat{\mathbb{P}}$ and $\mathbb{P}$ is $O(\varepsilon)$, and the optimal transport maps 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 ε. ♣

In practice, the following situation is often seen from data: the conditional distribution can be estimated accurately under a number of covariate values, but is largely unobserved for other values. For example, historical data may reveal an accurate estimate of the conditional demand distribution of the product sold at deployed vending machines, but the demand at some new location is unexplored. Nonetheless, it is conceivable that the conditional demand distribution should share some resemblance among 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 that share a similar conditional information structure with the nominal distribution. Example 2 demonstrates that the Wasserstein uncertainty set fails to preserve the conditional information structure, and in fact, we will show in Section 3.2 that this phenomenon also holds for the worst-case distribution. This raises the concern of overly conservativeness of Wasserstein robust solutions.

1.2. Our Contribution

To capture the conditional information, in this paper, we consider a new distributional uncertainty set based on *causal transport distance*, a notion that is related to Wasserstein distance but imposes 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, which require new analysis on tractable reformulations and

interpretations. Moreover, when the outer minimization over the class of decision rules is performed over a non-parametric class, additional computational challenges are presented due to the involved 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). To the best of our knowledge, this is the first paper that considers distributionally robust optimization with causal transport distance and finds a computationally tractable solution.
(II) Our proof is based on new analysis of the worst-case distribution, through which we demonstrate how our choice of distributional uncertainty set helps to capture the conditional information structure of the random variable given the side information (Section 3.2).
(III) We study tractable reformulations for finding the optimal decision rule when optimizing over (i) the affine class and (ii) all (non-parametric) decision rules. In the former case, we provide convex program reformulations when the cost function $\Psi(w,z)$ is linear in the decision w or bilinear in w and z (Section 4.1). In the latter case, we provide convex program reformulations when the cost function Ψ is convex in the one-dimensional decision w. This provides a new class of decision rule with no sub-optimality gap for adjustable robust optimization (Section 4.2).

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: *separate prediction and optimization*, *conditional stochastic optimization*, and *optimization over decision rules*.

(I) Separate prediction and optimization is a classical two-step process that first estimates a conditional distribution of Z given a new context $X=x$, and then optimizes for the conditional expectation $\min_{w\in\mathcal{D}} \mathbb{E}[\Psi(w,Z)|X=x]$ (e.g., [70, 76]). There are some theoretical guarantees in this approach discussed in [23, 40]. One main issue of this framework, as discussed in [51, 6], is that the statistical estimation error and model mis-specification error may propagate to the decision optimization model and thus lead to a sub-optimal performance. Recent developments in contextual decision-making highlights the need for integrating the prediction and optimization.
(II) Conditional stochastic optimization avoids estimating the conditional distribution by directly estimating the conditional expected objective $\mathbb{E}[\Psi(w,Z)|X=x]$. Various estimation approaches have been studied, for example, based on Dirichlet process [37], Nadaraya-Watson kernel regression [36, 6, 67], local regression and classification [13, 15], smart prediction-then-optimization [25, 23, 24, 38], trees and forests [5, 43], robustness optimization and regularization [71, 77, 17, 52, 27], regret minimization [28], empirical residuals [44, 45], bilevel optimization [53, 21], etc. This approach requires solving a decision optimization problem for each individual context, which could be computationally prohibitive when numerous contexts are presented.
(III) Optimization over decision rules is an end-to-end formulation which 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 [20, 6, 8, 16] or in a reproduce kernel Hilbert space [14] and weight matrices in a neural network [54, 63, 50]. Our formulation falls into this category, but our results in Section 4 do not restrict the class of decision rules on a parametric family. In this respect, the closest work to ours is [32], which considers robust optimization over decision rules with Wasserstein uncertainty set; see the last paragraph of this subsection for a detailed comparison.

We remark that in online setting, stochastic optimization with side information has also been considered under the umbrella of contextual bandits and reinforcement learning, which are beyond the scope of this paper.

On causal transport. The origin of the idea of causal transport could be traced back to the Yamada-Watanabe criterion for stochastic differential equations [74, 47, 42]. In the theory of optimal transport, Lassalle [48] studied the transport problem in continuous time under the so-called causality constraint, and [4] considers a discrete-time analogue. Causal transport has been applied to stochastic optimization in [2], as well as other areas such as stochastic control [1] and machine learning [73]. In discrete time stochastic programming, *nested distance*—a symmetrized analogue of casual transport distance—has been exploited to study the stability and sensitivity of multistage stochastic programming [55, 57, 58, 59, 60]. Our problem can be viewed as a causal transport with two time periods.

On distributionally robust optimization. Distributionally robust optimization (DRO) has received significant attentions recently as a tool for decision-making under uncertainty, and different approaches mainly differ in how the uncertainty set is constructed. Our choice of uncertainty set is aligned with DRO with transport distance, such as Wasserstein distance [56, 72, 26, 19, 18, 31, 30, 29] and nested distance [3, 61, 65]. To our best knowledge, our distributionally robust formulation based on the causal transport distance has not been studied in the literature. We refer to [64] for a thorough review on other choices of uncertainty set.

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 [22, 11, 12, 10, 41, 39, 16, 33], k-adaptability [34, 35, 69], iterative splitting of uncertainty sets [62], binary decision rules [9], non-parametric Markovian stopping rules [68], etc. Most of these works do not consider covariate in their problem. [16] consider dynamic decision-making with side information using affine decision rules where as we consider general decision rules in a static setting; and [32] consider 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 Shapely 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 Section . 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 Appendices.

2. Distributionally Robust Optimization with Causal Transport Distance

In this section we briefly introduce notations 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 normed vector spaces. 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 $\mathcal{Z}$ 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

Causal transport distance and its associated optimal transport problem were introduced in [48], whose main motivation is to investigate optimal transportation problems with filtrations and their applications to stochastic calculus. The discrete-time counterpart was investigated in [4]. The definition of causal transport distance, specialized to our considered setting, is as follows.

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

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

Like Wasserstein distance, causal transport distance finds the minimal transport cost between two distributions, where norms captures the geometry of the data space and 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 restrict on causal transport plans as defined in Definition 1.

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

EXAMPLE 3 (CAUSAL TRANSPORT BETWEEN COLOR 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). A bitmap image stores the position-color information of an image via a $H \times H \times 3$ 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 $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_1$ and $\|\cdot\|_{\mathcal{Z}} = c\mathbf{1}\{\cdot = 0\}$, where c is a scaling parameter.

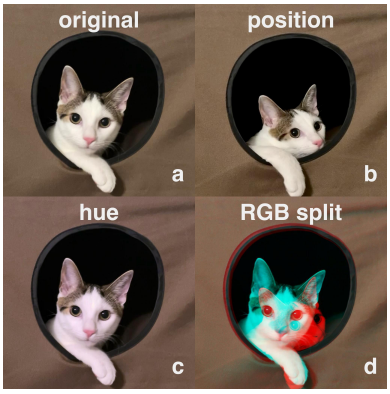

Figure 2 A color image (a) and its variations by shifting the position (b), adjusting the hue (c), or splitting the RGB channels (d)

Figure 2 contains four images of a cat: (a)(b)(c) can be viewed as real natural images with different poses or color portions, whereas (d) can be viewed as a fake 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 a causal transport.

(II) The cats in (a) and (c) have identical poses but different hue values. Changing hue values of an image would affect its RGB values and thus 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 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 be dependent on both $\widehat{X}$ and $\widehat{Z}$, that is, at each position of the image, changes in the color portions are permitted.

(III) The green and blue channels of (d) has 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 $\widehat{Z} \in \{G, B\}$, and transport it according to the cat's movement if $\widehat{Z} = R$, 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	2.303	2.044	0.495
Causal transport distance	2.767	2.535	**6.388**

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, under causal transport distance between Fig. 2(a) and the fake image Fig. 2(d) is much larger than that between Fig. (a) and the real images Fig. 2(b)(c). In contrast, Wasserstein distance fails to capture such intuition. ♣

As hinted in Example 3, one of the main advantage 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 4 (Revisit of Example 2). We compute the causal transport distance and the Wasserstein distance between $\widehat{\mathbb{P}}$ and $\mathbb{P}^\varepsilon$ shown in Example 2. Since the conditional distribution of $\mathbb{P}^\varepsilon$ is a Dirac measure for every x, the causal transport distance between $\widehat{\mathbb{P}}$ and $\mathbb{P}^\varepsilon$ 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 $\widehat{\mathbb{P}} \otimes \mathbb{P}^\varepsilon$, so

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

As a result, $\mathbb{P}$ would not belong to the uncertainty set induced from the causal transport distance with a small radius. This is consistent to 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 structure from the nominal one, and therefore may lead to an overly conservative solution. ♣

Worth to be noticed that, 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 the Fig. 1.17 in [58].

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

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

Below, we list a few examples.

EXAMPLE 5 (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 6 (FEATURE-BASED NEWSVENDOR). Let h and b represent the unit holding cost and the unit backordering 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 7 (PERSONALIZED PRICING). Consider a demand model $D_0 - z(w - w_0)$, where w_0 is an incumbent price under which the expected demand is known accurately to be D_0. This can be thought of as a reference price where the decision maker has sufficient data to learn the market response. z is a parameter interpreted as the price sensitivity coefficient. In practice, the price sensitivity coefficient z may exhibit heterogeneity among populations. As such, we model it as a random variable Z, which is affected by the contextual information X, based on which the decision maker can adjust the price directly or indirectly through personalized promotion. The revenue is calculated as $w(D_0 - z(w - w_0))$. Consider a revenue maximization with personalized pricing

$$\inf_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathfrak{M}} \mathbb{E}_{(X,Z)\sim\mathbb{P}}\left[-f(X)(D_0 - Z(f(X) - w_0))\right]. \qquad \clubsuit$$

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

EXAMPLE 8 (PORTFOLIO OPTIMIZATION WITH AFFINE DECISION RULE). Consider a portfolio optimization involving m assets. The return rate of the ith asset is modeled as a random variable Z_i. Suppose a weight $w \in \mathbb{R}^m$ is allocated on the assets with the restriction $\sum_{i=1}^m w_i = 1$, thereby the random loss of a portfolio is given by $w^\top Z$. Again, the weight w can be chosen based on the contextual information X. Consider the portfolio optimization problem

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

Here $\mathcal{F}$ is a class of functions $f : \mathcal{X} \to \mathcal{D}$, where

$$\mathcal{X} = \mathbb{R}^d, \qquad \mathcal{D} = \left\{ w \in \mathbb{R}^m : \mathbf{1}^\top w = 1 \right\}.$$

Here $\mathbf{1}$ is the m-dimensional all-one vector. In case of affine policies, we require $f \in \mathcal{F}$ to be affine. We can write

$$\mathcal{F} = \left\{ x \mapsto \left(\mathrm{Id} - \frac{1}{m} \mathbf{1}\mathbf{1}^\top \right) Ax + \frac{1}{m}\mathbf{1} : A \in \mathbb{R}^{m \times d} \right\}.$$

Here Id is the m-dimensional identity matrix. ♣

3. Evaluating 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, which demonstrates how our choice of causal transport distance-based distributional uncertainty set helps to preserves 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. Our proof technique can be extended to a general metric space with additional technical treatment.

Assumption 1. *$\mathcal{X}$, $\mathcal{Z}$, $\mathcal{D}$ are 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*

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

3.1. Strong Duality Reformulation

We begin by developing a tractable reformulation through deriving its strong dual.

For a fixed decision rule f, we define the primal problem as

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

and the dual problem as

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

The dual variable λ corresponds to the Lagrangian multiplier of the causal constraint in the primal problem. We will show that (P^f) and (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

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

By the tower property,

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

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

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

This completes the proof for the weak duality. □

The main result for this section states as follows.

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

The proof idea of Theorem 1 is to construct a nearly worst-case distribution of the primal problem based on the first-order optimality condition of the weak dual problem. Conceptually it shares some similar aspects as the duality proof for Wasserstein DRO [31], but differs from it in terms of the construction of a nearly worst-case distribution. Specifically, the nearly worst-case distribution is obtained by moving $\widehat{z}_{ki}$ toward the maximizer of the innermost maximization problem $\Upsilon(\lambda; x, \widehat{z}_{ki}) := \sup_{z\in\mathcal{Z}}\{\Psi(f(x),z) - \lambda\|z - \widehat{z}_{ki}\|\}$, and moving $\widehat{x}_k$ toward the maximizer of the maximization problem $\sup_{x\in\mathcal{X}}\{\Upsilon(\lambda; x, \widehat{z}_{ki}) - \lambda\|x - \widehat{x}_k\|\}$. One can see that such a transport plan is causal: where $\widehat{x}_k$ is transported depends only on $\widehat{x}_k$ but not on $\widehat{z}_{ki}$. If both maximizers over x and over z exist and are unique, then the transport plan would induce a worst-case distribution. If the maximizers do not exist or are not unique, we can still find two transport plans such that one induces a feasible yet suboptimal distribution, while the other induces an infeasible yet superoptimal distribution. By interpolating these two distributions, we can obtain a near-optimal feasible solution to the primal problem. 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 [31, 75]

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

Comparing it with the dual problem (D^f) of causal transport distance DRO, the difference is the switching 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, thus 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. ◇

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 folowing result we require $\mathcal{X}$ and $\mathcal{Z}$ to be finite dimensional and thus locally compact, and require some continuity assumptions on Ψ, 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 reformulation (D^f) is achieved at some $\lambda^* > \kappa$ for κ specified in Lemma EC.2, then a worst case distribution exists and has the following form*

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

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

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

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

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 $(\hat{p}_{ki})_{i=1,2,3}$. 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 $(\hat{p}_{2i})_{i=1,2,3}$.

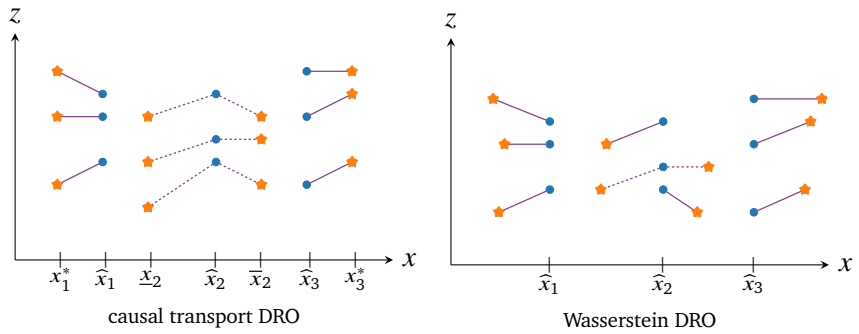

Figure 3 Structure of the worst-case distributions

As a comparison, on the right side of Figure 3, we plot a worst-case distribution resulting from Wasserstein DRO. According to [31], 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.

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:

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

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

$$v_{\mathsf{D}} := \inf_{\substack{f\in\mathcal{F}\\ \lambda\geq 0}} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Psi(f(x),\widehat{Z}) | \widehat{X} \right] - \lambda\|x-\widehat{X}\|^p \right\} \right] \right\}. \tag{4}$$

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 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 programing 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 4.2).

4.1. Optimizing over Affine Decision Rules

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

$$\mathcal{F}_\Theta = \left\{ x\mapsto B^\top x+\delta :\ (B,\delta)\in\Theta \right\} \tag{5}$$

where Θ is a finite-dimensional convex set.

Our first result shows that (4) is tractable when Ψ is affine in the decision variable w.

Corollary 1. *Suppose $\mathcal{F}=\mathcal{F}_\Theta$ defined in (5), and $\Psi(\cdot,z)$ is affine for every z:*

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

Set

$$\beta_k := \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\beta(\widehat{Z})|\widehat{X}=\widehat{x}_k], \quad b_k := \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[b(\widehat{Z})|\widehat{X}=\widehat{x}_k].$$

Then the dual problem (4) is equivalent to the following convex program

$$\inf_{\lambda\geq 0,(B,\delta)\in\Theta} \left\{ \lambda\rho^p + \sum_{k=1}^K \left(\sum_{i=1}^{n_k} \hat{p}_{ki} \right) \cdot \left(\beta_k^\top (B^\top\widehat{x}_k+\delta) + b_k + R_p\left(\lambda, |B\beta_k|\right) \right) \right\},$$

where $R_p:\mathbb{R}_+^2\to\mathbb{R}\cup\{\infty\}$ is a convex function defined as

$$R_p(\lambda,\mu) := \sup_{t\geq 0}\{\mu t-\lambda t^p\} = \begin{cases} \infty\mathbf{1}_{\{\lambda<\mu\}}, & p=1,\\ \lambda(p-1)\left(\frac{\mu}{\lambda p}\right)^{\frac{p}{p-1}}, & p>1. \end{cases} \tag{6}$$

Next, we consider the case when $\Psi(w,z)$ is bilinear. For the sake of tractability, we restrict ourselves to the case $p=2$.

COROLLARY 2. *Suppose $\mathcal{F}=\mathcal{F}_\Theta$ as defined in (5) and $\Psi(w,z)$ is bilinear:*

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

Set

$$\beta_k=\beta+A\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z}|\widehat{X}=\widehat{x}_k],\quad b_k=b+\alpha^\top\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z}|\widehat{X}=\widehat{x}_k].$$

Then (D) with $p=2$ is equivalent to the following convex program

$$\inf_{\substack{(B,\delta)\in\Theta\\ \frac{1}{2}\|BA\|_2<\lambda}}\left\{\lambda\rho^2+\sum_{k=1}^K\left(\sum_{i=1}^{n_k}\hat{p}_{ki}\right)\cdot\left(\beta_k^\top(B^\top\widehat{x}_k+\delta)+b_k+\frac{|A^\top(B^\top\widehat{x}_k+\delta)+\alpha|^2}{4\lambda}\right.\right.$$
$$\left.\left.+\frac{1}{4}\left[\frac{A(A^\top(B^\top\widehat{x}_k+\delta)+\alpha)}{2\lambda}+\beta_k\right]^\top\left[\frac{1}{4\lambda}(BA)(BA)^\top-\lambda\,\mathrm{Id}\right]^{-1}\left[\frac{A(A^\top(B^\top\widehat{x}_k+\delta)+\alpha)}{2\lambda}+\beta_k\right]\right)\right\}.$$

Note that the problem above can be written as a semi-definite program.

4.2. Optimizing over All (Non-parametric) Decision Rules

In this subsection, we consider $\mathcal{F}$ to be unrestricted and contains 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

$$v_{\mathsf{D}}=\min_{f:\mathcal{X}\to\mathcal{D}}\min_{\lambda\geq0}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{\varphi(f(x);\lambda,\widehat{X})-\lambda\|x-\widehat{X}\|\right\}\right]\right\},\tag{7}$$

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

$$v_{\hat{\mathsf{D}}}:=\min_{\substack{f:\mathcal{X}\to\mathcal{D}\\ \lambda\geq0}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq K}\left\{\varphi(f(x_k);\lambda,\widehat{X})-\lambda\|x_k-\widehat{X}\|\right\}\right]\right\}\tag{8}$$

$$=\min_{\substack{\hat{f}\in\hat{\mathcal{F}}\\ \lambda\geq0}}\left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq k\leq K}\left\{\varphi(\hat{f}(x_k);\lambda,\widehat{X})-\lambda\|x_k-\widehat{X}\|\right\}\right]\right\},\tag{9}$$

where the second equality holds because the objective value in (8) depends only on the value of f on $\operatorname{supp}\widehat{\mathbb{P}}$. Note that (9) 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 a convex subset, and $\Psi(w,z)$ is convex in w. Let $(\lambda^*,\hat{f}^*)$ be the minimizer to the in-sample dual problem (9). Denote $\varphi_k(y):=\varphi(y;\lambda^*,\widehat{x}_k)$, $y_k:=\hat{f}^*(\widehat{x}_k)$, and $\phi_k:=\max_j\{\varphi_k(y_j)-\lambda^*\mathsf{d}(\widehat{x}_k,\widehat{x}_j)\}$. For $x\in\mathcal{X}$, define*

$$I_k(x):=\{y\in\mathcal{D}:\varphi_k(y)\leq\lambda^*\mathsf{d}(x,x_k)+\phi_k\}.$$

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 (7).

Theorem 3 shows that problems (7) and (9) share the same optimal dual variable λ^*, and to solve the infinite-dimensional optimization over decision rules (7), it suffices to first solve a finite-dimensional robust in-sample optimization (9) 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 (9) is solved, the values y_k, ϕ_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 $\hat{f}$ to the entire space, as long as it belongs to the range of $\cap_k I_k(x)$.

The proof idea of Theorem 3 is as follows. Observe that $v_{\mathsf{D}} \geq v_{\hat{\mathsf{D}}}$, since the inner supremum in (7) is taken with respect to a larger set compared with the maximization in (8). To see the other direction, the main step is to show $I_k(x)$ has a nonempty intersection. Once this is shown, using simple algebra it is easy to verify that $f^*(x) \in \cap_k I_k(x)$ attains the value $v_{\hat{\mathsf{D}}}$, thereby v_{D} is dominated by the objective value of f^* which equals $v_{\hat{\mathsf{D}}}$. To show $I_k(x)$ has a nonempty intersection, it suffices to show they pairwise intersect because they are one-dimensional intervals. This can be established using the convexity of φ.

Remark 2 (Comparison with the Shapely policy in [32]). In [32], 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 Shapely policy, can be found by first solving for the in-sample Wasserstein robust optimal decision rule $\hat{f}_{\mathsf{W}}$, then extending to the entire space by solving

$$f^*(x) = \arg\min_{y \in \mathbb{R}} \max_k \frac{|y - \hat{f}_{\mathsf{W}}(\widehat{x}_k)|}{\|x - \widehat{x}_k\|},$$

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

$$f^*_\infty(x) := \arg\min_{y \in \mathbb{R}} \max_k \frac{|y - \hat{f}^*(\widehat{x}_k)|}{\|x - \widehat{x}_k\|}, \tag{10}$$

where $\hat{f}^*(\widehat{x}_k)$'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 $f^*_\infty(x)$ defined a robust optimal decision rule for (7). Note that we use the subscript ∞ to indicate the ∞-norm (maximum) of the slope function $k \mapsto \frac{|y - \hat{f}^*(\widehat{x}_k)|}{\|x - \widehat{x}_k\|}$.

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

$$f^\dagger_1(x) := \arg\min_{y \in \mathbb{R}} \sum_k \frac{|y - y_k|}{\|x - \widehat{x}_k\|}.$$

The resulting decision rule may not necessarily optimal, but we can always truncate its values to force them falling into $\cap_k I_k(x)$ and thereby making it robust optimal. Namely, if we use $\overline{I}(\cdot)$ and $\underline{I}(\cdot)$ to represent the upper and lower bound of the region $\cap_k I_k(x)$, the we define

$$\bar{f}^\dagger_1(x) := \max\left(\underline{I}(x),\ \min\left(f^\dagger_1(x), \overline{I}(x)\right)\right). \tag{11}$$

We denote the truncated decision rule as $\bar{f}^\dagger_1(x)$.

We illustrate the two robust optimal decision rules defined above using a conditional median estimate problem with $Z = \mu(X) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, $\mu(x) = \sin(2x) + 2\exp(-16x^2)$.

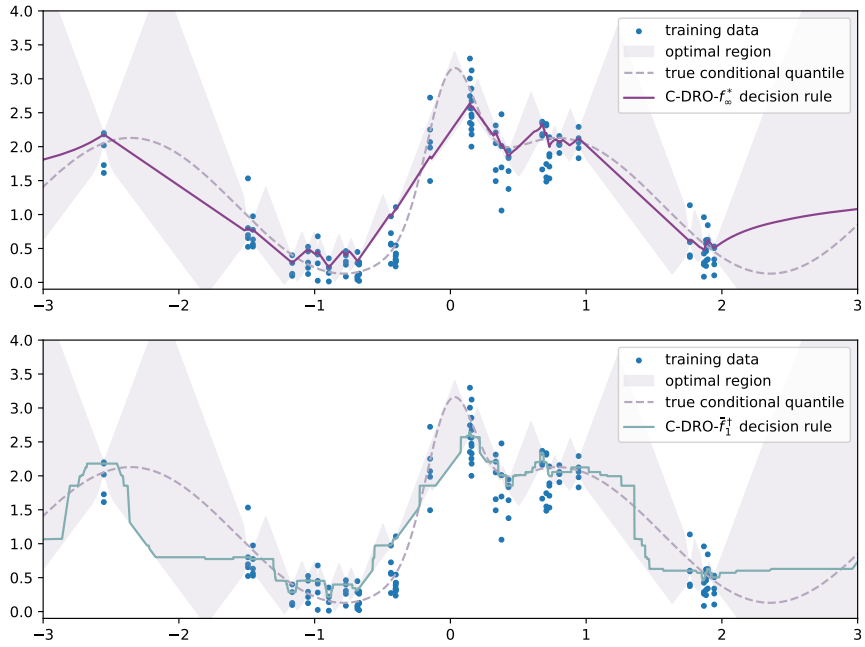

Figure 4 Two robust optimal decision rules f_∞^* and $\hat{f}_1^\dagger$ of a median estimation problem

Example 9 (Conditional median estimation). Consider the feature-based newsvendor problem in Example 6. When $h = b = 1$, this is equivalent to conditional median estimation. As detailed in EC.5, the in-sample dual problem (9) can be transformed into a linear programming problem

$$\begin{aligned}
\inf_{\{w_k\}_k, \lambda \geq 0} \ & \lambda\rho + \frac{1}{n}\sum_{k=1}^{K} c_j \\
\text{s.t.} \quad & c_j \geq \sum_{i=1}^{n_j} c_{kji} - \lambda n_j \mathsf{d}(\widehat{x}_k, \widehat{x}_j), \forall j, k \\
& c_{kji} \geq w_k - \widehat{z}_{ji}, \forall k, j, i \\
& c_{kji} \geq \widehat{z}_{ji} - w_k, \forall k, j, i \\
& \lambda \geq 1
\end{aligned}$$

Example 10 (Personalized Pricing). Consider the personalized pricing problem in Example 7. By Theorem 1, its strong dual problem can be written as

$$\inf_{\substack{f:\mathcal{X}\to\mathbb{R} \\ \lambda \geq 0}} \left\{ \lambda\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}} \left[\sup_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[\sup_{z\in\mathcal{Z}} \left\{ -f(x)(D_0 - z(f(x) - w_0) - \lambda \mathsf{d}(z, \widehat{Z})^p \right\} | \widehat{X} \right] - \lambda \mathsf{d}(x, \widehat{X})^p \right\} \right] \right\}.$$

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

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

where

$$\varphi(w; \lambda; \widehat{x}) = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}} \left[\sup_{z\in\mathcal{Z}} \left\{ -w(D_0 - z(w - w_0)) - \lambda \mathsf{d}(z, \widehat{Z}) \right\} | \widehat{X} = \widehat{x} \right].$$

A detailed calculation could be found in EC.5. In case when $\|z-\hat{z}\|_{\mathcal{Z}} = \infty \mathbf{1}_{\{z\neq\hat{z}\}}$, by adding dummy variables, this can be transformed into a quadratic programming problem

$$\begin{aligned} \inf_{w_k,\lambda} \quad & \lambda\rho + \sum_{k=1}^{K} \hat{p}_k c_k \\ s.t. \quad & c_k \geq -w_j \left(D_0 - \bar{z}_k (w_j - w_0)\right) - \lambda \mathrm{d}(\widehat{x}_j, \widehat{x}_k), \forall j,k \end{aligned}$$

where $\hat{p}_k = \sum_{i=1}^{n_k} \hat{p}_{ki}$ and $\bar{z}_k = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\widehat{Z}|\widehat{X}=\widehat{x}_k\right]$.

5. Numerical Experiments

In this section, we illustrate our proposed approach in the context of feature-based newsvendor. We consider a similar setup as in [32]. 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 β 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 holding cost $h \in \{0.2, 0.5, 0.8, 1\}$ while fixing the unit backordering 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 $\|\cdot\|_{\mathcal{Z}} = \infty \cdot \mathbf{1}_{\{z\neq\hat{z}\}}$. To generate the boxplots, we run 20 repeated experiments (except for $K = 10$ we run 50 experiments to get 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).

In our first set of experiments, we compare the performance between W-DRO and C-DRO-f_∞^* as defined in (10). Figure 5 shows the relative difference in the out-of-sample expected cost — a negative number indicates that C-DRO-f_∞^* outperforms W-DRO.

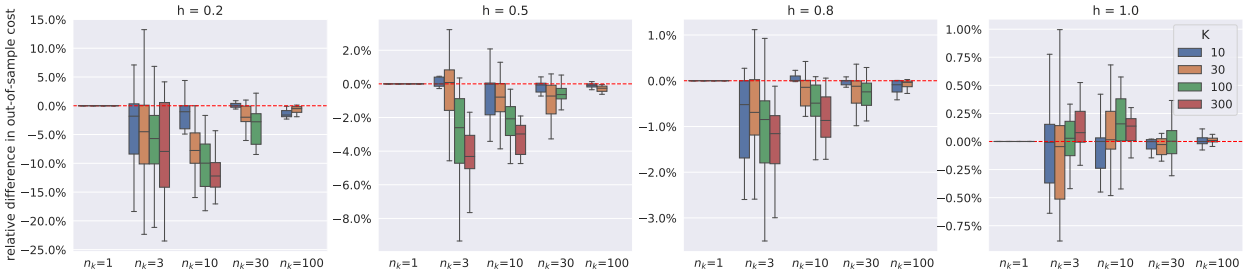

Figure 5 Boxplots of the relative differences in the out-of-sample performance between W-DRO and C-DRO-f_∞^*

We have the following observations.

(I) When $n_k = 1$, C-DRO-f_∞^* and W-DRO have same performance because the two formulation are equivalent (Remark 1).

(II) When $n_k = 3, 10, 30, 100$, C-DRO-f_∞^* performs relatively better than W-DRO when the loss function is skewed ($h = 0.2, 0.5, 0.8$). The advantage is mostly clear when $n_k = 3, 10$, which shows the value of (even a little) conditional information. When n_k is very large, the advantage of is lesssC-DRO-f_∞^*. One explanation is that the worst-case distribution of W-DRO does not deteriorate the conditional information structure greatly when there are many samples at the same covariate value.

(III) The advantage of C-DRO-f_∞^* over W-DRO increases as K increases in most cases. An explanation is that when K is large, C-DRO can fully take advantages of these conditional information to extrapolate other conditional distributions.

Next, in our second set of experiments, we compare the performance between C-DRO-f_∞^* and C-DRO-$\bar{f}_1^\dagger$ as defined in (11). Figure 6 shows the relative differences in out-of-sample expected cost between C-DRO-f_∞^* — a positive number indicates that C-DRO-f_∞^* outperforms C-DRO-$\bar{f}_1^\dagger$.

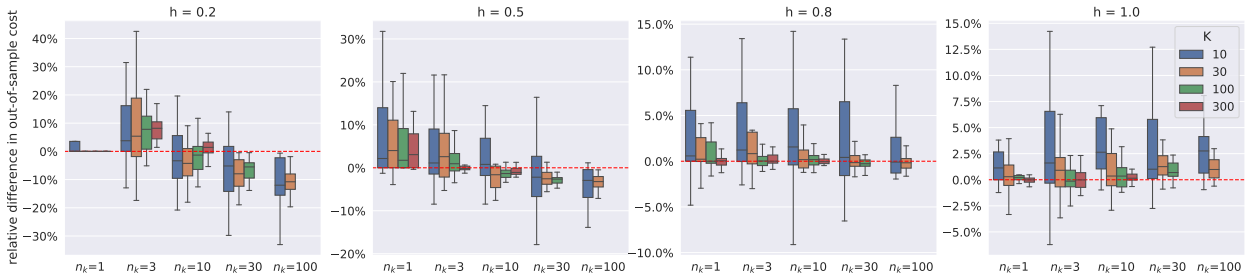

Figure 6 Boxplots of the relative differences in the out-of-sample performance between C-DRO-$-f_\infty^*$ and C-DRO-$\bar{f}_1^\dagger$

We observe that both C-DRO-f_∞^* and C-DRO-$\bar{f}_1^\dagger$ have their own competitive advantages. C-DRO-f_∞^* has better performance when the sample size is relatively small. Since it minimizes ∞-norm, it is more conservative than C-DRO-$\bar{f}_1^\dagger$, which deteriorate the performance when the sample size is large.

6. Concluding Remarks

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

References

[1] Beatrice Acciaio, Julio Backhoff-Veraguas, and René Carmona. Extended mean field control problems: stochastic maximum principle and transport perspective. *SIAM journal on Control and Optimization*, 57(6):3666–3693, 2019.

[2] Beatrice Acciaio, Julio Backhoff-Veraguas, and Anastasiia Zalashko. Causal optimal transport and its links to enlargement of filtrations and continuous-time stochastic optimization. *Stochastic Processes and their Applications*, 130(5):2918–2953, 2020.

[3] Bita Analui and Georg Ch Pflug. On distributionally robust multiperiod stochastic optimization. *Computational Management Science*, 11(3):197–220, 2014.

[4] Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

[5] Gah-Yi Ban, Jérémie Gallien, and Adam J Mersereau. Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing & Service Operations Management*, 21(4):798–815, 2019.

[6] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

[7] Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *INFORMS TutORials in Operations Research*, pages 1–19. INFORMS, 2015.

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

[9] Dimitris Bertsimas and Angelos Georghiou. Design of near optimal decision rules in multistage adaptive mixed-integer optimization. *Operations Research*, 63(3):610–627, 2015.

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

[11] Dimitris Bertsimas, Dan A Iancu, and Pablo A Parrilo. Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research*, 35(2):363–394, 2010.

[12] Dimitris Bertsimas, Dan Andrei Iancu, and Pablo A Parrilo. A hierarchy of near-optimal policies for multistage adaptive optimization. *IEEE Transactions on Automatic Control*, 56(12):2809–2824, 2011.

[13] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[14] Dimitris Bertsimas and Nihal Koduri. Data-driven optimization: A reproducing kernel hilbert space approach. *Operations Research*, 70(1):454–471, 2022.

[15] Dimitris Bertsimas and Christopher McCord. From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637*, 2019.

[16] Dimitris Bertsimas, Christopher McCord, and Bradley Sturt. Dynamic optimization with side information. *European Journal of Operational Research*, 2022.

[17] Dimitris Bertsimas and Bart Van Parys. Bootstrap robust prescriptive analytics. *Mathematical Programming*, pages 1–40, 2021.

[18] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

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

[20] Michael W Brandt, Pedro Santa-Clara, and Rossen Valkanov. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447, 2009.

[21] Junyu Cao and Rui Gao. Contextual decision-making under parametric uncertainty and data-driven optimistic optimization. *Available at Optimization Online*, 2021.

[22] Xin Chen, Melvyn Sim, Peng Sun, and Jiawei Zhang. A linear decision-based approximation approach to stochastic programming. *Operations Research*, 56(2):344–357, 2008.

[23] Othman El Balghiti, Adam N Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. *Advances in Neural Information Processing Systems*, 32:14412–14421, 2019.

[24] Adam Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. Decision trees for decision-making under the predict-then-optimize framework. In *International Conference on Machine Learning*, pages 2858–2867. PMLR, 2020.

[25] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 2021.

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

[27] Adrián Esteban-Pérez and Juan M Morales. Distributionally robust stochastic programs with side information based on trimmings. *arXiv preprint*, page arXiv:2009.10592, 2020.

[28] Alexander Estes. Slow rates of convergence in optimization with side information. *Available at SSRN 3803427*, 2021.

[29] Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

[30] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *arXiv preprint*, page arXiv:1712.06050, 2017.

[31] Rui Gao and Anton J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv e-prints*, page arXiv:1604.02199, April 2016.

[32] Rui Gao, Jincheng Yang, and Luhao Zhang. Optimal robust policy for feature-based newsvendor. *Management Science, Forthcoming*, 2022.

[33] Angelos Georghiou, Angelos Tsoukalas, and Wolfram Wiesemann. On the optimality of affine decision rules in robust and distributionally robust optimization. *Optimization Online*, 2021.

[34] Grani A Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. K-adaptability in two-stage robust binary programming. *Operations Research*, 63(4):877–891, 2015.

[35] Grani A Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. K-adaptability in two-stage distributionally robust binary programming. *Operations Research Letters*, 44(1):6–11, 2016.

[36] Grani Adiwena Hanasusanto and Daniel Kuhn. Robust data-driven dynamic programming. *Advances in Neural Information Processing Systems*, 26, 2013.

[37] Lauren Hannah, Warren Powell, and David Blei. Nonparametric density estimation for stochastic optimization with an observable state variable. *Advances in Neural Information Processing Systems*, 23:820–828, 2010.

[38] Nam Ho-Nguyen and Fatma Kılınç-Karzan. Risk guarantees for end-to-end prediction and optimization processes. *Management Science*, 2022.

[39] Omar El Housni and Vineet Goyal. On the optimality of affine policies for budgeted uncertainty sets. *arXiv preprint*, page arXiv:1807.00163, 2018.

[40] Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science (Forthcoming)*, 2022.

[41] Dan A Iancu, Mayank Sharma, and Maxim Sviridenko. Supermodularity and affine policies in dynamic robust optimization. *Operations Research*, 61(4):941–956, 2013.

[42] Jacod Jean. Weak and strong solutions of stochastic differential equations. *Stochastics*, 3(1-4):171–191, 1980.

[43] Nathan Kallus and Xiaojie Mao. Stochastic optimization forests. *arXiv preprint*, page arXiv:2008.07473, 2020.

[44] Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Data-driven sample average approximation with covariate information. *Optimization Online*, page preprint, 2020.

[45] Rohit Kannan, Güzin Bayraksan, and James R Luedtke. Residuals-based distributionally robust optimization with covariate information. *arXiv preprint*, page arXiv:2012.01088, 2020.

[46] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[47] Thomas Kurtz. Weak and strong solutions of general stochastic models. *Electronic Communications in Probability*, 19:1–16, 2014.

[48] Rémi Lassalle. Causal transference plans and their monge-kantorovich problems. *arXiv preprint arXiv:1303.6925*, 2013.

[49] Rémi Lassalle. Causal transport plans and their monge–kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, 2018.

[50] Mo Liu, Meng Qi, and Zuo-Jun Max Shen. End-to-end deep learning for inventory management with fixed ordering cost and its theoretical analysis. *Available at SSRN 3888897*, 2021.

[51] Liwan H Liyanage and J George Shanthikumar. A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4):341–348, 2005.

[52] Gar Goei Loke, Qinshen Tang, and Yangge Xiao. Decision-driven regularization: Harmonizing the predictive and prescriptive. *Available at SSRN*, page 3623006, 2020.

[53] Miguel Angel Muñoz, Salvador Pineda, and Juan Miguel Morales. A bilevel framework for decision-making under uncertainty with contextual information. *Omega*, 108:102575, 2022.

[54] Afshin Oroojlooyjadid, Lawrence V Snyder, and Martin Takáč. Applying deep learning to the newsvendor problem. *IISE Transactions*, 52(4):444–463, 2020.

[55] G Ch Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20(3):1406–1420, 2010.

[56] Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[57] Georg Ch Pflug and Alois Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012.

[58] Georg Ch Pflug and Alois Pichler. *Multistage stochastic optimization*. Springer, 2014.

[59] Georg Ch Pflug and Alois Pichler. Dynamic generation of scenario trees. *Computational Optimization and Applications*, 62(3):641–668, 2015.

[60] Georg Ch Pflug and Alois Pichler. From empirical observations to tree models for stochastic optimization: convergence properties. *SIAM Journal on Optimization*, 26(3):1715–1740, 2016.

[61] Alois Pichler and Alexander Shapiro. Mathematical foundations of distributionally robust multistage optimization. *arXiv preprint arXiv:2101.02498*, 2021.

[62] Krzysztof Postek and Dick den Hertog. Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. *INFORMS Journal on Computing*, 28(3):553–574, 2016.

[63] Meng Qi, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu, and Zuojun (Max) Shen. A practical end-to-end inventory management model with deep learning. *Management Science*, page (Forthcoming), 2021.

[64] Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de Mello. Distributionally robust newsvendor problems with variation distance. *Optimization Online*, page preprint, 2017.

[65] Ludger Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.

[66] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[67] Prateek R Srivastava, Yijie Wang, Grani A Hanasusanto, and Chin Pang Ho. On data-driven prescriptive analytics with side information: A regularized nadaraya-watson approach. *arXiv preprint arXiv:2110.04855*, 2021.

[68] Bradley Sturt. A nonparametric algorithm for optimal stopping based on robust optimization. *arXiv preprint arXiv:2103.03300*, 2021.

[69] Anirudh Subramanyam, Chrysanthos E Gounaris, and Wolfram Wiesemann. K-adaptability in two-stage mixed-integer robust optimization. *Mathematical Programming Computation*, pages 1–32, 2019.

[70] L Beril Toktay and Lawrence M Wein. Analysis of a forecasting-production-inventory system with stationary demand. *Management Science*, 47(9):1268–1281, 2001.

[71] Theja Tulabandhula and Cynthia Rudin. Machine learning with operational costs. *Journal of Machine Learning Research*, 14:1989–2028, 2013.

[72] David Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[73] Tianlin Xu, Li K Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. *arXiv preprint arXiv:2006.08571*, 2020.

[74] Toshio Yamada and Shinzo Watanabe. On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University*, 11(1):155–167, 1971.

[75] Luhao Zhang, Jincheng Yang, and Rui Gao. A simple duality proof for wasserstein distributionally robust optimization. *arXiv preprint arXiv:2205.00362*, 2022.

[76] Kaijie Zhu and Ulrich W Thonemann. An adaptive forecasting algorithm and inventory policy for products with short life cycles. *Naval Research Logistics (NRL)*, 51(5):633–653, 2004.

[77] Taozeng Zhu, Jingui Xie, and Melvyn Sim. Joint estimation and robustness optimization. *Management Science*, 2021.

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) *γ 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

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

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 γ 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 $\gamma \in \Gamma(\widehat{\mathbb{P}}, \mathbb{P})$.

EXAMPLE EC.1. Here we show a few examples of causal transport. A transport plan induced by a causal transport map $\boldsymbol{T} : \mathcal{X} \times \mathcal{Z} \to \mathcal{X} \times \mathcal{Z}$ is causal. Recall that $\boldsymbol{T}$ is causal if it is in the form $\boldsymbol{T}(x,z) = (T_1(x), T_2(x,z))$, where $T_1 : \mathcal{X} \to \mathcal{Z}$ and $T_2 : \mathcal{X} \times \mathcal{Z} \to \mathcal{Z}$ are measurable. To see why this is true, let $\gamma = (\mathrm{Id} \times T)_{\#}\widehat{\mathbb{P}}$, then take any two points $(\widehat{x}, \widehat{z})$, we have

$$(\mathrm{Proj}_X)_{\#}\gamma(\mathrm{d}x|\widehat{x},\widehat{z}) = \delta_{T_1(\widehat{x})}$$

which is independent of the choice of $\widehat{z}$.

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 (2), 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 γ is causal.

EC.2. Sup of Convex Functions

LEMMA EC.2 **(Dual Objective Function)**. *The dual objective function h has the following properties. Let $\mathcal{I} = \{h < \infty\}$. Then*

(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 monotonously decreasing in λ, therefore we can find κ such that h is infinite for smaller λ, and finite for greater λ.

(II) h is a combination of supremums and expectations of convex functions, therefore h is convex. Since $h < \infty$ in $\mathcal{I}$, h is continuous in $\mathcal{I}$ with only a possible exception at $\kappa \in \mathcal{I}$. Notice that

$$\begin{aligned}
\liminf_{\lambda\downarrow\kappa} F_{(x)}(\lambda;\widehat{x}) &= \liminf_{\lambda\downarrow\kappa} \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{G_{(z)}(\lambda;x,\widehat{Z})\right\} \mid \widehat{X}=\widehat{x}\right] - \kappa \mathsf{d}(x,\widehat{x})^p \\
&\geq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\liminf_{\lambda\downarrow\kappa}\sup_{z\in\mathcal{Z}}\left\{G_{(z)}(\lambda;x,\widehat{Z})\right\} \mid \widehat{X}=\widehat{x}\right] - \kappa \mathsf{d}(x,\widehat{x})^p \\
&\geq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{\liminf_{\lambda\downarrow\kappa} G_{(z)}(\lambda;x,\widehat{Z})\right\} \mid \widehat{X}=\widehat{x}\right] - \kappa \mathsf{d}(x,\widehat{x})^p \\
&\geq \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{G_{(z)}(\kappa;x,\widehat{Z})\right\} \mid \widehat{X}=\widehat{x}\right] - \kappa \mathsf{d}(x,\widehat{x})^p = F_{(x)}(\kappa;\widehat{x}).
\end{aligned}$$

Similarly

$$\begin{aligned}
\liminf_{\lambda\downarrow\kappa} h(\lambda) &= \kappa\rho^p + \liminf_{\lambda\downarrow\kappa} \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{F_{(x)}(\lambda;\widehat{X})\right\}\right] \\
&\geq \kappa\rho^p + \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}}\left\{F_{(x)}(\kappa;\widehat{X})\right\}\right] = h(\kappa).
\end{aligned}$$

Therefore 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 (1)-(3).

LEMMA EC.3 **(Exchange sup and derivative for Convex Functions)**. *Let Λ 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] semicontinous envelope of f. For every $\varepsilon > 0$, define the ε-argmax set Ω_ε and $\overline{D}, \underline{D}$ by*

$$\begin{aligned}
&\Omega_\varepsilon(\lambda) := \{\alpha\in\Lambda : F_\alpha(\lambda) \geq \Phi(\lambda) - \varepsilon\}, \\
&\overline{D}_\varepsilon(\lambda) := \sup_{\alpha\in\Omega_\varepsilon(\lambda)} f^*_\alpha(\lambda), \qquad \overline{D}(\lambda) = \lim_{\varepsilon\to 0}\overline{D}_\varepsilon(\lambda), \\
&\underline{D}_\varepsilon(\lambda) := \inf_{\alpha\in\Omega_\varepsilon(\lambda)} f_{\alpha *}(\lambda), \qquad \underline{D}(\lambda) = \lim_{\varepsilon\to 0}\underline{D}_\varepsilon(\lambda).
\end{aligned}$$

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 λ, then we can find $\alpha\in\Lambda$ such that*

$$f^*_\alpha(\lambda_1) \leq \phi_*(\lambda) + \delta, \qquad F_\alpha(\lambda_2) \geq \Phi(\lambda) - \epsilon.$$

If $\lambda_2\in\mathcal{I}$ such that $\lambda_2 > \lambda$ is sufficiently close to λ, 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. Φ is the sup of a family of convex functions, so Φ is convex. Since Φ and F_α are convex and finite in $\mathcal{I}$, they have locally Lipschitz, monotonously increasing derivatives ϕ and f_α. Monotonicity implies f_α^* and ϕ^* [resp. $f_{\alpha *}$ and ϕ_*] are right [resp. left] continuous, and thus convexity implies for $\lambda < \mu$,

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

(I) ε-argmax set Ω_ε is never empty by definition. Therefore, $\underline{D}_\varepsilon(\lambda) \le \overline{D}_\varepsilon(\lambda)$ holds for all ε. As $\varepsilon \to 0$, $\Omega_\varepsilon(\lambda)$ shrinks, so $\overline{D}_\varepsilon(\lambda) \downarrow \overline{D}(\lambda)$, $\underline{D}_\varepsilon(\lambda) \uparrow \underline{D}(\lambda)$, we have $\underline{D}(\lambda) \le \overline{D}(\lambda)$.

(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

$$F_\alpha(\mu) - \varepsilon \le \Phi(\mu) - \varepsilon \le F_\beta(\mu) \le F_\beta(\lambda) + (\mu - \lambda) f_{\beta *}(\mu) \le \Phi(\lambda) + (\mu - \lambda) f_{\beta *}(\mu),$$
$$F_\beta(\lambda) - \varepsilon \le \Phi(\lambda) - \varepsilon \le F_\alpha(\lambda) \le F_\alpha(\mu) - (\mu - \lambda) f_\alpha{}^*(\lambda) \le \Phi(\mu) - (\mu - \lambda) f_\alpha{}^*(\lambda).$$

By these two inequalities, we conclude

$$-\varepsilon + (\mu - \lambda) f_\alpha^*(\lambda) \le \Phi(\mu) - \Phi(\lambda) \le \varepsilon + (\mu - \lambda) f_{\beta *}(\mu),$$
$$\Rightarrow -\frac{\varepsilon}{\mu - \lambda} + f_\alpha^*(\lambda) \le \frac{\Phi(\mu) - \Phi(\lambda)}{\mu - \lambda} \le \frac{\varepsilon}{\mu - \lambda} + f_{\beta *}(\mu).$$

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 (2), 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 Φ is continuous in the interior of $\mathcal{I}$, we can let λ_1 and λ_2 be close enough to λ 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

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

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

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

Lemma EC.4. *With the same notations as the previous lemma, let Λ be a Euclidean space. Suppose for each $\lambda \in \mathrm{Int}(\mathcal{I})$, $F_\alpha(\lambda)$ is upper semicontinuous in α, 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 δ, $\kappa' = \kappa - \delta$ and $\mu' = \mu + \delta$ are also inside $\mathrm{Int}(\mathcal{I})$.

(I) By Lemma EC.3 (2), $\phi_*(\lambda) \le \underline{D}(\lambda) \le \overline{D}(\lambda) \le \phi^*(\lambda)$, and since λ is in the interior of $\mathcal{I}$, Φ is locally Lipschitz, $\underline{D}(\lambda), \overline{D}(\lambda)$ are finite. Thus for some small ε, $\underline{D}_\varepsilon(\lambda)$ and $\overline{D}_\varepsilon(\lambda)$ are finite. This implies that Ω_ε is bounded, otherwise $|f_\alpha(\lambda)| \to \infty$ as $\alpha \to \infty$. Because F_α is upper semicontinuous, Ω_ε 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 λ, since $\Omega_0(\lambda) \subset \Omega_\varepsilon(\lambda)$ for any ε, 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 λ_n, $\Omega_0(\lambda_n)$ is nonempty, so we can find α_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_{α_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 Φ, 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_{α_n}, for $\lambda \in [\lambda_n, \mu]$,

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

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

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

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

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

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

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

by convexity of F_{α_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_{α_n} are increasing between κ and μ, so they are uniformly bounded, thus F_{α_n} are uniformly Lipschitz.

Since f_{α_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_{α_n} are uniformly bounded and equicontinuous in $[\kappa,\mu]$, by Arzelà-Ascoli Lemma it admits a subsequence uniformly converging to some F_∞, and since F_α is upper semicontinuous in α, 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 λ_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 β part is exactly symmetric to the α, so we omit here.

(III) This is the consequence of part (2) and (EC.3).

EC.3. Proofs for Section 3.1

Proof of Theorem 1. What remains to be proved is the strong duality $v_P \geq v_D$. For each $x\in\mathcal{X}$, $\widehat{z}\in\mathcal{Z}$ we denote

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

It is a linearly decreasing function of λ. Their supremum

$$\Upsilon(\lambda; x, \widehat{z}) := \sup_{z\in\mathcal{Z}} \left\{G_{(z)}(\lambda; x, \widehat{z})\right\} \tag{EC.5}$$

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

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

is a family of decreasing convex functions of λ. Their supremum

$$\Phi(\lambda;\widehat{x}) := \sup_{x\in\mathcal{X}} \left\{F_{(x)}(\lambda;\widehat{x})\right\} \tag{EC.6}$$

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 (κ,∞) and infinite in $[0,\kappa)$. Moreover, in the case $\kappa < \infty$, h attains its global minimum at $\lambda^* \geq \kappa$. Therefore we can separate into the following cases.

Case 1: $\kappa = \infty$

This means $h(\lambda) = \infty$ for any $\lambda \geq 0$, therefore $v_D = \infty$. Now fix $\lambda > 0$, then

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

We may assume

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

otherwise $v_P = \infty$ because $\widehat{\mathbb{P}}$ is feasible, and there is nothing else to prove. For each $\widehat{X}$ we can find an $X = T_1(\widehat{X}) \in \mathcal{X}$, such that

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

We can find $Z = T_2(\widehat{X},\widehat{Z}) \in \mathcal{Z}$, such that

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

Denote $\gamma_1 = ((T_1,T_2)\otimes \mathrm{id}_{\mathcal{X}\times\mathcal{Z}})_\sharp\widehat{\mathbb{P}}$, then $((X,Z),(\widehat{X},\widehat{Z})) \sim \gamma_1$, and denote the distance between $(\widehat{X},\widehat{Z})$ and (X,Z) by

$$D = \mathbb{E}_{\gamma_1}\left[\mathsf{d}(X,\widehat{X})^p + \mathsf{d}(Z,\widehat{Z})^p\right],$$

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 γ_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 γ_θ. Then

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

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

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

and since λ can be arbitrarily large, we have

$$v_P = \infty = v_D.$$

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

Fix some small $\delta > 0$, $\varepsilon > 0$. Applying Lemma EC.3 on (EC.6), for $\widehat{x} \in \mathcal{X}$ we can find $\overline{x}, \underline{x} \in \mathcal{X}$ such that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda^+}F_{(\underline{x})}(\lambda_1;\widehat{x}) &\leq \frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Phi(\lambda^*;\widehat{x}) + \delta, & \frac{\mathrm{d}}{\mathrm{d}\lambda^-}F_{(\overline{x})}(\lambda_2;\widehat{x}) &\geq \frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Phi(\lambda^*;\widehat{x}) - \delta, \\ F_{(\underline{x})}(\lambda_1,\widehat{x}) &\geq \Phi(\lambda^*,\widehat{x}) - \varepsilon, & F_{(\overline{x})}(\lambda_2,\widehat{x}) &\geq \Phi(\lambda^*,\widehat{x}) - \varepsilon\end{aligned}$$

for $\kappa < \lambda_1 < \lambda^* < \lambda_2$ and λ_1, λ_2 sufficiently close to λ^*. Fix $x \in \mathcal{X}$. Apply Lemma EC.3 on (EC.5), for $\widehat{z} \in \mathcal{Z}$ we can find $\overline{z}, \underline{z} \in \mathcal{Z}$ such that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda^+}G_{(\underline{z})}(\lambda_3;x,\widehat{z}) &\leq \frac{\mathrm{d}}{\mathrm{d}\lambda^-}\Upsilon(\lambda_1;x,\widehat{z}) + \delta, & \frac{\mathrm{d}}{\mathrm{d}\lambda^-}G_{(\overline{z})}(\lambda_4;x,\widehat{z}) &\geq \frac{\mathrm{d}}{\mathrm{d}\lambda^+}\Upsilon(\lambda_2;x,\widehat{z}) - \delta, \\ G_{(\underline{z})}(\lambda_3;x,\widehat{z}) &\geq \Upsilon(\lambda_1,x,\widehat{z}) - \varepsilon, & G_{(\overline{z})}(\lambda_4;x,\widehat{z}) &\geq \Upsilon(\lambda_2,x,\widehat{z}) - \varepsilon\end{aligned}$$

for $\kappa < \lambda_3 < \lambda_1 < \lambda^* < \lambda_2 < \lambda_4$ and λ_3, λ_4 sufficiently close to λ_1, λ_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 λ^* 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 $\hat{\gamma} = (\mathrm{id}_{\mathcal{X}\times\mathcal{Z}} \otimes \mathrm{id}_{\mathcal{X}\times\mathcal{Z}})_\sharp\widehat{\mathbb{P}}$ 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_D at some $\lambda^* \in \mathcal{I}$, so $h'(\lambda^*+) \geq 0$ and $h'(\lambda^*-) \leq 0$ (if $\lambda^* > \kappa$), so

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

where

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

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

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

Therefore,

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

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

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

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

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

In the mean time,

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

Similarly

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

Therefore,

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

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

$$v_P \ge v_D - 2\varepsilon.$$

Since ε can be taken arbitrarily small, $v_P \ge v_D$.

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

$$\overline{d} = \mathbb{E}_{\overline{\gamma}}\left[\mathsf{d}(\overline{X}, \widehat{X})^p + \mathsf{d}(\overline{Z}, \widehat{Z})^p\right] \le \rho^p + 2\delta.$$

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

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

Let $q^\delta := \frac{\rho^p}{\rho^p + 2\delta} \le 1$. Define $\gamma_\delta^\varepsilon := q^\delta \overline{\gamma} + (1 - q^\delta)\hat{\gamma}$, 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

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

thus

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

using $\lambda^* = 0$. Let $\delta \to 0$, $q^\delta \to 1$, we have $v_P \ge v_D - 2\varepsilon$, and by taking $\varepsilon \to 0$ we have $v_P \ge v_D$.

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

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

$$\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\Phi(\kappa'; \widehat{X}) - \Phi(\kappa; \widehat{X})\right] = h(\kappa') - h(\kappa) = \infty. \tag{EC.7}$$

We denote

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

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.7) 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

$$\begin{aligned}
R &< \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[F_{(\underline{X})}(\kappa'; \widehat{X}) - \Phi(\kappa; \widehat{X})\right] \\
&\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[F_{(\underline{X})}(\kappa'; \widehat{X}) - F_{(\underline{X})}(\kappa; \widehat{X})\right] \\
&= \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\kappa'; \underline{X}, \widehat{Z}) - \Upsilon(\kappa; \underline{X}, \widehat{Z}) \mid (\underline{X}, \widehat{X})\right] + (\kappa - \kappa')\mathsf{d}(\underline{X}, \widehat{X})^p\right].
\end{aligned} \tag{EC.8}$$

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

$$\begin{aligned} F_{(\widehat{X})}(\kappa';\widehat{X}) &\le F_{(\underline{X})}(\kappa';\widehat{X}),\\ \kappa'\mathsf{d}(\underline{X},\widehat{X})^p &\le \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\kappa';\underline{X},\widehat{Z})-\Upsilon(\kappa';\widehat{X},\widehat{Z})\mid \widehat{X}\right],\\ \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\kappa'\mathsf{d}(\underline{X},\widehat{X})^p\right] &\le \mathbb{E}_{(\underline{X},\widehat{X})\sim\underline{\gamma}_1}\left[\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\Upsilon(\kappa';\underline{X},\widehat{Z})-\Upsilon(\kappa';\widehat{X},\widehat{Z})\mid (\underline{X},\widehat{X})\right]\right]. \end{aligned} \tag{EC.9}$$

We denote

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

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.8) and (EC.9) as

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

Thus we can pick $\underline{Z}=\underline{T}_2(\widehat{X},\widehat{Z})\in\mathcal{Z}^*(\kappa';\underline{X},\widehat{Z})$, which induces γ_2, such that

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

and simultaneously ensure

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

In conclusion, we have

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

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

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

Therefore

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

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 limit, we have

$$v_P \ge \kappa'\rho^p + v_D - \kappa\rho^p - 2\varepsilon = v_D - 2\varepsilon - (\kappa-\kappa')\rho^p.$$

Since this is true for any $\kappa' < \kappa$ and $\varepsilon > 0$, we may take $\kappa' \to \kappa$ and $\varepsilon \to 0$ so $v_P \ge v_D$.

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 (2), 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 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

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

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

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

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

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,

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

as $x \to \infty$. By Lemma EC.4 (2) we can find $\overline{x}$ and $\underline{x}$ such that

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

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

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

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

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

and

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

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

$$\mathbb{E}_{\mathbb{P}^*}[\Psi(f(X),Z)] = q^*\mathbb{E}_{\overline{\mathbb{P}}}[\Psi(f(\overline{X}),\overline{Z})] + (1-q^*)\mathbb{E}_{\overline{\mathbb{P}}}[\Psi(f(\overline{X}),\overline{Z})] = h(\lambda^*) = v_D = v_P$$

it is optimal.

Note that this optimal solution is

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

Now we first consider the following linear optimization problem,

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

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, 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

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

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, Ψ can be written as

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

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

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

Denote

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

and

$$\ell_k(w) := \frac{1}{\sum_{i=1}^{n_k}\hat{p}_{ki}}\sum_{i=1}^{n_k}\hat{p}_{ki}\ell^{\widehat{z}_{ki}}(w) = \beta_k^\top w + b_k, \tag{EC.10}$$

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

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

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

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

Using notation introduced in (6),

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

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

Note that R_p is a convex function in λ and B, $\ell_k(f(\widehat{x}_k)) = \ell_k(B^\top \widehat{x}_k + \delta)$ is affine in B and δ, so the right hand side of the last expression is convex in λ and B as well. Hence (4) is a convex program:

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

where ℓ_k is an affine function defined by (EC.10) and R_p is a convex function defined by (6).

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

$$\begin{aligned}\sup_{z\in\mathcal{Z}} \left\{ \Psi(w,z) - \lambda\|z-\widehat{z}_{ki}\|^2 \right\} &= \Psi(w,\widehat{z}_{ki}) + \sup_{z\in\mathcal{Z}} \left\{ (A^\top w+\alpha)^\top (z-\widehat{z}_{ki}) - \lambda\|z-\widehat{z}_{ki}\|^2 \right\} \\ &= \Psi(w,\widehat{z}_{ki}) + \frac{|A^\top w+\alpha|^2}{4\lambda}.\end{aligned}$$

Since Ψ is affine in z, the conditional expectation of $\Psi(w,\widehat{z}_{ki})$ in $\widehat{Z}$ will be

$$\Psi_k(w) = \frac{\sum_{i=1}^{n_k} \hat{p}_{ki} \Psi(w,\widehat{z}_{ki})}{\sum_{i=1}^{n_k} \hat{p}_{ki}} = \beta_k^\top w + b_k,$$

where

$$\beta_k = \beta + A\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z}|\widehat{X}=\widehat{x}_k], \qquad b_k = b + \alpha^\top \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}[\widehat{Z}|\widehat{X}=\widehat{x}_k].$$

Then Ψ_k is affine in w. Now we can take supremum in x and expectation in $\widehat{X}$ as

$$\begin{aligned}&\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\sup_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}} \left\{ \Psi(f(x),z) - \lambda\|z-\widehat{Z}\|^2 \right\} | \widehat{X} \right] - \lambda\|x-\widehat{X}\|^2 \right\} \right] \\ &\quad = \sum_{k=1}^{K} \left(\sum_{i=1}^{n_k} \hat{p}_{ki} \right) \sup_{x\in\mathcal{X}} \left\{ \Psi_k(f(x)) + \frac{|A^\top f(x)+\alpha|^2}{4\lambda} - \lambda\|x-\widehat{x}_k\|^2 \right\}.\end{aligned}$$

Suppose now $f(x) = B^\top x + \delta$ is affine. Then

$$\begin{aligned}&\sup_{x\in\mathcal{X}} \left\{ \Psi_k(f(x)) + \frac{|A^\top f(x)+\alpha|^2}{4\lambda} - \lambda\|x-\widehat{x}_k\|^2 \right\} - \Psi_k(f(\widehat{x}_k)) - \frac{|A^\top f(\widehat{x}_k)+\alpha|^2}{4\lambda} \\ &\quad = \sup_{x\in\mathcal{X}} \left\{ (B\beta_k)^\top (x-\widehat{x}_k) + \frac{|(BA)^\top (x-\widehat{x}_k)|^2 + 2(A^\top f(\widehat{x}_k)+\alpha)^\top (BA)^\top (x-\widehat{x}_k)}{4\lambda} - \lambda\|x-\widehat{x}_k\|^2 \right\} \\ &\quad = \sup_{x+\widehat{x}_k\in\mathcal{X}} \left\{ (B\beta_k)^\top x + \frac{|(BA)^\top x|^2 + 2(A^\top f(\widehat{x}_k)+\alpha)^\top (BA)^\top x}{4\lambda} - \lambda\|x\|^2 \right\} \\ &\quad = \sup_{x\in\mathcal{X}} \left\{ x^\top \left[\frac{1}{4\lambda}(BA)(BA)^\top - \lambda\,\mathrm{Id} \right] x + \left[\frac{A(A^\top f(\widehat{x}_k)+\alpha)}{2\lambda} + \beta_k \right]^\top B^\top x \right\}.\end{aligned}$$

Recall that $f(\widehat{x}_k) = B\widehat{x}_k + \delta$. For the sup to be finite, we need $\|BA\|_2 \leq 2\lambda$. By continuity, we can assert the strict inequality constraint $\|BA\|_2 < 2\lambda$ instead, and the sup will be

$$\frac{1}{4} \left[\frac{A(A^\top f(\widehat{x}_k)+\alpha)}{2\lambda} + \beta_k \right]^\top \left[\frac{1}{4\lambda}(BA)(BA)^\top - \lambda\,\mathrm{Id} \right]^{-1} \left[\frac{A(A^\top f(\widehat{x}_k)+\alpha)}{2\lambda} + \beta_k \right].$$

This is convex in λ and B, with quadratic convex constraints. It could be solve by a quadratic constrained convex program, and the corollary follows.

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 ϕ_k implies

$$\varphi_k(y_k) \le \phi_k \le \lambda^* \mathsf{d}(x, x_k) + \phi_k,$$

so $y_k \in I_k(x)$. Note that each $I_k(x)$ is an interval, since it is the sub-level set of a convex function φ_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 $y_1 \in I_1(x)$, $y_2 \in I_2(x)$, we know that I_1 and I_2 are disjoint if and only if we can find y_3 in between y_1 and y_2 outside both intervals. This implies that

$$\begin{aligned}
&\varphi_1(y_3) > \lambda^* \mathsf{d}(x, x_3) + \phi_1 \ge \lambda^* \mathsf{d}(x, x_1) + \varphi_1(y_1),\\
&\varphi_1(y_3) > \lambda^* \mathsf{d}(x, x_3) + \phi_1 \ge \lambda^* \mathsf{d}(x, x_1) + \varphi_1(y_2) - \lambda^* \mathsf{d}(x_1, x_2),\\
&\varphi_2(y_3) > \lambda^* \mathsf{d}(x, x_3) + \phi_2 \ge \lambda^* \mathsf{d}(x, x_2) + \varphi_2(y_2),\\
&\varphi_2(y_3) > \lambda^* \mathsf{d}(x, x_3) + \phi_2 \ge \lambda^* \mathsf{d}(x, x_2) + \varphi_2(y_1) - \lambda^* \mathsf{d}(x_1, x_2).
\end{aligned}$$

Since y_3 is between y_1 and y_2, we can find $\alpha, \beta \in [0,1]$ with $\alpha + \beta = 1$ and $y_3 = \alpha y_1 + \beta y_2$. By multiplying the first/fourth inequality with α and the second/third inequality with β then taking the sum, we have

$$\begin{aligned}
(\varphi_1 + \varphi_2)(y_3) &> \lambda^* (\mathsf{d}(x, x_1) + \mathsf{d}(x, x_2)) + \alpha(\varphi_1 + \varphi_2)(y_1) + \beta(\varphi_1 + \varphi_2)(y_2) - \lambda^* \mathsf{d}(x_1, x_2)\\
&\ge \alpha(\varphi_1 + \varphi_2)(y_1) + \beta(\varphi_1 + \varphi_2)(y_2),
\end{aligned}$$

using the triangle inequality. However, this contradicts with 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 $\hat{f} = f|_{\widehat{\mathcal{X}}} \in \hat{\mathcal{F}}$ be the restriction of f on the set $\widehat{X}$, then

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

By taking the infimum over $f \in \mathcal{F}$ we would have $v_{\mathsf{D}} \ge v_{\hat{\mathsf{D}}}$. On the other hand, for the minimizer λ^* and $\hat{f}^* \in \hat{\mathcal{F}}$ of (9), let $f \in \mathcal{F}$ be an extension in $\cap_k I_k(x)$, then for every x we have

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

Therefore,

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

Thus we complete the proof of the theorem. □

EC.5. Proofs for Examples in Section 4

Proof of Example 6. Since f is real-valued and Ψ is convex in w, we use Theorem 3, so it has the following reformulation

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

with

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

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

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

This can be transformed into a linear programming problem

$$\begin{aligned}&\inf_{w_k,\lambda}\ \lambda\rho+\frac{1}{n}\sum_{k=1}^{K}c_j\\ &s.t.\begin{cases}c_j\geq\sum_{i=1}^{n_j}c_{kji}-\lambda n_j \mathsf{d}(\widehat{x}_k,\widehat{x}_j),\forall j,k\\ c_{kji}\geq w_k-\widehat{z}_{ji},\forall k,j,i\\ c_{kji}\geq\widehat{z}_{ji}-w_k,\forall k,j,i\\ \lambda\geq 1\end{cases}\end{aligned}$$

□

Proof of Example 7. Recall that the problem could be reformulated as

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

where

$$\varphi(w;\lambda;\widehat{x})=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\sup_{z\in\mathcal{Z}}\left\{-w(D_0-z(w-w_0))-\lambda \mathsf{d}(z,\widehat{Z})\right\}|\widehat{X}=\widehat{x}\right].$$

Observe that the supremum over z is infinite if $|w(w-w_0)|>\lambda$, otherwise the supremum is achieved at $z=\widehat{Z}$, thereby

$$\begin{aligned}\varphi(w;\lambda;\widehat{x})&=\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[-w(D_0-\widehat{Z}(w-w_0))|\widehat{X}=\widehat{x}\right]+\mathbf{1}_{\{|w(w-w_0)|>\lambda\}}\infty\\ &=-w\left(D_0-\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\widehat{Z}|\widehat{X}=\widehat{x}\right](w-w_0)\right)+\mathbf{1}_{\{|w(w-w_0)|>\lambda\}}\infty.\end{aligned}$$

Therefore, the dual problem is further equivalent to

$$\begin{aligned}&\inf_{\substack{f:\mathcal{X}\to\mathbb{R}\\ \lambda\geq\max\|f(f-w_0)\|}} \left\{\lambda\rho+\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{X}}}\left[\max_{1\leq j\leq K}\left\{-f(\widehat{x}_j)\left(D_0-\mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\widehat{Z}|\widehat{X}\right](w(\widehat{x}_j)-w_0)\right)-\lambda \mathsf{d}(\widehat{x}_j,\widehat{X})\right\}\right]\right\}\\ =&\inf_{\substack{w_k\in\mathbb{R}\\ \lambda\geq\max_k\|w_k(w_k-w_0)\|}} \left\{\lambda\rho+\sum_{k=1}^{K}\hat{p}_k\left[\max_{1\leq j\leq K}\left\{-w_j\left(D_0-\bar{z}_k(w_j-w_0)\right)-\lambda \mathsf{d}(\widehat{x}_j,\widehat{x}_k)\right\}\right]\right\},\end{aligned}$$

where $\hat{p}_k = \sum_{i=1}^{n_k} \hat{p}_{ki}$ and $\bar{z}_k = \mathbb{E}_{\widehat{\mathbb{P}}_{\widehat{Z}|\widehat{X}}}\left[\widehat{Z}|\widehat{X} = \widehat{x}_k\right]$. Then it could be transformed into a quadratic programming problem shown in Section 4 by adding dummy variables.

□