Optimal Transport in the Face of Noisy Data

Bart P.G. Van Parys*[1]

[1]*Sloan School of Management, MIT*

February 8, 2021

Abstract

Optimal transport distances are popular and theoretically well understood in the context of data-driven prediction. A flurry of recent work has popularized these distances for data-driven decision-making as well although their merits in this context are far less well understood. This in contrast to the more classical entropic distances which are known to enjoy optimal statistical properties. This begs the question when, if ever, optimal transport distances enjoy similar statistical guarantees. Optimal transport methods are shown here to enjoy optimal statistical guarantees for decision problems faced with noisy data.

1 Introduction

Let $\mathcal{P}$ be a family of probability measures over a space Ξ and let P be an unknown probability measure in this family. Many problems in the machine learning community attempt at heart to learn the unknown data generating probability measure from a finite collection of independent data observations

$$\xi_i \sim P \quad \forall i \in [1, \ldots, N]. \tag{1}$$

The most obvious application of this class of learning problems is perhaps density estimation [24]. Depending on whether the family $\mathcal{P}$ is the entire probability simplex $\mathcal{P}(\Xi)$ over Ξ or merely a parametrized subset such methods can be classified as nonparametric or parametric, respectively. That is not to say that all machine learning problems with the data observation model (1) are density estimation problems in disguise. Often, one is not interested in the probability measure P *per se* but rather only a certain aspect. In predictive problems, in which the data consists of both dependent and independent variables, typically only the conditional expectation of the dependent variable given the independent variables is of interest. In prescriptive problems, which are here the problems of primary interest, the problem of immediate concern is to find an $0 < \epsilon$-suboptimal[1] solution to a stochastic optimization problem, i.e.,

$$z(P) \in \arg_\epsilon \inf_{z\in Z} \{\mathbf{E}_P\left[\ell(z, \xi)\right] = \textstyle\int \ell(z, \xi)\, \mathrm{d}P(\xi)\}. \tag{2}$$

A wide spectrum of decision problems can be cast as instances of (2). Shapiro et al. [23] point out, that (2) can be viewed as the first stage of a two-stage stochastic program, where the loss function $\ell : Z \times \Xi \to \mathrm{R}$ embodies the optimal value of a subordinate second-stage problem. Alternatively, problem (2) may also be interpreted as a generic learning problem in the spirit of statistical learning theory. Rather than learning the unknown probability measure P, the primary objective in data-driven decision-making is to learn the cost function and perhaps even better an ϵ-suboptimal decision to (2) directly from data.

*vanparys@mit.edu

[1] Here $\arg_\epsilon \inf_{z\in Z} \mathbf{E}_P\left[\ell(z,\xi)\right]$ with $\epsilon > 0$ corresponds to the set $\{z \in Z \ : \ \mathbf{E}_P\left[\ell(z,\xi)\right] < \inf_{z\in Z} \mathbf{E}_P\left[\ell(z,\xi)\right] + \epsilon\}$.

In this paper we will denote the observational model described in Equation (1) as "noiseless". That is, the learner has access to uncorrupted independent samples from the probability measure of interest P. Clearly, given any finite amount of data the learner can not expect to learn the data generating probability measure exactly even in this noiseless regime. In case the probability measure P is only known to belong to the probability simplex $\mathcal{P}(\Xi)$, one reasonable substitute for P could be its empirically observed counterpart denoted here as $P_N := \sum_{i=1}^N \delta_{\xi_i}/N$. If on the other hand some prior information is available in the sense that the probability measure P is known to belong to a subset $\mathcal{P} \subset \mathcal{P}(\Xi)$, a maximum likelihood estimate [13] is often used instead. In the machine learning and robust optimization community such point estimates are widely known to be problematic when used naively in subsequent analysis. In particular, it is widely established both empirically as well as in theory that a sample average formulation

$$z(P_N) \in \arg_\epsilon \inf_{z \in Z} \mathbf{E}_{P_N}\left[\ell(z, \xi)\right] \tag{3}$$

which substitutes P with a mere point estimate P_N tends to disappoint ($\mathbf{E}_P\left[\ell(z(P_N), \xi)\right] > \mathbf{E}_{P_N}\left[\ell(z(P_N), \xi)\right]$) out of sample. That is, the actual cost observed out of sample breaks the predicted cost of the data-driven decision $z(P_N)$. This adversarial phenomenon is well known colloquially as the "Optimizer's Curse" [16] and is akin to the overfitting phenomenon in the context of prediction problems. Such adversarial phenomena related to over-calibration to observed data but poor performance on out-of-sample data can be attributed primarily to the treatment of mere point estimates as exact substitutes for the unknown probability measure.

Ambiguity sets consisting of all probability measures sufficiently compatible with the observed data can offer a better alternative to simple point estimates. As the data observations are here independent and identically distributed, their order is irrelevant, and ambiguity sets $\mathcal{A}_N(P_N) \subseteq \mathcal{P}$ can be made functions of the empirical probability measure P_N rather than the data itself. A large line of work in the robust optimization community, see [20] and references therein, focuses consequently on data-driven formulations of the form

$$z_{\mathcal{A}}(P_N) \in \arg_\epsilon \inf_{z \in Z} \sup \left\{\mathbf{E}_P\left[\ell(z, \xi)\right] \ : \ P \in \mathcal{A}_N(P_N)\right\}$$

which can be thought of as robust counterparts to the nominal sample average formulation stated in Equation (3). Robust formulations guard against over-calibrated decisions by forcing any decision to do well on all distributions sufficiently compatible with the observed data as opposed to only a single point estimate. The recent uptick in popularity of robust formulations is in no small part due to the fact that they are often just as tractable and typically enjoy superior statistical properties than their nominal counterparts. Much of the early literature [6, 28, 26] focused on ambiguity sets consisting of probability measures sharing certain given moments. More recent approaches [4] however consider ambiguity sets $\mathcal{A}_N(P_N) = \{P \in \mathcal{P} \ : \ D(P_N, P) \leq r_N\}$ which are based on a statistical distance $D : \mathcal{P}(\Xi) \times \mathcal{P}(\Xi) \to \mathrm{R} \cup \{+\infty\}$ instead. The latter ambiguity sets can hence be interpreted as the set of probability measures sufficiently close to the empirical probability measure P_N. Two qualitatively different statistical distances have recently positioned themselves as the front runners for data-driven decision-making and are now briefly discussed.

Optimal transport distances [10, 12] have received a lot of attention both in the context of data-driven decision-making as well as in the machine learning community at large [15, 18]. In the context of data-driven decision-making optimal transport distances have become very popular after [10] pointed out that their resulting robust formulations need not be intractable. Furthermore, the associated optimal transport ambiguity sets enjoy an interpretation as confidence intervals [11] for the unknown probability measure P when the radii r_N are judiciously chosen as function of the number N of observed data points. Perhaps the main competitor to optimal transport distances is the Kullback-Leibler or entropic divergence. We briefly recall the definition of entropic divergence between two measures μ and ν on the same space as

$$D_{\mathrm{KL}}(\mu,\nu) = \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

The entropic divergence is a particular member of the class of convex f-divergences which are well known [14] to yield tractable robust formulations. Interestingly, its associated ambiguity sets do not generally admit an interpretation as a confidence interval for the unknown probability measure. Unless the event set Ξ is finite, the associated entropic ambiguity set $\{P \in \mathcal{P} \;:\; D_{\mathrm{KL}}(P_N, P) \leq r_N\}$ does indeed not contain any continuous probability measure and hence also not necessarily P. Despite this observation, the interval

$$[\min_{P\in\mathcal{A}_N(P_N)} \mathbf{E}_P\left[\ell(z,\xi)\right], \max_{P\in\mathcal{A}_N(P_N)} \mathbf{E}_P\left[\ell(z,\xi)\right]]$$

nevertheless admits an interpretation [8] as a confidence interval for the unknown cost $\mathbf{E}_P\left[\ell(z,\xi)\right]$ for any decision $z \in Z$ when the radii r_N are judiciously chosen. Perhaps even more surprising, the associated entropic robust prescriptive formulation

$$z_{\mathrm{KL}}(P_N) \in \arg_\epsilon \inf_{z\in Z} \sup \left\{\mathbf{E}_P\left[\ell(z,\xi)\right] \;:\; P \in \mathcal{P},\; D_{\mathrm{KL}}(P_N, P) \leq r\right\} \tag{4}$$

can be shown [27] to enjoy optimal large deviation properties of a similar nature to those we will encounter in Section 5.

The previous discussion naturally begs the question when – if ever – optimal transport distances enjoy similar statistical guarantees in the context of prescriptive problems. Folklore belief suggests that optimal transport methods derive their superior empirical performance from their ability to guard against noisy or corrupted data. For instance, optimal transport methods can be interpreted as maximum likelihood estimation [21] in the context of predictive problems. This note indicates that optimal transport methods are similarly well suited for prescriptive problems facing noisy data. In that sense this note hopes to offer a theoretical justification of the perhaps surprising effectiveness and popularity of optimal transport methods for data-driven decision-making. We also show that any perceived dichotomy between entropic and optimal transport distances in the context of data-driven decision-making is in fact a false one and that a balance of both distances is better than either distance separately. Hence, we argue that entropic and optimal transport distance formulations should be perceived as complementary rather than as direct competitors for data-driven decision-making.

Organization In Section 2 we briefly recall the (entropic) optimal transport distance and its properties. We introduce our noisy data model and provide three illustrative examples in Section 3. In Section 4 we prove that the empirical distribution of noisy observational data satisfies a large deviation principle with a rate function which balances entropic divergence and entropic optimal transport distances. Finally, Section 5 illustrates the power of optimal transport distances in the context of both hypothesis testing and data-driven decision-making.

Topology We will assume that Ξ and Ξ' are Polish topological spaces and hence so is the product space $\Xi\times\Xi'$ when equipped with the product topology. We denote with $\mathcal{P}(\Xi)$, $\mathcal{P}(\Xi')$ and $\mathcal{P}(\Xi\times\Xi')$ as the sets of all Borel probability measures on the spaces Ξ, Ξ' and $\Xi\times\Xi'$, respectively. Following [7, Section 6.2] the probability simplices $\mathcal{P}(\Xi)$, $\mathcal{P}(\Xi')$ and $\mathcal{P}(\Xi\times\Xi')$ when equipped with the topology of weak convergence of probability measures are Polish spaces too. We denote with $D_{M'} : \mathcal{P}(\Xi')\times\mathcal{P}(\Xi') \to \mathrm{R}_+$ a metric compatible with the weak topology[2] on $\mathcal{P}(\Xi')$. Finally, we take Z to be a finite dimensional linear vector space equipped with its classical norm topology.

[2] A classical choice is to take $D_{M'} : \mathcal{P}(\Xi')\times\mathcal{P}(\Xi') \to [0,1]$ as the Lévy-Prokhorov [19] metric on $\mathcal{P}(\Xi')$.

2 Optimal Transport Distances

Given two measures μ and ν we will denote with $\mu \otimes \nu$ their product measure. Conversely, given a probability measure T on $\mathcal{P}(\Xi \times \Xi')$ we denote its marginal projection on $\mathcal{P}(\Xi)$ and $\mathcal{P}(\Xi')$ as $\Pi_\Xi T$ and $\Pi_{\Xi'} T$ respectively. We define

$$\mathcal{T}(\mu, \nu) := \{T \in \mathcal{P}(\Xi \times \Xi') \ : \ \Pi_\Xi T = \mu, \ \Pi_{\Xi'} T = \nu\}$$

as the set of all joint probability measures with given marginal distributions μ and ν. Measures in the set $\mathcal{T}$ can be interpreted to transport marginal μ to marginal ν. Furthermore, this set is nonempty as clearly we have $\mu \otimes \nu \in \mathcal{T}(\mu, \nu)$.

Definition 2.1 (Entropic Optimal Transport Distance)**.** Given a distance function $d : \Xi \times \Xi' \to \mathrm{R} \cup \{+\infty\}$. The entropic optimal transport distance between μ on Ξ and ν on Ξ' is defined as

$$D_W(\mu, \nu) := \inf_{T \in \mathcal{T}(\mu,\nu)} \int d(\xi, \xi') \, \mathrm{d}T(\xi, \xi') + D_{\mathrm{KL}}(T, \Pi_\Xi T \otimes \Pi_{\Xi'} T).$$

Note that we do not explicitly require the spaces Ξ and Ξ' to coincide here. Assume for a moment however that $\Xi = \Xi' = \mathrm{R}^n$ and $d(\xi', \xi) = \|\xi' - \xi\|_2^2$. Then, the entropic optimal transport distance modulo the regularization term $D_{\mathrm{KL}}(T, \Pi_\Xi T \otimes \Pi_{\Xi'} T)$ coincides with the classical Wasserstein distance between probability measures [22]. We remark that the entropic optimal transport distance is not a metric although it still enjoys distance-like properties [5]. Historically, the entropy term $D_{\mathrm{KL}}(T, \Pi_\Xi T \otimes \Pi_{\Xi'} T)$ has been considered primarily for its beneficial smoothing effect as indeed it allows the entropic optimal transport distance to be computed efficiently using the Sinkhorn Algorithm [5] in case the event sets $\Xi = \Xi'$ are finite. As for any $T \in \mathcal{T}(\mu, \nu)$ we have $D_{\mathrm{KL}}(T, \Pi_\Xi T \otimes \Pi_{\Xi'} T) = D_{\mathrm{KL}}(T, \mu \otimes \nu)$, the regularization term can be interpreted to encourage transportation plans which are not too different from the independent product coupling $\mu \otimes \nu$.

3 Noisy Observational Data

In deconvolution problems the learner attempts to estimate P on the basis of noisy observations, i.e.,

$$\xi'_i \sim_{\text{i.i.d.}} O_{\xi_i} \quad \forall i \in [1, \ldots, N], \tag{5}$$

where $\xi_1, \ldots, \xi_N$ are independent and identically distributed according to P. Hence, rather than having direct access to samples $\xi_i \in \Xi$ from the probability measure of interest, the learner must do with indirect noisy data $\xi'_i \in \Xi'$ instead. We do assume here that the observational map O is known. Furthermore, for some of our results the observational process will be required to be continuous in the sense of Assumption 3.1. We remark that this assumption is without much loss of generality. In particular, if Ξ and Ξ' are finite Assumption 3.1 is trivially satisfied.

Assumption 3.1. The map $O : \Xi \to \mathcal{P}(\Xi')$ is absolutely continuous with respect to a base measure m', i.e., $O_\xi \ll m'$ for all $\xi \in \Xi$. Consequently, there exists a density function $d : \Xi \times \Xi' \to \mathrm{R} \cup \{+\infty\}$ so that

$$\frac{\mathrm{d}O_\xi}{\mathrm{d}m'}(\xi') = \exp(-d(\xi, \xi')) \quad \forall \xi' \in \Xi'.$$

The relationship between the probability measure P' of the noisy observations ξ'_i and the probability measure P of the unobserved noiseless data ξ_i can be characterized as the convolution

$$P'(B) = (O \star P)(B) := \int O_\xi(B) \mathrm{d}P(\xi)$$

for all measurable sets $B \in \mathcal{B}(\Xi')$. Clearly, the unknown probability measure P is identifiable from its counterpart P' only if this convolution transformation is invertible. We will denote with the set $\mathcal{P}' := \{O \star P \in \mathcal{P}(\Xi') \,:\, P \in \mathcal{P}\}$ the family of potential distributions of our noisy data. We conclude this section by pointing out that the presented observational model is quite flexible and captures a wide variety of interesting settings.

Example 3.2 (Noiseless Data)**.** The choice $O_\xi = \delta_\xi$ for all $\xi \in X$ can be identified with a noiseless observation regime where δ_ξ denotes the Dirac measure at ξ, i.e., $\delta_\xi(B) = \mathbb{1}\{\xi \in B\}$ for all $B \in \mathcal{B}(\Xi')$. Here, the probability measure P' of the observation $\xi'_1, \ldots, \xi'_N$ coincides with the unknown probability measure P. That is,

$$P'(B) = (O \star P)(B) := \int \mathbb{1}\{\xi \in B\}\, \mathrm{d}P(\xi) = P(B)$$

for all measurable sets $B \in \mathcal{B}(\Xi')$. This setting corresponds to the observational setting described by Equation (1) in which data sampled from the unknown probability measure is observed directly uncorrupted by any noise.

Example 3.3 (Irrelevant Data)**.** The case $O_\xi = P'$ for some probability measure P' on $\mathcal{P}(\Xi')$ represents a setting in which the data is irrelevant. Here the observations are independent of the unknown probability measure P and consequently are wholly irrelevant. Indeed, the distribution of the noisy data is independent of P and given as

$$P'(B) = (O \star P)(B) := \int P'(B) \mathrm{d}P(\xi) = P'(B)$$

for all measurable subsets $B \in \mathcal{B}(\Xi')$. Clearly, under these circumstances the unknown probability measure P is simply not identifiable from the observed data unless $\mathcal{P}$ is a singleton in which case the learning problem is trivial.

Example 3.4 (Gaussian Noise)**.** Most practical examples are situated somewhere between the previously discussed corner cases. For the sake of exposition we will consider the case of Gaussian noise as a final example. Let Assumption 3.1 hold here with $\Xi = \Xi' = \mathrm{R}^n$, and that $d(\xi, \xi') = \|\xi - \xi'\|_2^2/(2\sigma^2)$ with noisy power $\sigma \geq 0$ and $m' = \mu'/(\sigma\sqrt{(2\pi)^n})$ with μ' the Lebesgue measure on R^n. Consequently here $Q_\xi = N(\xi, \sigma^2)$ a normal distribution with mean ξ and variance σ^2. The noisy data ξ'_i in Equation (5) follows the same distribution as

$$\xi_i + z_i \quad \forall i \in [1, \ldots, N],$$

where $\xi_1, \ldots, \xi_N$ and $z_1, \ldots, z_N$ are independent and identically distributed as P and $N(0, \sigma^2)$, respectively. This class of noisy observations interpolates between the noiseless regime in Example 3.2 when $\sigma^2 \to 0$ and the uninformative data of Example 3.3 as $\sigma^2 \to \infty$.

4 A Large Deviation Property

We will attempt to infer the unknown probability measure P from our noisy data based on its empirical probability measure $P'_N = \sum_{i=1}^N \delta_{\xi'_i}/N$. Clearly, considering the empirical probability measure rather than the noisy data directly imposes no loss of information as the order of the data points is of no consequence here. Our main observation will be that this sufficient statistic P'_N satisfies a large deviation principle [7], with a rate function which carefully balances an entropic optimal transport distance and an entropic divergence.

Theorem 4.1. *Let Assumption 3.1 hold. Then, the sufficient statistic P'_N satisfies for any open subset $\mathcal{O} \subseteq \mathcal{P}(\Xi')$ the large deviation lower bound*

$$-\inf_{P' \in \mathcal{O}} I(P', P) \;\leq\; \liminf_{N \to \infty} \frac{1}{N} \log \mathrm{Prob}_P[P'_N \in \mathcal{O}] \tag{6a}$$

and for any closed subset $\mathcal{C} \subseteq \mathcal{P}(\Xi')$ the large deviation upper bound

$$\limsup_{N\to\infty} \frac{1}{N} \log \operatorname{Prob}_P[P'_N \in \mathcal{C}] \leq -\inf_{P'\in\mathcal{C}} I(P', P). \tag{6b}$$

for the good[3] *and convex rate function*

$$I(P', P) = \inf_{Q\in\mathcal{P}(\Xi)} D_W(P', Q) + D_{\mathrm{KL}}(Q, P) + D_{\mathrm{KL}}(P', m') \geq 0. \tag{7}$$

Proof. Consider first the statistic $T'_N = \sum_{i=1}^N \delta_{(\xi_i,\xi'_i)}/N$. We first show that this statistic satisfies a large deviation property [7, Section 1.2]. Second, as we have that $P'_N = \Pi_{\Xi'} T'_N$, the large deviation inequalities (6a) and (6b) for P'_N can be established via a contraction princple [7, Theorem 4.2.1].

Let $T(P)$ be the joint distribution of the random variables (ξ_i, ξ'_i) and note that under Assumption 3.1 we have

$$T(P)(B) := \int \mathbb{1}\left\{(\xi,\xi') \in B\right\} \exp(-d(\xi,\xi'))\, \mathrm{d}m'(\xi')\mathrm{d}P(\xi) \quad \forall B \in \mathcal{B}(\Xi\times\Xi'). \tag{8}$$

Equivalently, $\mathrm{d}T(P)/\mathrm{d}(P\otimes m')(\xi,\xi') = \exp(-d(\xi,\xi'))$. Clearly, the noisy observational model in Equation (5) guarantees that each of the samples in the sequence $\{(\xi_i,\xi'_i)\}$ is independent and identically distributed. An empirical distribution T'_N of independent and identically distributed samples following distribution $T(P)$ enjoys a large deviation property with rate function $D_{\mathrm{KL}}(T', T(P))$ [7, Theorem 6.2.10]. Assume that $T' \ll T(P)$ as otherwise $D_{KL}(T', T(P)) = +\infty$. Under Assumption 3.1 the condition $T' \ll T(P)$ is furthermore equivalent to $\Pi_\Xi T' \ll P$ and $\Pi_{\Xi'} T' \ll m'$. We have under this assumption that

$$\begin{aligned}
&D_{KL}(T', T(P))\\
:=&\int \log\left(\frac{\mathrm{d}T'}{\mathrm{d}T(P)}(\xi,\xi')\right) \mathrm{d}T'(\xi,\xi')\\
=&\int \log\left(\frac{\mathrm{d}T'}{\mathrm{d}P\otimes m'}(\xi,\xi')\right) \mathrm{d}T'(\xi,\xi') - \int \log\left(\frac{\mathrm{d}T(P)}{\mathrm{d}P\otimes m'}(\xi,\xi')\right) \mathrm{d}T'(\xi,\xi')\\
=&D_{\mathrm{KL}}(T', P\otimes m') - \int \log\left(\exp(-d(\xi,\xi'))\right) \mathrm{d}T'(\xi,\xi')\\
=&D_{\mathrm{KL}}(T', P\otimes m') + \int d(\xi,\xi')\, \mathrm{d}T'(\xi,\xi')\\
=&\int d(\xi,\xi')\, \mathrm{d}T'(\xi,\xi') + D_{\mathrm{KL}}(T', \Pi_\Xi T'\otimes \Pi_{\Xi'}T') + D_{\mathrm{KL}}(\Pi_\Xi T', P) + D_{\mathrm{KL}}(\Pi_{\Xi'}T', m') \geq 0
\end{aligned}$$

where we use [21, Lemma 6.1] to establish the final equality. As we have that $P'_N = \Pi_{\Xi'} T_N$, a large deviation property for P'_N can now be established via a contraction principle [7, Theorem 4.2.1] as the projection operator $\Pi_{\Xi'} : \mathcal{P}(\Xi\times\Xi') \to \mathcal{P}(\Xi')$ is continuous. For any sequence $T_k \in \mathcal{P}(\Xi\times\Xi')$ with limit $\bar{T} \in \mathcal{P}(\Xi\times\Xi')$ we have by definition of the weak topology that

$$\int c(\xi,\xi')\, \mathrm{d}T_k(\xi,\xi') \to \int c(\xi,\xi')\, \mathrm{d}\bar{T}(\xi,\xi')$$

for all bounded and continuous functions $c : \Xi\times\Xi' \to \mathrm{R}$. Consequently, for any bounded and continuous function $c' : \Xi' \to \mathrm{R}$ we have that

$$\int c'(\xi')\, \mathrm{d}\Pi_{\Xi'}T_k(\xi') = \int c'(\xi')\, \mathrm{d}T_k(\xi,\xi') \to \int c'(\xi')\, \mathrm{d}\bar{T}(\xi,\xi') = \int c'(\xi')\, \mathrm{d}\Pi_{\Xi'}\bar{T}(\xi')$$

where we use that also $(\xi,\xi') \mapsto c'(\xi')$ is bounded and continuous as a map from $\Xi\times\Xi'$ to R. Marginal projection

[3] A rate function I is good if its sublevel sets $\{P' \in \mathcal{P}(\Xi') \ : \ I(P', P) \leq r\}$ for any $r \geq 0$ and $P \in \mathcal{P}$ are compact [7].

$T \mapsto \Pi_{\Xi'} T$ is indeed continuous as we have $\Pi_{\Xi'} T_k \to \Pi_{\Xi'} \bar{T}$ for any converging sequence $T_k \in \mathcal{P}(\Xi \times \Xi')$ with limit $\bar{T} \in \mathcal{P}(\Xi \times \Xi')$. Consequently, via a contraction princple [7, Theorem 4.2.1], the large deviation inequalities (6a) and (6b) of the statistic P'_N can be established for the rate function

$$\begin{aligned} & I(P', P) \\ = & \inf \{ D_{KL}(T', T(P)) \ : \ T' \text{ s.t. } \Pi_{\Xi'} T' = P' \} \\ = & \inf \{ D_{\mathrm{KL}}(T', T(P)) \ : \ Q \in \mathcal{P}(\Xi), \ T' \ll T(P) \text{ s.t. } \Pi_{\Xi} T' = Q, \Pi_{\Xi'} T' = P' \} \\ = & \inf \Big\{ \textstyle\int d(\xi, \xi') \, \mathrm{d}T'(\xi, \xi') + D_{\mathrm{KL}}(T', \Pi_{\Xi} T' \otimes \Pi_{\Xi'} T') + D_{\mathrm{KL}}(Q, P) + D_{\mathrm{KL}}(P', m') : \\ & \hfill Q \in \mathcal{P}(\Xi), \ T' \in \mathcal{T}(Q, P') \Big\} \\ = & \inf_{Q \in \mathcal{P}(\Xi)} D_W(P', Q) + D_{\mathrm{KL}}(Q, P) + D_{\mathrm{KL}}(P', m'). \end{aligned}$$

The joint convexity of the rate function $I(P', P) = \inf \{ D_{KL}(T', T(P)) \ : \ T' \text{ s.t. } \Pi_{\Xi'} T' = P' \}$ is inherited [17, Theorem 1.31] from the joint convexity of the objective function $D_{KL}(T', T)$ in (T', T), the linearity of $T(P)$ in P evident from Equation (8) and the linearity of the projection $\Pi_{\Xi'} T'$ in T'. □

We remark that large deviation inequalities generally are quite rough in nature as indeed (6a) and (6b) only pertain to open or closed sets, respectively. Theorem 4.1 states that the rate function $I(P', P)$ is always nonnegative and the fact that $I(P', P) = 0$ if and only if $P' = O \star P$ can easily be deduced from its proof and the observation that $D_{\mathrm{KL}}(T', T) = 0 \iff T' = T$. For any $\epsilon > 0$, the large deviation inequality (6b) despite its rough nature nevertheless implies

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathrm{Prob}_P [D_{M'}(P'_N, O \star P) \geq \epsilon] \leq - \min \{ I(P', P) \ : \ P' \in D_{M'}(P', P) \geq \epsilon \} < 0 \quad \forall P \in \mathcal{P}$$

where the minimum is indeed achieved as our good rate function has compact sublevel sets and the set of all $P' \in \mathcal{P}(\Xi')$ such that $D_{M'}(P', O \star P) \geq \epsilon$ is by definition closed and does not contain the distribution $O \star P$. Hence, our large deviation property immediately implies that the empirical probability measure P'_N converges in probability to $P' = O \star P$ with an increasing number of observations. In fact, the rate function can be interpreted as the appropriate yardstick with which to measure how fast this convergence takes place.

5 Applications

In this section we present two distinct statistical problems in which optimal transport distances become appropriate in the face of noisy observational data. That is, we indicate that our rate function I induces optimal statistical properties in two distinct problem settings. A hypothesis testing problem class will serve to illustrate the optimality of our considered optimal transport distance in a first predictive setting. This example enforces the findings of Rigollet and Weed [21] in so far that an optimal transport distance is shown to be sensible in a predictive context. The optimality of ambiguity sets associated with the rate function I is also established in a second prescriptive setting. Hence, we illustrate by means of example that optimal transport distances enjoy optimal statistical guarantees in both predictive and prescriptive settings in the face of noisy observational data.

5.1 Optimal Hypothesis Testing

We consider a hypothesis testing problem in which we need to determine whether the sequence of noisy data points $\xi'_1, \ldots, \xi'_N$ defined in Equation (5) is produced by the unobserved probability measure P_0 or alternatively by P_1. Hypothesis testing problems can be regarded as simple prediction problems in which $\mathcal{P} = \{P_0, P_1\}$

consists of only two probability measures. Classically, the considered decision rule to choose between these alternatives is denoted as a hypothesis test.

Definition 5.1 (Hypothesis test)**.** A hypothesis test $\tilde{h}$ is a measurable functions $\tilde{h} : \mathcal{P}(\Xi') \to \{P_0, P_1\}$. We denote with

$$\mathcal{R}(\tilde{h}) = \left\{P' \in \mathcal{P}(\Xi') \ : \ \tilde{h}(P') \neq P_0\right\}$$

its rejection region.

A hypothesis test $\tilde{h}$ has as interpretation that for data with empirical distribution P'_N, the null hypothesis (P_0 generated the data) is accepted if $\tilde{h}(P'_N) = P_0$, while its alternative is accepted (P_1 generated the data) if $\tilde{h}(P'_N) = P_1$. We associate with each hypothesis test its asymptotic exponential error rates

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_0}\left[\tilde{h}(P'_N) \neq P_0\right] = \limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_0}\left[P'_N \in \mathcal{R}(\tilde{h})\right], \tag{9}$$

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_1}\left[\tilde{h}(P'_N) \neq P_1\right] = \limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_1}\left[P'_N \notin \mathcal{R}(\tilde{h})\right]. \tag{10}$$

Clearly, the statistical performance of a hypothesis test should be based on how well it can balance the desire to keep both the first and second error probabilities small. It is quite common to only consider hypothesis tests which suffer type I errors at rate at most $-r$, i.e., hypothesis tests for which (9) $\leq -r$. Such hypothesis tests guarantee that the null hypothesis is not erroneously rejected all that often. Given this requirement, we would now like to find hypothesis tests which additionally enjoy an optimal type II error rate, i.e., a hypothesis tests for which the probabilities $\mathrm{Prob}_{P_1}\left[P'_N \notin \mathcal{R}(\tilde{h})\right]$ decay exponentially as fast as possible.

To that end let the δ-smoothed rate function be defined as

$$I^\delta(P', P) := \inf\left\{I(P'', P) \ : \ P'' \in \mathcal{P}(\Xi'),\ D_{M'}(P'', P') \leq \delta\right\}. \tag{11}$$

Fix a radius $r > 0$ and consider a family of hypothesis tests $\tilde{h}^\delta$ such that for all $\delta > 0$ we have that

$$\tilde{h}^\delta(P'_N) = \begin{cases} P_0 & \text{if } I^\delta(P'_N, P_0) \leq r, \\ P_1 & \text{otherwise.} \end{cases}$$

The proposed family of hypothesis tests can be shown to be almost asymptotically optimal using merely a large deviations argument. Due to the rough nature of the large deviation property, we will consider in the same spirit as [7, Section 7.1] for any hypothesis test $\tilde{h}$ also its $0 < \epsilon$-open inflated rejection regions

$$\mathcal{R}^\epsilon(\tilde{h}) = \left\{P'' \in \mathcal{P}(\Xi') \ : \ P' \in \mathcal{R}(\tilde{h}),\ D_{M'}(P'', P') < \epsilon\right\}.$$

Theorem 5.2. *The family of hypothesis tests $\tilde{h}^\delta$ satisfies for any $\delta > 0$ a type I error which satisfies*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_0}\left[P'_N \in \mathcal{R}(\tilde{h}^\delta)\right] \leq -r. \tag{12}$$

For any other hypothesis test $\tilde{h}$ that satisfies a type I error with

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_0}\left[P'_N \in \mathcal{R}^\epsilon(\tilde{h})\right] \leq -r \tag{13}$$

for some $\epsilon > 0$, there exists furthermore a $\delta' > 0$ (independent from P_1) so that for all $0 < \delta \leq \delta'$ we have

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_1}\left[P'_N \notin \mathcal{R}(\tilde{h})\right] \geq \liminf_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_1}\left[P'_N \notin \mathcal{R}(\tilde{h}^\delta)\right]. \tag{14}$$

Proof. We start by proving that our family of formulations is feasible and satisfies inequality (12). Note that

$$\mathcal{R}(\tilde{h}^\delta) = \{P' \in \mathcal{P}(\Xi') : I^\delta(P', P_0) > r\}.$$

We may assume without loss of generality that we have $\mathcal{R}(\tilde{h}^\delta) \neq \emptyset$ for otherwise the error probability $\mathrm{Prob}_{P_0}[P'_N \in \mathcal{R}(\tilde{h}^\delta)] = 0$ for $N \geq 1$ clearly satisfies inequality (12). Next we show that $P' \in \mathrm{cl}\,\mathcal{R}(\tilde{h}^\delta) \implies I(P', P_0) > r$. For the sake of contradiction assume that we have found $P' \in \mathrm{cl}\,\mathcal{R}(\tilde{h}^\delta)$ for which $I(P', P_0) \leq r$. There must exist now a sequence $P'_k \in \mathcal{R}(\tilde{h}^\delta)$ which converges to P' and hence $D_{M'}(P'_k, P')$ tends to zero. However, from the definition of the smooth rate function I^δ we have that in fact for all $Q' \in \mathcal{P}(\Xi')$ such that $D_{M'}(Q', P'_k) \leq \delta$ we have that $I(Q', P) > r$. Take now k large enough such that $D_{M'}(P'_k, P') = D_{M'}(P', P'_k) \leq \delta$ then we must have $I(P', P) > r$; a contradiction. The above reasoning implies using the large deviation inequality (6b) that

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_0}\left[P'_N \in \mathcal{R}(\tilde{h}^\delta)\right] \leq -\inf_{P' \in \mathrm{cl}\,\mathcal{R}(\tilde{h}^\delta)} I(P', P_0) \leq -r$$

establishing inequality (12).

We will now prove that for any hypothesis test $\tilde{h}$ which satisfies inequality (13) for some $\epsilon > 0$ we have that there must exist a $\delta' > 0$ so that

$$\mathcal{R}(\tilde{h}) \subseteq \mathcal{R}(\tilde{h}^{\delta'}) \tag{15}$$

from which inequality (14) follows immediately as we note that indeed $\mathcal{R}(\tilde{h}^{\delta'}) \subseteq \mathcal{R}(\tilde{h}^\delta)$ for all $0 < \delta \leq \delta'$.

It remains to prove inequality (15). Assume for the sake of contradiction that $Q''(\delta) \in \mathcal{R}(\tilde{h})$ and $Q''(\delta) \notin \mathcal{R}(\tilde{h}^\delta)$ for all $\delta > 0$. By definition of the smooth rate function I^δ stated in Equation (11) and the hypothesis test $\tilde{h}^\delta$ we have that we can find an auxiliary sequence $Q^\star(\delta) \in \mathcal{P}(\Xi')$ so that $D_{M'}(Q^\star(\delta), Q''(\delta)) \leq \delta$ and $I(Q^\star(\delta), P_0) \leq r$. As the rate function I is good there exist a sequence δ_k to that $Q^\star(\delta_k)$ converges to some $Q^\star \in \mathrm{cl}\,\mathcal{R}(\tilde{h}) \subseteq \mathcal{R}^\epsilon(\tilde{h})$ with $I(Q^\star, P_0) \leq r$. Define now the continuous function $Q' : [0,1] \to \mathcal{P}(\Xi')$, $\lambda \mapsto \lambda \cdot O \star P_0 + (1-\lambda) \cdot Q^\star$ and recall that $I(O \star P_0, P_0) = 0$. From convexity of the rate function I and the fact that $\mathcal{R}^\epsilon(\tilde{h})$ is an open set containing $\mathcal{R}(\tilde{h})$, there must exist $\lambda' \in (0,1]$ sufficiently small so that with $Q' = Q'(\lambda')$ we have using the secant inequality $I(Q', P_0) \leq \lambda I(O \star P_0, P_0) + (1-\lambda')I(Q^\star, P_0) \leq (1-\lambda')r = r' < r$ and $Q' \in \mathcal{R}^\epsilon(\tilde{h})$. From the large deviation inequality (6a) we have

$$-r' \leq -I(Q', P_0) < -\inf_{P' \in \mathcal{R}^\epsilon(\tilde{h})} I(P', P_0) \leq \liminf_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_{P_0}\left[P'_N \in \mathcal{R}^\epsilon(\tilde{h})\right]$$

directly contradicting inequality (13) as we have established that $r' < r$. □

Inequality (12) guarantees that our proposed family of hypothesis tests does not make type I errors all that often. That is, the probability of making an type I error decays to zero exponentially fast at rate at least $-r$. Inequalities (13) and (14) state that our proposed family almost dominates any other hypothesis test. That is, any other hypothesis test which enjoys a type I error guarantee which is stronger by even the smallest amount $\epsilon > 0$ is dominated by all members of our optimal family for the smoothing parameter $0 < \delta \leq \delta'$ and $0 < \delta'$ sufficiently small. Furthermore, remark that our family of tests does not depend on the alternative hypothesis P_1 in any way. It is hence a universally optimal family in that among all feasible tests it suffers an essentially minimal type II error probability rate whatever the alternative hypothesis P_1 may be.

Remark 5.3. In view of the previous discussion it would perhaps be tempting to consider the hypothesis test $\tilde{h}^0$ such that

$$\tilde{h}^0(P'_N) = \begin{cases} P_0 & \text{if } I(P'_N, P_0) \leq r, \\ P_1 & \text{otherwise.} \end{cases}$$

However, remark that when the base measure m' defined in Assumption 3.1 fails to be atomic, then $I(P'_N, P_0) = +\infty$ for the (atomic) empirical probability distribution P'_N of our noisy data. Consequently, the null hypothesis is never accepted by the test $\tilde{h}^0$ which is clearly undesirable as we have under such circumstances

$$\mathrm{Prob}_{P_0}\left[P'_N \in \mathcal{R}(\tilde{h}^0)\right] = 1 \quad \forall N \geq 1.$$

Hence, some degree of smoothing of the rate function as done in Equation (11) seems unavoidable.

5.2 Optimal Data-Driven Decisions

Consider a prescriptive problem in which we attempt to learn the solution to the stochastic optimization problem stated in Equation (2) from the noisy observational data defined in Equation (5). Let us denote with P_N^{ml} the maximum likelihood estimate proposed by [21] for the unobserved probability distribution P. A straightforward extension of the sample average formulation in Equation (3) to this noisy data would be to consider

$$z(P_N^{\mathrm{ml}}) \in \arg_\epsilon \inf_{z\in Z} \mathbf{E}_{P_N^{\mathrm{ml}}}\left[\ell(z,\xi)\right]. \tag{16}$$

Many other formulations based on different distributional estimates are evidently possible as well. The kernel deconvolution estimate P^{kd} proposed in [25] states one such alternative and yields yet another data-driven formulation

$$z(P_N^{\mathrm{kd}}) \in \arg_\epsilon \inf_{z\in Z} \mathbf{E}_{P_N^{\mathrm{kd}}}\left[\ell(z,\xi)\right]. \tag{17}$$

This naturally leads us to question if between these two data-driven formulations one ought to be preferred over the other from a statistical point of view? To answer this question more broadly we must of course first define precisely what constitutes a data-driven formulation and secondly agree on how its statistical performance should be quantified. We follow the framework presented in [27] and define a data-driven formulation as consisting of a predictor and prescriptor.

Definition 5.4 (Data-driven predictors and prescriptors)**.** A measurable function $\tilde{c} : Z \times \mathcal{P}(\Xi') \to \mathrm{R}$ is called a data-driven predictor. A measurable function $\tilde{z} : \mathcal{P}(\Xi') \to Z$ is called a data-driven prescriptor if there exists a data-driven predictor $\tilde{c}$ that induces $\tilde{x}$ in the sense that $\tilde{z}(P') \in \arg_\epsilon \inf_{z\in Z} \tilde{c}(z, P')$ for all $P' \in \mathcal{P}(\Xi')$. That is, we have $\tilde{c}(\tilde{z}(P'), P') - \epsilon < \tilde{v}(P') := \inf_{z\in Z} \tilde{c}(z, P')$ where we denote the function $\tilde{v} : \mathcal{P}(\Xi') \to \mathrm{R}$ as the optimal value function of the formulation.

The maximum likelihood formulation (16) and the kernel deconvolution formulation (17) employ the cost predictors $\mathbf{E}_{P_N^{\mathrm{ml}}}\left[\ell(z,\xi)\right]$ and $\mathbf{E}_{P_N^{\mathrm{kd}}}\left[\ell(z,\xi)\right]$ to prescribe their decisions $z(P_N^{\mathrm{ml}})$ and $z(P_N^{\mathrm{kd}})$, respectively. However, both the maximum likelihood and kernel deconvolution formulation are based on a point estimate for the unobserved probability distribution P and consequently can be expected to suffers similar shortcomings as the sample average formulation. That is, the cost budgeted for its prescribed decision is likely to disappoint out of sample. Here we say a formulation based on a predictor prescriptor pair $(\tilde{c}, \tilde{z})$ disappoints if the event

$$P'_N \in \mathcal{D}(\tilde{c}, \tilde{z}; P) := \{P' \in \mathcal{P}(\Xi') \ : \ c(\tilde{z}(P'), P) > \tilde{c}(\tilde{z}(P'), P')\}$$

occurs with $c(z, P) = \mathbf{E}_P\left[\ell(z,\xi)\right]$. Such disappointment events in which the actual cost of our decision $c(\tilde{z}(P'_N), P)$ breaks the predicted cost or budget $\tilde{c}(\tilde{z}(P'_N), P'_N)$ may result in severe financial consequences and should be avoided by the decision-maker. Consequently, we would prefer formulations which keep the disappointment rates

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_P[P_N \in \mathcal{D}(\tilde{c}, \tilde{z}; P)] \tag{18}$$

as small as possible for all $P \in \mathcal{P}$. Evidently, this can be achieved trivially by simply inflating the cost budgeted for each decision by some large nonnegative amount $\rho > 0$. Indeed, the disappointment probabilities

$$\mathrm{Prob}_P[P'_N \in \mathcal{D}(\tilde{c} + \rho, \tilde{z}; P)] = \mathrm{Prob}_P[c(\tilde{z}(P'_N), P) > \tilde{c}(\tilde{z}(P'_N), P'_N) + \rho] \tag{19}$$

can be made arbitrarily small by choosing a large enough additive inflation constant ρ. However, data-driven formulations with overly conservative cost estimates are clearly undesirable as the ultimate budgeted cost for their optimal decision would compare unfavorable to that of a competitor using more aggressive pricing and should hence also be avoided. We will only denote here formulations feasible if their out-of-sample disappointment probability decays sufficiently fast, i.e., (18) $\leq -r$. Naturally, we would prefer formulations which promise minimal long term costs $\tilde{c}(\tilde{z}(O \star P), O \star P)$ for all $P \in \mathcal{P}$ as indeed we have that the observed random cost $\tilde{c}(\tilde{z}(P'_N), P'_N)$ converges almost surely to $\tilde{c}(\tilde{z}(O \star P), O \star P)$ as the empirical distribution P'_N converges almost surely to $O \star P$ following [9, Theorem 11.4.1] for every $P \in \mathcal{P}$.

We consider therefore the family of robust formulations utilizing predictor prescriptor pairs

$$\tilde{c}^\delta(z, P'_N) := \sup\left\{\mathbf{E}_P\left[\ell(z,\xi)\right] \;:\; P \in \mathcal{P},\; I^\delta(P'_N, P) \leq r\right\}, \quad \tilde{z}^\delta(P'_N) \in \arg_\epsilon \inf_{z \in Z} \tilde{c}^\delta(z, P'_N) \tag{20}$$

based on our smooth large deviation rate function defined in Equation (11). We will show using a large deviation argument that this family dominates the very rich class of formulations based on regular predictors and prescriptors.

Definition 5.5 (Regular predictors and prescriptors)**.** A data-driven predictor $\tilde{c} : Z \times \mathcal{P}(\Xi') \to \mathrm{R}$ is called regular if it is continuous on $Z \times \mathcal{P}(\Xi')$. A data-driven prescriptor $\tilde{z} : \mathcal{P}(\Xi') \to Z$ is called a regular if it is continuous and there exists a regular data-driven predictor $\tilde{c}$ that induces $\tilde{x}$ in the sense that $\tilde{x}(P') \in \arg_\epsilon \inf_{z \in Z} \tilde{c}(z, P')$ for all $P' \in \mathcal{P}(\Xi')$.

Remark that the class of all predictor prescriptor pairs is very rich as Definition 5.5 imposes only mild structural restrictions. The Berge maximum theorem [3, p. 116] indeed implies that the optimal value function $\tilde{v}(P') = \min_{z \in Z} \tilde{c}(z, P')$ of any regular formulation is a continuous function on $\mathcal{P}'(\Xi)$ already when the constraint set Z is merely compact. The correspondence $P' \mapsto \{z \in Z \;:\; \tilde{c}(z, P') < \tilde{v}(P') + \epsilon\}$ of ϵ-suboptimal solutions in a regular formulation is consequently lower semicontinuous [2, Corollary 4.2.4.1] for any $\epsilon > 0$. Hence, for formulations employing a convex predictor $\tilde{c}$ and $\mathcal{P}(\Xi')$ a compact set, an associated regular predictor can always be found [1, Theorem 9.1.]. Should a regular formulation admit unique optimal decisions, such decisions will constitute a regular prescriptor as well following [3, p. 117]. The need to focus on this restricted but nevertheless quite rich class of regular formulations is necessary due to the rough nature of the employed large deviation argument.

Assumption 5.6. The cost function $c : Z \times \mathcal{P} \to \mathrm{R},\; (z, P) \mapsto \mathbf{E}_P\left[\ell(z,\xi)\right]$ is continuous.

Theorem 5.7. *Let Assumption 5.6 hold. Then, the family of predictor prescriptor pairs* $(\tilde{c}^\delta, \tilde{z}^\delta)$ *is feasible for any* $\delta > 0$*, i.e.,*

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathrm{Prob}_P\left[P'_N \in \mathcal{D}(\tilde{c}^\delta, \tilde{z}^\delta; P)\right] \leq -r \quad \forall P \in \mathcal{P}. \tag{21}$$

Furthermore, for any regular predictor prescriptor pair $(\tilde{c}, \tilde{z})$ *that satisfies*

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathrm{Prob}_P\left[P'_N \in \mathcal{D}(\tilde{c}, \tilde{z}; P)\right] \leq -r \qquad \forall P \in \mathcal{P}, \tag{22}$$

we have that for all $\epsilon > 0$ *there exists* $0 < \delta'$ *so that for all* $0 < \delta \leq \delta'$ *we have*

$$\tilde{c}(\tilde{z}(O \star P), O \star P) + 2\epsilon \geq \tilde{c}^\delta(\tilde{z}^\delta(O \star P), O \star P) \quad \forall P \in \mathcal{P}. \tag{23}$$

Proof. We start by proving that our family of formulations is feasible and satisfies inequality (21). To this end define the sets

$$\mathcal{D}^\delta(P) = \{P' \in \mathcal{P}(\Xi') : \sup_{z \in Z}\; c(z, P) - \tilde{c}^\delta(z, P') > 0\} \quad \text{and} \quad \mathcal{B}^\delta(P) = \{P' \in \mathcal{P}(\Xi') : I^\delta(P', P) > r\}.$$

We may assume without loss of generality that we have $\mathcal{D}^\delta(P) \neq \emptyset$ for otherwise the out-of-sample disappointment $\mathrm{Prob}_P[P'_N \in \mathcal{D}(\tilde{c}^\delta_\delta, \tilde{z}^\delta_\delta; P) \subseteq \mathcal{D}^\delta(P)]$ clearly vanishes for all $N \geq 1$ and inequality (21) holds trivially. We will now show that $\mathcal{D}^\delta(P) \subseteq \mathcal{B}^\delta(P)$. For the sake of contradiction, choose any $P' \in \mathcal{D}^\delta(P)$, and assume that $I^\delta(P', P) \leq r$. Thus, we have for some $z \in Z$ that

$$c(z, P) > \tilde{c}^\delta(z, P') = \sup\left\{c(z, Q) \;:\; Q \in \mathcal{P},\; I^\delta(P', Q) \leq r\right\} \geq c(z, P);$$

a contradiction. As $P' \in \mathcal{D}^\delta(P)$ was chosen arbitrarily, we have thus shown that $\mathcal{D}^\delta(P) \subseteq \mathcal{B}^\delta(P)$ and hence also $\mathrm{cl}\,\mathcal{D}^\delta(P) \subseteq \mathrm{cl}\,\mathcal{B}^\delta(P)$. Next we show that $P' \in \mathrm{cl}\,\mathcal{B}^\delta(P) \implies I(P', P) > r$. For the sake of contradiction assume that we have found $P' \in \mathrm{cl}\,\mathcal{B}^\delta(P)$ for which $I(P', P) \leq r$. There must exist a sequence $P'_k \in \mathcal{B}^\delta(P)$ which converge to P' and hence $D_{M'}(P'_k, P')$ tends to zero. However, from the definition of $\mathcal{B}^\delta(P)$ we have that in fact for all $Q' \in \mathcal{P}(\Xi')$ such that $D_{M'}(Q', P'_k) \leq \delta$ we have that $I(Q', P) > r$. Take now k large enough such that $D_{M'}(P'_k, P') = D_{M'}(P', P'_k) \leq \delta$ then we must have $I(P', P) > r$; a contradiction. The above reasoning implies that

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_P \left[c(\tilde{z}^\delta(P'_N), P) > \tilde{c}^\delta(\tilde{z}^\delta(P'_N), P'_N)\right] \\
&\qquad\qquad \leq \limsup_{N\to\infty} \frac{1}{N} \log \mathrm{Prob}_P \left[\sup_{z\in Z} c(z, P) - \tilde{c}^\delta(z, P'_N) > 0\right] \leq -\inf_{P'\in \mathrm{cl}\,\mathcal{B}^\delta(P)} I(P', P) \leq -r
\end{aligned}$$

establishing inequality (21) as $P \in \mathcal{P}$ was arbitrary.

We will now prove that for any regular predictor prescriptor pair $(\tilde{c}, \tilde{z})$ which satisfies inequality (22) we have

$$\lim_{\delta\to 0} \inf_{P'\in\mathcal{P}'} \tilde{c}(\tilde{z}(P'), P') - \tilde{c}^\delta(\tilde{z}(P'), P') \geq 0. \tag{24}$$

From inequality (24) we can take $\delta' > 0$ sufficiently small so that $\tilde{c}(\tilde{z}(P'), P') \geq \tilde{c}^{\delta'}(\tilde{z}(P'), P') - \epsilon \geq \tilde{c}^{\delta'}(\tilde{z}^{\delta'}(P'), P') - 2\epsilon$ uniformly for all $P' \in \mathcal{P}' = \{O \star P \;:\; P \in \mathcal{P}\}$. Remark that we have clearly $\tilde{c}^\delta(z, P') \leq \tilde{c}^{\delta'}(z, P')$ for all $z \in Z$, $P' \in \mathcal{P}(\Xi')$ and $0 < \delta \leq \delta'$ from which inequality (23) follows immediately.

We will now establish inequality (24) by showing that

$$\lim_{\delta\to 0} \inf_{P'\in\mathcal{P}'} \tilde{c}(\tilde{z}(P'), P') - \tilde{c}^\delta(\tilde{z}(P'), P') \geq -3\rho \tag{25}$$

for any $\rho > 0$. Assume for the sake of contradiction that $\lim_{\delta\to 0} \inf_{P'\in\mathcal{P}'} \tilde{c}(\tilde{z}(P'), P') - \tilde{c}^\delta(\tilde{z}(P'), P') < -3\rho$. There must hence exist a $\delta' > 0$ such that for all $\delta \leq \delta'$ we have $\inf_{P'\in\mathcal{P}'} \tilde{c}(\tilde{z}(P'), P') - \tilde{c}^\delta(\tilde{z}(P'), P') < -2\rho$. Consequently, there exists a distribution $Q''(\delta) \in \mathcal{P}'$ such that

$$\begin{aligned}
&\tilde{c}(\tilde{z}(Q''(\delta)), Q''(\delta)) + 2\rho < c^\delta(\tilde{z}(Q''(\delta)), Q''(\delta)) \quad \forall \delta \leq \delta' \\
&\qquad\qquad < \sup\left\{c(\tilde{z}(Q''(\delta)), P) \;:\; P \in \mathcal{P},\, I^\delta(Q''(\delta), P) \leq r\right\} \quad \forall \delta \leq \delta'.
\end{aligned}$$

Hence, there exists for all $\delta \leq \delta'$ a distribution $P^\star(\delta) \in \mathcal{P}$ such that $I^\delta(Q''(\delta), P^\star(\delta)) \leq r$ and $\tilde{c}(\tilde{z}(Q''(\delta)), Q''(\delta)) + 2\rho < c(\tilde{z}(Q''(\delta)), P^\star(\delta))$. From the definition of the smooth rate function I^δ stated in Equation (11) this implies that there exists an auxiliary sequence $Q'''(\delta) \in \mathcal{P}(\Xi')$ such that $I(Q'''(\delta), P^\star(\delta)) \leq r$ with $D_{M'}(Q'''(\delta), Q''(\delta)) \leq \delta$ for all $\delta \leq \delta'$. Remark that

$$\begin{aligned}
\lim_{\delta\to 0} \tilde{c}(\tilde{z}(Q'''(\delta)), Q'''(\delta)) &= \lim_{\delta\to 0} \tilde{c}(\tilde{z}(Q''(\delta)), Q''(\delta)) \\
\lim_{\delta\to 0} c(\tilde{z}(Q'''(\delta)), P^\star(\delta)) &= \lim_{\delta\to 0} c(\tilde{z}(Q''(\delta)), P^\star(\delta))
\end{aligned}$$

as the functions $c : Z \times \mathcal{P} \to \mathrm{R}$, $\tilde{c} : Z \times \mathcal{P}(\Xi') \to \mathrm{R}$ and $\tilde{z} : \mathcal{P}(\Xi') \to Z$ are continuous and $D_{M'}(Q'''(\delta), Q''(\delta)) \leq \delta$ implies $\lim_{\delta\to 0} Q'''(\delta) = \lim_{\delta\to 0} Q''(\delta)$. Consequently, there exists a $\delta^\star \in (0, \delta']$ so that with $P^\star = P^\star(\delta^\star) \in \mathcal{P}$

and $Q^\star = Q'''(\delta^\star) \in \mathcal{P}(\Xi')$ we have both

$$|\tilde{c}(\tilde{z}(Q''(\delta^\star)), Q''(\delta^\star)) - \tilde{c}(\tilde{z}(Q^\star), Q^\star)| < \rho \text{ and } |c(\tilde{z}(Q''(\delta^\star)), P^\star) - c(\tilde{z}(Q^\star), P^\star)| < \rho.$$

Hence, we have both $\tilde{c}(\tilde{z}(Q^\star), Q^\star) < c(\tilde{z}(Q^\star), P^\star)$ and $I(Q^\star, P^\star) \leq r$. Define the continuous function $Q' : [0,1] \to \mathcal{P}(\Xi'),\ \lambda \mapsto Q \star P^\star \cdot \lambda + Q^\star \cdot (1-\lambda)$ and recall that $I(O \star P^\star, P^\star) = 0$. As we have that the functions $c : Z \times \mathcal{P} \to \mathrm{R}$, $\tilde{c} : Z \times \mathcal{P}(\Xi') \to \mathrm{R}$ and $\tilde{z} : \mathcal{P}(\Xi') \to Z$ are continuous there hence exists $Q' = Q'(\lambda')$ for $\lambda' \in (0,1]$ sufficiently small so that $I(Q', P^\star) \leq \lambda I(O \star P^\star, P^\star) + (1-\lambda) I(Q^\star, P^\star) \leq (1-\lambda') \cdot r = r' < r$ using the convexity of the rate function and $\tilde{c}(\tilde{z}(Q'), Q') < c(\tilde{z}(Q'), P^\star)$. Consequently, we have that $Q' \in \mathcal{D}(\tilde{c}, \tilde{x}; P^\star) = \operatorname{int} \mathcal{D}(\tilde{c}, \tilde{x}; P^\star)$ as again we remark for a final time that the functions $c : Z \times \mathcal{P} \to \mathrm{R}$, $\tilde{c} : Z \times \mathcal{P}(\Xi') \to \mathrm{R}$ and $\tilde{z} : \mathcal{P}(\Xi') \to Z$ are continuous. From the large deviation inequality (6a) it follows now that

$$-r' \leq -I(Q', P^\star) < -\inf_{P' \in \operatorname{int} \mathcal{D}(\tilde{c}, \tilde{x}; P^\star)} I(P', P^\star) \leq \liminf_{N \to \infty} \frac{1}{N} \log \operatorname{Prob}_P \left[P'_N \in \mathcal{D}(\tilde{c}, \tilde{z}; P^\star) \right]$$

which is in direct contradiction with the feasibility inequality (22) as $r' < r$ which establishes our inequality (25) as $\rho > 0$ was arbitrary. □

Inequality (21) indicates that any formulation in our family is feasible. Furthermore, inequalities (22) and (23) argue that any regular formulation is dominated by members of our optimal family for all parameters $0 < \delta \leq \delta'$ with $0 < \delta'$ sufficiently small. The previous theorem hence indicates that our family dominates any regular formulation in terms of balancing the desire for small out-of-sample disappointment as well as minimal budget costs under Assumption 5.6. Finally, we remark that Assumption 5.6 is rather mild and is already satisfied when the loss function $\ell : Z \times \Xi \to \mathrm{R}$ is merely bounded and uniformly continuous.

Remark 5.8. In view of the previous discussion it is again tempting to consider the data-driven formulation with predictor and prescriptor

$$\tilde{c}^0(z, P'_N) := \sup \left\{ \mathbf{E}_P \left[\ell(z, X) \right] \ : \ P \in \mathcal{P},\ I(P'_N, P) \leq r \right\}, \quad \tilde{z}^0(P'_N) \in \arg\min_{z \in Z} \tilde{c}^0(z, P'_N) \tag{26}$$

based directly on our rate function I rather than its smooth counterpart I^δ. Van Parys et al. [27] proves indeed that when given access to the noiseless data in Equation (1) with empirical distribution P_N the appropriate rate function is precisely the entropic distance $D_{\mathrm{KL}}(P_N, P)$ and formulation (4) enjoys optimal statistical properties very similar to those found in Theorem 5.7. However, recall again that for noisy data when the base measure m' defined in Assumption 3.1 fails to be atomic, the ambiguity set $\{P \in \mathcal{P} \ : \ I(P'_N, P) < \infty\} = \emptyset$ is trivial and consequently the associated data-driven formulation is here infeasible. Hence, considering the smooth rate function I^δ instead of the rate function I directly seems unavoidable when faced with noisy observational data.

References

[1] J.-P. Aubin and H. Frankowska. **Set-Valued Analysis**. Springer Science & Business Media, 2009.

[2] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer. **Non-Linear Parametric Optimization**. Springer, 1982.

[3] C. Berge. **Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity**. Courier Corporation, 1997.

[4] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. **Mathematical Programming**, 167(2):235–292, 2018.

[5] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. **Advances in neural information processing systems**, 26:2292–2300, 2013.

[6] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. **Operations research**, 58(3):595–612, 2010.

[7] A. Dembo and O. Zeitouni. **Large Deviations Techniques and Applications**, volume 38. Springer Science & Business Media, 2009.

[8] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. **Mathematics of Operations Research**, 2021.

[9] R. M. Dudley. **Real Analysis and Probability**. CRC Press, 2018.

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

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

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

[13] P. Groeneboom and J. A. Wellner. **Information bounds and nonparametric maximum likelihood estimation**, volume 19. Springer Science & Business Media, 1992.

[14] Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. **Available at Optimization Online**, 2013.

[15] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. 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.

[16] R. O. Michaud. The Markowitz optimization enigma: Is “optimized” optimal? **Financial Analysts Journal**, 45(1):31–42, 1989.

[17] T. Pennanen. **Introduction to Convex Optimization**. 2019. URL `https://nms.kcl.ac.uk/teemu.pennanen/co-new.pdf`.

[18] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. **Foundations and Trends® in Machine Learning**, 11(5-6):355–607, 2019.

[19] Y. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. **Theory of Probability & Its Applications**, 1(2):157–214, 1956.

[20] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. **arXiv preprint arXiv:1908.05659**, 2019.

[21] P. Rigollet and J. Weed. Entropic optimal transport is maximum-likelihood deconvolution. **Comptes Rendus Mathematique**, 356(11-12):1228–1235, 2018.

[22] F. Santambrogio. **Optimal Transport for Applied Mathematicians**. Birkäuser, NY, 2015.

[23] A. Shapiro, D. Dentcheva, and A. Ruszczyńsk. **Lectures on Stochastic Programming: Modeling and Theory**. SIAM, 2014.

[24] B. W. Silverman. **Density Estimation for Statistics and Data Analysis**, volume 26. CRC press, 1986.

[25] L. A. Stefanski and R. J. Carroll. Deconvolving kernel density estimators. **Statistics**, 21(2):169–184, 1990.

[26] B. P. Van Parys, P. J. Goulart, and D. Kuhn. Generalized Gauss inequalities via semidefinite programming. **Mathematical Programming**, 156(1-2):271–302, 2016.

[27] B. P. Van Parys, P. M. Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. **Management Science**, 2020. doi: https://doi.org/10.1287/mnsc.2020.3678.

[28] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. **Operations Research**, 62(6):1358–1376, 2014.