Operations Research Letters 54 (2024) 107089

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Efficient data-driven optimization with noisy data

Bart P.G. Van Parys

MIT Sloan School of Management, 100 Main Street, Cambridge, 02142, MA, USA

ARTICLE INFO

Article history:
Received 15 April 2022
Received in revised form 10 July 2023
Accepted 31 January 2024
Available online 6 February 2024

Keywords:
Noisy data
Distributionally robust optimization
Statistical efficiency
Large deviations
Entropic optimal transport

ABSTRACT

The classical Kullback-Leibler distance is known to enjoy desirable statistical properties in the context of decision-making with noiseless data. However, in most practical situations data is subject to a certain amount of measurement noise. We hence study here data-driven prescription problems in which the data is corrupted by a known noise source. We derive efficient data-driven formulations in this noisy regime and indicate that they enjoy an entropic optimal transport interpretation.

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. We are interested in finding an $0 < \epsilon$-suboptimal solution to the stochastic optimization problem

$$z(P) \in \arg_\epsilon \inf_{z \in Z} \{\mathbf{E}_P[\ell(z,\xi)] = \int_\Xi \ell(z,\xi)\,\mathrm{d}P(\xi)\}. \tag{1}$$

In practice, however, the distribution P which describes the uncertain parameter ξ is not known. The stochastic optimization community has in recent years focused on making decisions directly based on a finite collection of independent data observations

$$\xi_i \sim P \quad \forall i \in [1,\ldots,N], \tag{2}$$

instead. In this paper we will denote this observational model, considered by the overwhelming majority of the literature on data-driven decisions, described in Equation (2) as *noiseless*. That is, the decision maker has access to uncorrupted independent samples from the probability measure of interest P. However, it is often the case in practice that the data itself is not observed directly but rather through a measurement device with known limitations. Such noisy data is better known as censored data in statistics [16]. An observational model in which the noisy data

$$\xi_i' = \xi_i + n_i \quad \forall i \in [1,\ldots,N], \tag{3}$$

is observed instead where the noise terms n_i are independent and share a known distribution has been widely studied in the system control and identification literature. We will come back to this observational model and further generalizations in Section 2.1. In the remainder of this section we first briefly discuss various results of interest in the noiseless observational model.

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 [19] may be considered 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

$$z(P_N) \in \arg_\epsilon \inf_{z \in Z} \mathbf{E}_{P_N}[\ell(z,\xi)] \tag{4}$$

which substitutes P with a mere point estimate P_N tends to disappoint out of sample. That is, the actual cost ($\mathbf{E}_P[\ell(z(P_N),\xi)]$) observed out of sample exceeds the predicted cost ($\mathbf{E}_{P_N}[\ell(z(P_N),\xi)]$) of the data-driven decision $z(P_N)$. This adversarial phenomenon is well known colloquially as the "Optimizer's Curse" [25] or overfitting. 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.

E-mail address: vanparys@mit.edu.
URL: https://www.vanparys.xyz.

https://doi.org/10.1016/j.orl.2024.107089
0167-6377/

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, pioneered by [29], focuses consequently on data-driven formulations of the form

$$z_{\mathcal{A}}(P_N) \in \arg_{\epsilon} \inf_{z\in Z} \sup \{\mathbf{E}_P[\ell(z,\xi)] : P \in \mathcal{A}_N(P_N)\} \tag{5}$$

which can be thought of as robust counterparts to the nominal sample average formulation stated in Equation (4) and where by convention consider the supremum to take the value $-\infty$ in case its feasible set $\mathcal{A}_N(P_N)$ is empty. The recent uptick in popularity of such 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 counterpart. Earlier work [5,11,17,32,34] focused on ambiguity sets consisting of probability measures sharing certain given moments. More recent approaches [3,6,18,20,21] however consider ambiguity sets $\mathcal{A}_N(P_N) = \{P \in \mathcal{P} : D(P_N, P) \leq r_N\}$ which are based on some statistical distance $D : \mathcal{P}(\Xi) \times \mathcal{P}(\Xi) \to \mathrm{R} \cup \{+\infty\}$. Such ambiguity sets are interpretable as the set of probability measures sufficiently close to the empirical probability measure P_N. Two qualitatively different distances have recently positioned themselves as the front runners for data-driven decision-making and are now briefly discussed.

The optimal transport distance between a measure μ on Ξ and a measure ν on another set Ξ' can be defined as

$$W_0(\mu,\nu) := \inf_{T\in\mathcal{T}(\mu,\nu)} \int_{\Xi\times\Xi'} d(\xi,\xi')\,\mathrm{d}T(\xi,\xi') \tag{6}$$

for a given the cost function $d : \Xi \times \Xi' \mapsto \mathrm{R} \cup \{+\infty\}$. Here $\mathcal{T}(\mu,\nu) := \left\{T \in \mathcal{P}(\Xi\times\Xi') : \Pi_{\Xi'}T = \mu,\ \Pi_{\Xi}T = \nu\right\}$ is the so called transport polytope and consists of all transport measures with given marginal measures μ and ν. Assume for a moment that $\Xi = \Xi' = \mathrm{R}^m$ and $d(\xi,\xi') = \|\xi - \xi'\|_2$. Then, the optimal transport distance coincides with the classical Wasserstein distance between probability measures [28]. Optimal transport distances have received a lot of attention both in the context of prescriptive analytics [7,8,15,30,35] as well as in the machine learning community at large [23,27]. In the context of prescriptive analytics such distances have become very popular after the seminal work [15] pointed out that the resulting robust formulation $z_{W,r}(P_N) \in \min_{z\in Z}\{c_{W,r}(z,P_N) := \sup\{\mathbf{E}_P[\ell(z,\xi)] : W_0(P_N,P) \leq r\}\}$ need not be intractable and enjoys strong out-of-sample guarantees. For $r < 1$ and $2 \leq \dim(\Xi) < \infty$, we have [23] for any P with $A = \mathbf{E}_P[\exp(\|\xi\|_2^a)] < \infty$ for some $a > 1$ that with $r' = (r/C)^{1/\dim(\Xi)}$ we have

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \Pr\left[\mathbf{E}_P\left[\ell(z_{W,r'}(P_N),\xi)\right] > c_{W,r'}(z_{W,r'}(P_N),P_N)\right] \leq r \tag{7}$$

where C is a known positive constant depending only on a, A and $\dim(\Xi)$. In other words, the probability of being disappointed (the actual cost $\mathbf{E}_P\left[\ell(z_{W,r'}(P_N),\xi)\right]$ of the decision $z_{W,r'}(P_N)$ exceeds the predicted worst-case cost $c_{W,r'}(z_{W,r'}(P_N),P_N)$) decays exponentially fast in the number of samples.

Optimal transport distances are however computationally quite challenging as even determining the optimal transport distance between two given measures requires the solution of the associated linear optimization problem in Equation (6). In fact, optimal transport distances are often computed numerically, cf., [9], by considering their embedding in the larger family of entropic optimal transport distances $W_\epsilon(\mu,\nu)$ defined as

$$\inf_{T\in\mathcal{T}(\mu,\nu)} \int_{\Xi\times\Xi'} d(\xi,\xi')\mathrm{d}T(\xi,\xi') + \epsilon \mathrm{KL}(T, \Pi_{\Xi}T \otimes \Pi_{\Xi'}T) \tag{8}$$

as the limit $\epsilon \downarrow 0$ where we denote the entropic divergence between two finite measures μ and ν on the same space as

$$\mathrm{KL}(\mu,\nu) = c\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu - \int \mathrm{d}\mu + \int \mathrm{d}\nu$$

where the random variable $\mathrm{d}\mu/\mathrm{d}\nu$ denotes the Radon-Nikodym derivative between μ and ν. By convention if $\mu \not\ll \nu$ then the entropic distance is taken as $+\infty$. The previously discussed entropic divergence is a particular member of the class of convex f-divergences which like the Wasserstein distance are well known [21] to yield tractable robust formulations. Moreover, when the loss function $\ell(z,\xi)$ is continuous on the compact set $Z \times \Xi$ the associated robust prescriptive formulation (5) can be specialized to $z_{\mathrm{KL},r}(P_N) \in \min_{z\in Z}\{c_{\mathrm{KL},r}(z,P_N) := \max\{\mathbf{E}_P[\ell(z,\xi)] : \mathrm{KL}(P_N,P) \leq r\}\}$ and enjoys strong statistical out-of-sample guarantees. One can prove [33, Theorem 11] for all $P \in \mathcal{P}(\Xi)$ that indeed

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \Pr\left[\mathbf{E}_P\left[\ell(z_{\mathrm{KL},r}(P_N),\xi)\right] > c_{\mathrm{KL},r}(z_{\mathrm{KL},r}(P_N),P_N) + \epsilon\right] \leq -r \tag{9}$$

for any $\epsilon > 0$. In other words, the probability of being disappointed (the actual cost $\mathbf{E}_P\left[\ell(z_{\mathrm{KL},r}(P_N),\xi)\right]$ of the decision $z_{\mathrm{KL},r}(P_N)$ exceeds the predicted worst-case cost $\sup\left\{\mathbf{E}_Q\left[\ell(z_{\mathrm{KL},r}(P_N),\xi)\right] : \mathrm{KL}(P_N,Q) \leq r\right\}$ by any small amount $\epsilon > 0$) decays exponentially fast in the number of samples with rate precisely the size r of the considered ambiguity set.

Equations (7) and (9) reflect the fact that when properly calibrated robust entropic and Wasserstein formulations enjoy essentially the same out-of-sample guarantees. The classical sample average formulation can also be made to enjoy similar out-of-sample guarantees by naively inflating its objective in Equation (4) by some bias term $b > 0$. Indeed, taking the bias term b sufficiently large we can guarantee that for all $P \in \mathcal{P}$ we have

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \Pr\left[\mathbf{E}_P[\ell(z_N(P_N),\xi)] > \mathbf{E}_{P_N}[\ell(z_N(P_N),\xi)] + b\right] \leq -r. \tag{10}$$

When several prescription formulations enjoy the same out-of-sample guarantees, we should prefer that formulation which inflates the cost prediction the least. However, any formulation which enjoys for all $P \in \mathcal{P}$ the out-of-sample guarantee

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \Pr\left[\mathbf{E}_P[\ell(z_N(P_N),\xi)] > \tilde{c}_r(\tilde{z}_r(P_N),P_N)\right] \leq -r \tag{11}$$

must be more conservative in its cost predictions compared to an entropic formulation [33, Theorem 11], i.e., $\tilde{c}_r(\tilde{z}_r(P_N),P_N) \geq c_{\mathrm{KL},r}(z_{\mathrm{KL},r}(P_N),P_N)$ indicating that the entropic formulation is universally least conservative or *efficient*. We remark however that the efficiency of the entropic formulation is intimately tied to the noiseless data model. Indeed, the efficiency of the entropic formulation is established [33] by pointing out that the Kullback-Leibler divergence is precisely the rate function characterizing large deviations between the empirical distribution P_N and the noiseless data generating distribution P [12].

1.1. Notation

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. Given any set $S \subset \Xi'$ we denote with $S^\delta = \left\{s' \in \Xi' \,:\, \left\|s - s'\right\| \leq \delta,\ s \in S\right\}$ its δ-inflation. We denote with $\mathcal{M}_+(\Xi)$, $\mathcal{M}_+(\Xi')$ and $\mathcal{M}_+(\Xi \times \Xi')$ the sets of all positive Borel measures on the spaces Ξ, Ξ' and $\Xi \times \Xi'$, respectively. Similarly, we denote with $\mathcal{P}(\Xi)$, $\mathcal{P}(\Xi')$ and $\mathcal{P}(\Xi \times \Xi')$ the sets of all Borel probability measures on the spaces Ξ, Ξ' and $\Xi \times \Xi'$, respectively. Given two measures μ and ν we denote with $\mu \otimes \nu$ as their product measure. 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 [12, Section 6.2].

1.2. Contributions

In this paper we generalize the framework of efficient formulations introduced by [33] to problems with noisy data. We present three contributions.

First, we prove in Theorem 2.9 that the family of robust formulations

$$\tilde{z}^\delta(P'_N) \in \arg_\epsilon \inf_{z \in Z} \sup \left\{\mathbf{E}_Q\left[\ell(z,\xi)\right] \,:\, Q \in \mathcal{P},\ I^\delta(P'_N, Q) \leq r\right\} \tag{12}$$

parameterized in $\delta > 0$ and where $P'_N = \sum_{i=1}^N \delta_{\xi'_i}/N$ denotes the empirical distribution of the noisy data essentially dominates all other formulations mirroring the notion of efficiency enjoyed by the entropic distance in the noiseless regime. Perhaps surprisingly we show that the rate function I in the noisy setting is too irregular and its δ-smoothed counterparts

$$I^\delta(\mu,\nu) := \inf\left\{I(\upsilon,\nu) \,:\, \upsilon \in \mathcal{P}(\Xi'),\ \mathrm{LP}(\upsilon,\mu) \leq \delta\right\} \tag{13}$$

must be considered instead. This implies in particular that in stark contrast to the noiseless setting there is not any longer a single most efficient formulation but rather a family of increasingly more efficient formulations.

Second, we show in Section 2.3 that by reducing the smoothing parameter δ_N and robustness parameter r_N at an appropriate rate, consistent formulations with finite sample guarantees can be obtained under an identifiability condition and a mild assumption on the loss function ℓ.

We finally state tractable reformulations of the proposed novel family of efficient robust prescriptive formulations (12) under certain technical conditions on the loss function ℓ and space Ξ in Section 3. In particular, we exploit a classical representation result [31] which seems to be novel in this context and derive a dual formulation whose size is independent of the event set Ξ'.

2. Decision-making with noisy data

2.1. Noisy data

As stated in Equation (3) the distribution of the noisy data ξ' may be distinct from the distribution of the unobserved noiseless data ξ. We will allow the noisy data to take value in Ξ' which may be different from Ξ. We only assume here that the noisy observations are drawn independently as

$$\xi'_i \sim O_{\xi_i} \quad \forall i \in [1,\ldots,N].$$

That is, each noisy data point ξ'_i is obtained as an independent draw from a distribution $O_{\xi_i} \in \mathcal{P}(\Xi')$ given a noiseless observation ξ_i. We stress here again that we assume that the mapping $O : \Xi \to \mathcal{P}(\Xi')$ which characterizes our observational model is given. In other words, the distributional nature of the noise corrupting the unobserved data points is known. We will refer to O as our observational model as it precisely characterizes how the noisy data is derived from the noiseless data.

To establish the decomposition result in Theorem 2.10 the observational model will be required to satisfy the following technical condition.

Assumption 2.1. The measure O_ξ is absolutely continuous with respect to a base measure m' for all $\xi \in \Xi$, i.e., $O_\xi \ll m'$ for all $\xi \in \Xi$. Furthermore, there exists a measurable function $d : \Xi \times \Xi' \to \mathrm{R}$ so that $\mathrm{d}O_\xi/\mathrm{d}m'(\xi') = \exp(-d(\xi,\xi'))$ for all $\xi' \in \Xi'$.

The relationship between the probability measure P' of the noisy observations and the probability measure P of the unobserved noiseless data can be characterized as the convolution $P' := O \star P$ and is given explicitly as $P'(B) = (O \star P)(B) := \int O_\xi(B)\mathrm{d}P(\xi)$ for all measurable sets $B \in \mathcal{B}(\Xi')$. We will denote with the set $\mathcal{P}' := \left\{O \star P \in \mathcal{P}(\Xi') \,:\, P \in \mathcal{P}\right\}$ the family of potential distributions of our noisy data.

We point out that this noisy data model is quite flexible and captures a wide variety of settings.

Example 2.2 *(Additive noise).* Practical measurements are typically corrupted by some amount of measurement error. We consider here independent additive measurement error e_i from some distribution $E \in \mathcal{P}(\Xi')$ with $\Xi' = \mathrm{R}^{\dim(\Xi)}$ as an example. In this case we observe the noisy data $\xi'_i = \xi_i + e_i$ instead of the data ξ_i itself. This observational model is characterized by the map $O^{AE} : \xi \mapsto E(\xi)$ where $E(\xi)$ denotes the error distribution translated by ξ, i.e., $E(\xi)[B] = \int \mathbb{1}\{\xi + e \in B\}\,\mathrm{d}E(e)$ for every measurable set B in Ξ'.

Example 2.3 *(Gaussian noise).* Perhaps the most classical example of the additive noise model discussed previously is the simple case of independent zero mean Gaussian additive noise z_i with variance $\sigma^2 I$. This observational model is characterized by the map $O^{GN} : \xi \mapsto N(\xi, \sigma^2 I)$ where $N(\mu, V)$ denotes here a normal distribution with mean vector $\mu \in \mathrm{R}^{\dim(\Xi)}$ and variance matrix $V \in \mathrm{R}^{\dim(\Xi)\times\dim(\Xi)}$. Assumption 2.1 holds for O^{GN} with $\Xi' = \mathrm{R}^{\dim(\Xi)}$ and $d(\xi,\xi') = \left\|\xi - \xi'\right\|_2^2/(2\sigma^2)$ and $m' = \mu'/(\sigma\sqrt{(2\pi)^{\dim(\Xi')}})$ with μ' the Lebesgue measure on Ξ'.

Example 2.4 *(Clipping noise).* Most measurement devices have a limited output range, i.e., $\Xi' = [a,b]$ with $a < b$, which is a strict subset of all potential outcomes $\Xi = \mathrm{R}$. In this case we may only observe the censured data $\xi''_i = \max(\min(\xi'_i, b), a)$ instead of data ξ'_i which itself has been corrupted by Gaussian noise as discussed in Example 2.3. This observational model is characterized by the map $O^{CN} : \xi \mapsto N(\xi,\sigma^2)[-\infty,a)\cdot\delta_a + N(\xi,\sigma^2)(b,\infty]\cdot\delta_b + N(\xi,\sigma^2)[a,b]\cdot N_{[a,b]}(\xi,\sigma^2)$ where $N_{[a,b]}(\xi,\sigma^2)$ denotes here a normal distribution truncated to the interval $[a,b]$. Assumption 2.1 holds here with

$$d(\xi,\xi') = \begin{cases} -\log N(\xi,\sigma^2)[-\infty,a] & \text{if } \xi' = a, \\ \left\|\xi - \xi'\right\|_2^2/(2\sigma^2) & \text{if } \xi' \in (a,b), \\ -\log N(\xi,\sigma^2)[b,\infty] & \text{if } \xi' = b \end{cases}$$

for all $\xi \in \Xi$ and $m' = \mu'/(\sigma\sqrt{(2\pi)^{\dim(\Xi')}}) + \delta_a + \delta_b$ with μ' the Lebesgue measure on R and δ_a and δ_b two Dirac measures at locations a and b, respectively.

Example 2.5 *(Quantization noise).* Digital measurements quantize the noisy measurements of Example 2.3 further in the sense that

Ξ' is necessarily only a finite subset of Ξ. Let $\Xi_{\xi'}$ denote the collection of all inputs ξ in Ξ which get quantized into the digital symbol $\xi' \in \Xi'$. This observational model is characterized by the map $O^{QN} : \xi \mapsto \sum_{\xi' \in \Xi'} N(\xi, \sigma^2)[\Xi_{\xi'}] \cdot \delta_{\xi'}$. Assumption 2.1 holds here with

$$d(\xi, \xi') = -\log N(\xi, \sigma^2)[\Xi_{\xi'}]$$

for all $\xi' \in \Xi'$ and $\xi \in \Xi$ with μ' the counting measure on Ξ'.

We will attempt to infer the unknown probability measure P from the noisy data based on its empirical probability measure P'_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. Sanov's theorem [12, Theorem 6.2.10] ensures also here that the sufficient statistic P'_N enjoys a large deviation property. That is, the empirical probability measure P'_N satisfies for any open subset $O \subseteq \mathcal{P}(\Xi')$ the large deviation lower bound

$$-\inf_{\hat{P}' \in O} \mathrm{KL}(\hat{P}', O \star P) \leq \varliminf_{N\to\infty} \frac{1}{N} \log \Pr[P'_N \in O] \tag{14a}$$

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

$$\varlimsup_{N\to\infty} \frac{1}{N} \log \Pr[P'_N \in C] \leq -\inf_{\hat{P}' \in C} \mathrm{KL}(\hat{P}', O \star P), \tag{14b}$$

for the good rate function [12] $I(\hat{P}', P) := \mathrm{KL}(\hat{P}', O \star P)$. We remark that large deviation inequalities generally are quite rough in nature as indeed (14a) and (14b) only pertain to open or closed sets (in the topology of weak convergence), respectively. The rate function is observed to be nonnegative and in fact $I(\hat{P}', P') = 0$ if and only if $\hat{P}' = P'$. For any $\epsilon > 0$, the large deviation inequality (14b) implies that for all $P \in \mathcal{P}$ we have

$$\varlimsup_{N\to\infty} \frac{1}{N} \log \Pr[\mathrm{LP}(P'_N, P') \geq \epsilon] \leq -\min\left\{ I(\hat{P}', P) : \mathrm{LP}(\hat{P}', P') \geq \epsilon \right\} < 0$$

where the minimum is indeed achieved as our good rate function has compact sublevel sets and the set of all $\hat{P}' \in \mathcal{P}(\Xi')$ such that the Lévy-Prokhorov metric ball $\{\hat{P}' : \mathrm{LP}(\hat{P}', P') \geq \epsilon\}$ is closed and does not contain the distribution P'. Hence, the large deviation property immediately implies that the empirical probability measure P'_N converges in probability to P' with an increasing number of observations. In fact, the rate function can be interpreted to quantify the exponential speed with which this convergence in probability takes place.

2.2. Efficiency

We consider a prescriptive problem in which we attempt to learn the solution to the stochastic optimization problem stated in Equation (1) from the noisy observational data described in Section 2.1. Let us denote with P_N^{ml} the maximum likelihood estimate for the unobserved probability distribution P. A straightforward extension of the sample average formulation in Equation (4) to this noisy data would be to consider

$$z(P_N^{\mathrm{ml}}) \in \arg_\epsilon \inf_{z\in Z} \mathbf{E}_{P_N^{\mathrm{ml}}}[\ell(z, \xi)]. \tag{15}$$

Many other formulations based on different distributional estimates are evidently possible as well. This naturally leads us to question if between these many alternative 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 [33] and define a data-driven formulation as consisting of a predictor and prescriptor.

Definition 2.6 *(Predictors and prescriptors).* A measurable function $\tilde{c} : Z \times \mathcal{P}(\Xi') \to \mathrm{R}$ is called a predictor. A measurable function $\tilde{z} : \mathcal{P}(\Xi') \to Z$ is called a prescriptor if there exists a predictor $\tilde{c}$ that induces $\tilde{z}$ in the sense that $\tilde{z}(\hat{P}') \in \arg_\epsilon \inf_{z\in Z} \tilde{c}(z, \hat{P}')$ for all $\hat{P}' \in \mathcal{P}(\Xi')$. That is, we have $\tilde{c}(\tilde{z}(\hat{P}'), \hat{P}') - \epsilon < \tilde{v}(\hat{P}') := \inf_{z\in Z} \tilde{c}(z, \hat{P}')$ 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 (15) employs the cost predictor $\mathbf{E}_{P_N^{\mathrm{ml}}}[\ell(z, \xi)]$ to prescribe its decision $z(P_N^{\mathrm{ml}})$. However, the maximum likelihood is a mere point estimate of the unobserved probability distribution P. The maximum likelihood formulation can consequently be expected to suffer similar shortcomings as the sample average formulation in the noiseless regime. 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

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

occurs with $c(z, P) = \mathbf{E}_P[\ell(z, \xi)]$ the unknown out-of-sample cost. Such disappointment events in which the actual cost of our decision, i.e., $c(\tilde{z}(P'_N), P)$, breaks the predicted cost or budget, i.e., $\tilde{c}(\tilde{z}(P'_N), P'_N)$, are undesirable and should be avoided by the decision-maker. Consequently, we prefer formulations which keep the disappointment rates

$$\varlimsup_{N\to\infty} \frac{1}{N} \log \Pr[P'_N \in \mathcal{D}(\tilde{c}, \tilde{z}; P)] \tag{16}$$

as small as possible for all $P \in \mathcal{P}$. We will only denote here formulations as feasible if their out-of-sample disappointment probability decays sufficiently fast, i.e., (16) $\leq -r$. Evidently, sufficiently fast disappointment probability decay can be achieved trivially by simply inflating the cost budgeted for each decision by some large nonnegative amount. We would hence prefer those formulations which promise minimal biased long term cost prediction $\lim_{N\to\infty} \tilde{c}(\tilde{z}(P'_N), P'_N)$ for all $P \in \mathcal{P}$.

We consider here the family of robust formulations defined by the predictor prescriptor pairs

$$\begin{aligned} \tilde{c}^\delta(z, P'_N) &:= \sup\left\{ \mathbf{E}_Q[\ell(z, \xi)] : Q \in \mathcal{P},\ I^\delta(P'_N, Q) \leq r \right\}, \\ \tilde{z}^\delta(P'_N) &\in \arg_\epsilon \inf_{z\in Z} \tilde{c}^\delta(z, P'_N) \end{aligned} \tag{17}$$

based on our smooth large deviation rate function defined in Equation (13). We will show using a large deviation argument that this family dominates the very rich class of regular formulations.

Definition 2.7 *(Regular predictors and prescriptors).* A 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 prescriptor $\tilde{z} : \mathcal{P}(\Xi') \to Z$ is called regular if it is continuous and there exists a regular predictor $\tilde{c}$ that induces $\tilde{z}$ in the sense that $\tilde{z}(\hat{P}') \in \arg_\epsilon \inf_{z\in Z} \tilde{c}(z, \hat{P}')$ for all $\hat{P}' \in \mathcal{P}(\Xi')$.

For regular predictors 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 [14, Theorem 11.4.1] for every $P \in \mathcal{P}$. Remark that the

class of all regular formulations is very rich as Definition 2.7 imposes only mild structural restrictions. The Berge maximum theorem [4, p. 116] indeed implies that the optimal value function $\tilde{v}(\hat{P}') = \min_{z\in Z} \tilde{c}(z, \hat{P}')$ of any regular formulation is a continuous function on $\mathcal{P}'(\Xi)$ already when the constraint set Z is merely compact. The correspondence $\hat{P}' \mapsto \{z \in Z \,:\, \tilde{c}(z, \hat{P}') < \tilde{v}(\hat{P}') + \epsilon\}$ 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 [4, 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 2.8. The cost function $c : Z \times \mathcal{P} \to \mathrm{R}, \quad (z, P) \mapsto \mathbf{E}_P[\ell(z, \xi)]$ is continuous.

We remark that Assumption 2.8 is rather mild and is already satisfied when the loss function $\ell : Z \times \Xi \to \mathrm{R}$ is merely bounded and uniformly continuous. The proofs of all results are found in Appendix A.

Theorem 2.9. *Let Assumption 2.8 hold. Then, the family of predictor prescriptor pairs $(\tilde{c}^\delta, \tilde{z}^\delta)$ is feasible for any $\delta > 0$, i.e.,*

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \Pr\left[P'_N \in \mathcal{D}(\tilde{c}^\delta, \tilde{z}^\delta; P)\right] \leq -r \quad \forall P \in \mathcal{P}. \tag{18}$$

Furthermore, consider any regular predictor prescriptor pair $(\tilde{c}, \tilde{z})$ which satisfies

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \log \Pr\left[P'_N \in \mathcal{D}(\tilde{c}, \tilde{z}; P)\right] \leq -r \qquad \forall P \in \mathcal{P}. \tag{19}$$

Then, we have that for all $\epsilon > 0$ there exists $0 < \delta'$ so that any $0 < \delta \leq \delta'$ we have almost surely

$$\lim_{N\to\infty} \tilde{c}(\tilde{z}(P'_N), P'_N) + 3\epsilon \geq \lim_{N\to\infty} \tilde{c}^\delta(\tilde{z}^\delta(P'_N), P'_N) \tag{20}$$

when the noisy data is generated by any distribution $P' \in \mathcal{P}'$.

Inequality (18) guarantees that any formulation in our family is feasible. Inequalities (19) and (20) guarantee that any other feasible regular formulation is dominated (modulo the small constant $3\epsilon > 0$) by members of our efficient 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 bias under Assumption 2.8 and is thus efficient.

In view of the previous discussion it is tempting to consider the data-driven formulation (17) with $\delta = 0$. However, recall again that for noisy data when the base measure m' defined in Assumption 2.1 fails to be atomic, the ambiguity set $\left\{Q \in \mathcal{P} \,:\, I(P'_N, Q) < \infty\right\} = \emptyset$ is trivial for any empirical distribution P'_N and consequently (17) with $\delta = 0$ is here obviously infeasible. Hence, considering a somewhat smoothed rate function I^δ instead of the rate function I directly seems unavoidable when faced with noisy observational data.

We conclude here by providing an interesting connection between the actual rate function I and the entropic optimal transport which sheds a light on the role entropic optimal transport plays in this noisy observational data regime.

Theorem 2.10. *Let Assumption 2.1 hold. The jointly convex rate function can be decomposed as*

$$I(\hat{P}', P) = \inf_{Q\in\mathcal{P}(\Xi)} \mathrm{KL}(Q, P) + \mathrm{KL}(\hat{P}', m') + W_1(\hat{P}', Q) \geq 0. \tag{21}$$

In the proof of the previous result we indicate that if $\hat{P}' \ll m'$ the infimum in Problem (21) is achieved at $Q^\star$ whose Radon-Nikodym derivative with respect to P is $\mathrm{d}Q^\star/\mathrm{d}P(\xi) = \int_{\Xi'} [\exp(-d(\xi, \xi'))/\int_\Xi \exp(-d(\xi'', \xi'))\mathrm{d}P(\xi'')]\mathrm{d}\hat{P}'(\xi')$ for all $\xi \in \Xi$. The optimization variable Q in Theorem 2.10 can be interpreted to represent the unobserved empirical distribution P_N of the noiseless data points. With this interpretation in mind the first term $\mathrm{KL}(Q, P)$ ensures that $\Pr[P_N \approx Q] \asymp \exp(-N \cdot \mathrm{KL}(Q, P))$ following Sanov's theorem and accounts for the fact that the empirical distribution P_N of the noiseless data may differ from unknown the distribution P when the number of training data points is finite. The last two terms account for the fact that we only observe the empirical distribution P'_N of the noisy data. One can show that this term quantifies indeed $\Pr[P'_N \approx \hat{P}' | P_N \approx Q] \asymp \exp(-N \cdot (\mathrm{KL}(\hat{P}', m') + W_1(\hat{P}', Q)))$. Informally, we have using the law of total probability that

$$\begin{aligned}
&\Pr[P'_N \approx \hat{P}'] \\
&= \int \Pr[P'_N \approx \hat{P}' | P_N \approx Q] \Pr[P_N \approx Q] \\
&\asymp \int \exp(-N \cdot (\mathrm{KL}(\hat{P}', m') + W_1(\hat{P}', Q))) \exp(-N \cdot \mathrm{KL}(Q, P)) \\
&\asymp \exp(-N \cdot \inf_{Q\in\mathcal{P}} \mathrm{KL}(Q, P) + [\mathrm{KL}(\hat{P}', m') + W_1(\hat{P}', Q)]).
\end{aligned}$$

We remark that the entropic optimal transport term $W_1(\hat{P}', Q)$ is defined in Equation (8) where its marginal transport cost d is identified here as the logarithm of the density function of the noise corrupting our data as indicated in Assumption 2.1. The term $\mathrm{KL}(\hat{P}', m')$ can be interpreted as a compensation term for the fact that any density function arbitrarily depends on the base measure m' considered.

2.3. Consistency

Strong out-of-sample guarantees such as those we impose in Equation (18) yield conservative formulations even when a large amount of data points are observed. Indeed, even in the noiseless regime [33], imposing the out-of-sample guarantee (11) with $r > 0$ necessarily leaves any feasible formulation to be inconsistent, i.e., $\mathbf{E}_P\left[\ell(z_{\mathrm{KL},r}(P_N), \xi)\right]$ does not necessarily converge to $\inf_z \mathbf{E}_P[\ell(z, \xi)]$. Intuitively, this is a direct result from the fact that the employed ambiguity set $\{Q \in \mathcal{P} \,:\, \mathrm{KL}(P_N, Q) \leq r\}$ does not reduce to $\{P\}$ when N is large as the robustness radius r is here constant. Nevertheless, if the probability of disappointment is only required to be remain bounded rather than to decay exponentially as in Equation (9) a consistent formulation can be derived [13,24] by simply reducing the robustness radius r_N with increasing N under mild technical conditions on the cost function ℓ.

We remark here that by imposing the strong out-of-sample guarantee in Equation (18), our family of efficient formulations is also inconsistent as their associated ambiguity sets $\{Q \in \mathcal{P} : I^\delta(P'_N, Q) \leq r\}$ do not shrink to $\{P\}$ even when N is large as both here both r as well as δ are constant. Motivated by the previous discussion we consider simply reducing r_N and δ_N as N grows and consider the robust formulation

$$\begin{aligned}
&\tilde{c}_N(z, P'_N) := \sup\left\{\mathbf{E}_Q[\ell(z, \xi)] : Q \in \mathcal{P},\ I^{\delta_N}(P'_N, Q) \leq r_N\right\}, \\
&\tilde{z}_N(P'_N) \in \arg_\epsilon \inf_{z\in Z} \tilde{c}_N(z, P'_N).
\end{aligned} \tag{22}$$

We first show that the ambiguity set associated with the previously introduced formulation contains indeed P with high probability as the number of observations is large and both r_N and δ_N are reduced at appropriate rates with increasing N.

Proposition 2.11. Assume that $P' = O \star P$ has bounded fourth moment and consider two nonincreasing sequences $r_N > 0$ and $\delta_N > 0$ with $r_N = \Omega(N^{\gamma-1})$ and $\delta_N = \Omega(N^{-\gamma'/(2\dim(\Xi'))})$ for some $0 < \gamma' < \gamma < 1$. Then, $\lim_{N\to\infty} \Pr[\exists Q \in \mathcal{P}(\Xi') \text{ s.t. } \mathrm{LP}(P'_N, Q) \leq \delta_N, I(Q, P) \leq r_N] = 1$.

When there exists $Q \in \mathcal{P}(\Xi')$ so that $\mathrm{LP}(P'_N, Q) \leq \delta_N$ and $I(Q, P) \leq r_N$, the distribution P must be contained in the ambiguity set of the predictor $\tilde{c}_N$ which by definition implies that $c_N(z, P'_N) > \mathbf{E}_P[\ell(z, \xi)]$ for any $z \in Z$. The previous proposition hence immediately implies that $\Pr[\mathbf{E}_P[\ell(\tilde{z}_N(P'_N), \xi)] > \tilde{c}_N(\tilde{z}_N(P'_N), P'_N)]$ decays to zero, however, not necessarily exponentially fast.

Remark 2.12 *(Finite sample guarantees).* We remark that Equation (A.9) in the proof of Proposition 2.11 provides a finite sample guarantee on the disappointment probability $\Pr\left[\mathbf{E}_P\left[\ell(\tilde{z}_N(P'_N), \xi)\right] > \tilde{c}_N(\tilde{z}_N(P'_N), P'_N)\right]$. This enables the construction for any desired maximum disappointment probability $\beta \in (0, 1]$ of two particular sequences $r^\star_N > 0$ and $\delta^\star_N > 0$ tending to zero at an appropriate speed (with in fact $r^\star_N = O(N^{\gamma-1})$ and $\delta^\star_N = O(N^{-\gamma'/(2\dim(\Xi'))})$ for some $0 < \gamma' < \gamma < 1$) so that $\Pr[\mathbf{E}_P[\ell(\tilde{z}_N(P'_N), \xi)] > \tilde{c}_N(\tilde{z}_N(P'_N), P'_N)] \leq \beta$ for all $N \geq 1$. Furthermore, we remark that our efficiency guarantee in Theorem 2.10 on the formulation proposed in Equation (17) is asymptotic in nature. The result does hence not say after what number of samples our asymptotic guarantees are supposed to kick in. In particular, for small $\delta > 0$, the associated ambiguity set $\{Q \in \mathcal{P} : I^\delta(P_N, Q) \leq r\}$ may be empty and hence the formulation disappoints as by convention here $\mathbf{E}_P[\ell(\tilde{z}^\delta(P'_N), \xi)] > \tilde{c}^\delta(\tilde{z}^\delta(P'_N), P'_N) = -\infty$. If finite sample guarantees are a concern, we propose the robust formulation in Equation (22) with $\delta_N = \delta/2 + \delta^\star_N$ and $r_N = r + r^\star_N$ which trivially satisfies both the finite sample guarantee $\Pr[\mathbf{E}_P[\ell(\tilde{z}_N(P'_N), \xi)] > \tilde{c}_N(\tilde{z}_N(P'_N), P'_N)] \leq \beta,\ \forall N \geq 1$ as $\tilde{c}_N(\tilde{z}_N(P'_N), P'_N) \geq \tilde{c}^{\delta/2}(\tilde{z}^{\delta/2}(P'_N), P'_N)$ and the asymptotic guarantee $\overline{\lim}_{N\to\infty} \frac{1}{N} \log \Pr[\mathbf{E}_P[\ell(\tilde{z}_N(P'_N), \xi)] > \tilde{c}_N(\tilde{z}_N(P'_N), P'_N)] \leq -r$ as $\delta_N \geq \delta^\star_N$ and $r_N \geq r^\star_N$. Furthermore, the proposed formulation does not asymptotically impose any additional conservatism as indeed $\lim_{N\to\infty} \tilde{c}^\delta(\tilde{z}^\delta(P'_N), P'_N) \geq \lim_{N\to\infty} \tilde{c}_N(\tilde{z}_N(P'_N), P'_N)$ due to $\lim_{N\to\infty} \delta_N = \delta/2 < \delta$.

When there exists $Q \in \mathcal{P}(\Xi')$ so that $\mathrm{LP}(P'_N, Q) \leq \delta_N$ and $I(Q, P) = \mathrm{KL}(Q, P') \leq r_N$, in fact all distributions in the set $\left\{Q \in \mathcal{P} : O \star Q = P'\right\}$ are contained in the ambiguity set of the predictor $\tilde{c}_N$. Proposition 2.11 establishes that with high probability the ambiguity set will not only contain P but in fact all distributions in the set $\left\{Q \in \mathcal{P} : O \star Q = P'\right\}$. Unsurprisingly, consistency demands at a bare minimum that the distribution P is identifiable from P', i.e., we must impose $\left\{Q \in \mathcal{P} : O \star Q = P'\right\} = \{P\}$ on the observational model O. Not all observational models satisfy this assumption as certain types of measurement noise may lead to information loss. In an extreme case where $O_\xi = P'$ independent of ξ then clearly all information regarding the noiseless data is lost and in fact $\left\{Q \in \mathcal{P} : O \star Q = P'\right\} = \mathcal{P}$. We remark that our efficiency notion is independent of identifiability. We prove here consistency for the additive error setting under a mild assumption on the loss function ℓ which is slightly stronger than the condition imposed in Assumption 2.8.

Assumption 2.13 *(Identifiability).* Let $\varphi_E : \mathrm{R}^{\dim(\Xi)} \to \mathrm{C},\ t \mapsto \int \exp(i \langle t, e \rangle)\, \mathrm{d}E(e)$ be the characteristic function of the error distribution E. Assume that φ_E has no roots, i.e., $\varphi_E(t) \neq 0$ for $t \in \mathrm{R}^{\dim(\Xi)}$.

Assumption 2.14 *(Bounded Lipschitz loss).* Assume that the loss function ℓ is uniformly bounded and Lipschitz, i.e., we have $L > 0$ and $\mathcal{L} > 0$ so that $\sup_{\xi\in\Xi} |\ell(z, \xi)| \leq L$ as well as $\sup_{z\in Z, \xi_1\neq\xi_2\in\Xi} |\ell(z, \xi_1) - \ell(z, \xi_2)|/\|\xi_1 - \xi_2\| \leq \mathcal{L}$ for all $z \in Z$.

Recall that any distribution in $\mathcal{P}(\mathrm{R}^{\dim(\Xi)})$ is uniquely determined by its characteristic function and that the characteristic function of a convolution between two distributions is given as the product of their characteristic functions [22, Chapter 6]. Assumption 2.13 guarantees that from the noisy distribution $P' = O^{AE} \star P$ we can identify the noiseless distribution P via its characteristic function $\varphi_P = \varphi_{P'}/\varphi_E$. Remark that the characteristic function of a zero mean normal distribution with variance σ^2 is given as $\varphi_{N(0,\sigma^2 I)}(t) = \exp(-\|\sigma t\|_2^2/2) > 0$ and hence Example 2.3 satisfies Assumption 2.13.

Theorem 2.15 *(Consistency). Consider an additive error model O^{AE} for which Assumption 2.13 is satisfied. Assume that $P' = O^{AE} \star P$ has bounded fourth order moment, the loss ℓ satisfies Assumption 2.14 and assume that r_N and δ_N decay to zero with $r_N = \Omega(N^{\gamma-1})$ and $\delta_N = \Omega(N^{-2\gamma'/\dim(\Xi')})$ for some $0 < \gamma' < \gamma < 1$. Then, $\lim_{N\to\infty} \Pr[\mathbf{E}_P[\ell(\tilde{z}_N(P'_N), \xi)] \leq \inf_{z\in Z} \mathbf{E}_P[\ell(z, \xi)] + 2\epsilon] = 1$.*

3. Finite formulations

Our family of efficient robust formulations in Equation (17) is stated in terms of a saddle-point problem which may be difficult to solve in general. Indeed, even the original stochastic optimization problem (1) may not be easy to solve. For the sake of simplicity we assume here that nothing is known about P and hence $\mathcal{P} = \mathcal{P}(\Xi)$. We remark though that the presented analysis generalizes to the case where $\mathcal{P}$ is a convex subset of $\mathcal{P}(\Xi)$ with only minor modifications. If the loss function $\ell(z, \xi)$ is convex in the decision variable z for any ξ then the robust formulations in Equation (17) only require the solution of a convex optimization problem in the decision variable. Whether or not the convex optimization problem characterizing the optimal decision is tractable depends on whether the prediction function $\tilde{c}^\delta(z, P'_N)$ can be evaluated efficiently which we will discuss now in more depth.

The maximization problem characterizing the prediction function $\tilde{c}^\delta(z, P'_N)$ in Equation (17) is convex in the distribution P. Indeed, the closed metric balls $\{\hat{P}' \in \mathcal{P}(\Xi') : \mathrm{LP}(\hat{P}', P'_N) \leq \delta\} := \{\hat{P}' \in \mathcal{P}(\Xi') : P'_N(B) \leq \hat{P}'(B^\delta) + \delta \quad \forall B \subseteq \Xi'\}$ are convex. Consequently, we have that the predictor is characterized here as

$$\begin{aligned} \tilde{c}^\delta(z, P'_N) = \sup\ & \mathbf{E}_Q[\ell(z, \xi)] \\ \text{s.t.}\ & Q \in \mathcal{P}(\Xi),\ \hat{P}' \in \mathcal{P}(\Xi'), \\ & I(\hat{P}', Q) \leq r,\ \mathrm{LP}(\hat{P}', P'_N) \leq \delta. \end{aligned} \tag{23}$$

However, even in the case where both event sets Ξ and Ξ' have finite cardinality the terminal Lévy-Prokhorov constraint is characterized using $2^{|\Xi'|}$ linear inequalities which becomes prohibitive even for moderately sized event sets Ξ'.

3.1. Strassen representation

Surprisingly, by exploiting the Strassen [31] representation the Lévy-Prokhorov metric need not result in intractable formulations. To the best of our knowledge, the application of the Strassen representation to derive a tractable reformulation of the Lévy-Prokhorov distance is novel.

Theorem 3.1 *(Strassen representation). Let $\Xi'_N = \mathrm{supp}(P'_N)$ denote the support of P'_N. Then, $\tilde{c}^\delta(z, P'_N)$ is equal to*

$$\begin{array}{ll} \sup & \mathbf{E}_Q\left[\ell(z,\xi)\right] \\ \text{s.t.} & Q \in \mathcal{P}(\Xi),\ \hat{P}' \in \mathcal{M}_+(\Xi'),\ T \in \mathcal{M}_+(\Xi' \times \Xi'_N), \\ & I(\hat{P}', Q) \leq r, \\ & \Pi_{\Xi'} T = \hat{P}', \Pi_{\Xi'_N} T = P'_N, \\ & \int_{\Xi' \times \Xi'_N} \mathbb{1}\left\{\left\|\xi' - \xi'_i\right\| \leq \delta\right\} \mathrm{d}T(\xi', \xi'_i) \geq 1 - \delta. \end{array} \tag{24}$$

The equivalent formulation stated in Theorem 3.1 can be solved efficiently using an off-the-shelf exponential cone optimization solver [10] when both event sets Ξ and Ξ' have finite cardinality. Perhaps the only complication is that the size of this equivalent formulation counts $O(|\Xi'||\Xi'_N|)$ variables for $O(|\Xi'| + |\Xi'_N|)$ constraints which may limit its practicality.

We now indicate that even if the event set Ξ' is not finite, optimization problem (24) still admits a finite reduction. Consider a finite partition of Ξ' which is generated by the closed balls around the observed data points, i.e.,

$$\Xi'_k = \cap_{i=1}^{|\Xi'_N|} \left\{ \xi' \in \Xi' \ : \ \begin{array}{ll} \left\|\xi' - \xi'_i\right\| \leq \delta & \text{if } b(k,i) = 1 \\ \left\|\xi' - \xi'_i\right\| > \delta & \text{if } b(k,i) = 0 \end{array} \right\}$$

for $0 \leq k \leq K-1$ with $K = 2^{|\Xi'_N|}$ and $b(k,i)$ the ith digit of the natural number k written down in binary notation. Clearly, we have $\Xi'_k \cap \Xi'_{k'} = \emptyset$ if $k \neq k'$ and $\cup_{k=0}^{K-1} \Xi'_k = \Xi'$. Note that it may happen that $\Xi'_k = \emptyset$ for some $0 \leq k \leq K-1$.

Lemma 3.2. *Let $\Xi'_N = \mathrm{supp}(P'_N)$ denote the support of P'_N and let Ξ be finite. Then, $\tilde{c}^\delta(z, P'_N) = \tilde{c}^\delta_f(z, P'_N)$ with*

$$\begin{array}{rll} \tilde{c}^\delta_f(z, P'_N) := \sup & \sum_{\xi \in \Xi} \ell(z,\xi) q_\xi & \\ \text{s.t.} & q_\xi \geq 0 \ \forall \xi \in \Xi,\ \hat{p}' \in \mathrm{R}^K,\ t \in \mathrm{R}^{K \times |\Xi'_N|} & \\ & \sum_{k=0}^{K-1} \hat{p}'_k \log\left(\frac{\hat{p}'_k}{\sum_{\xi \in \Xi} O_\xi(\Xi'_k) q_\xi}\right) \leq r, & \\ & \sum_{i=1}^{|\Xi'_N|} t_{k,i} = \hat{p}'_k \quad 0 \leq k \leq K-1, & \\ & \sum_{k=0}^{K-1} t_{k,i} = P'_N(\xi'_i) \quad 1 \leq i \leq \left|\Xi'_N\right|, & \\ & \sum_{i=1}^{|\Xi'_N|} \sum_{k=0}^{K-1} b(k,i) t_{k,i} \geq 1 - \delta. & \end{array} \tag{25}$$

It should be remarked although optimization problem (25) is of finite size it may be very large as indeed we have that $K = O(2^{|\Xi'_N|})$. This clearly limits the practical usefulness of (25) to situations where the number of observed distinct data points Ξ'_N is very small. Surprisingly, using a dual equivalent formulation instead the computational burden of evaluating $c^\delta(z, P'_N)$ can be further reduced to $O(|\Xi'_N|)$ variables and $O(1)$ constraints which is the subject of the following section.

3.2. Dual representation

Consider the minimization problem

$$\begin{array}{ll} \inf & \beta r - \int_{\Xi'} v(\xi') \mathrm{d}P'_N(\xi') + \gamma(\delta - 1) \\ & \quad + \max_{\xi \in \Xi}\left[\ell(z,\xi) + \beta \int_{\Xi'} \exp\left(u(\xi')/\beta - 1\right) \mathrm{d}O_\xi(\xi')\right] \\ \text{s.t.} & \beta \geq 0,\ u : \Xi' \to \mathrm{R},\ v : \Xi'_N \to \mathrm{R},\ \gamma \geq 0, \\ & \gamma \mathbb{1}\left\{\left\|\xi' - \xi'_i\right\| \leq \delta\right\} + v(\xi'_i) \leq u(\xi')\ \forall \xi' \in \Xi',\ \xi'_i \in \Xi'_N. \end{array} \tag{26}$$

We label the previous problem as the dual problem of the primal problem (24) which is nontrivial if $\tilde{c}^\delta(z, P'_N)$ is finite.

Theorem 3.3 *(Dual representation). Let Ξ be finite. Suppose that (23) enjoys the Slater constraint qualification condition, i.e., there exist $Q_s \in \mathcal{P}(\Xi)$, $\hat{P}'_s \in \mathcal{P}(\Xi')$ with $I(\hat{P}'_s, Q_s) < r$ and $\mathrm{LP}(\hat{P}'_s, P'_N) < \delta$. Then, we have $\tilde{c}^\delta(z, P'_N) =$ (26).*

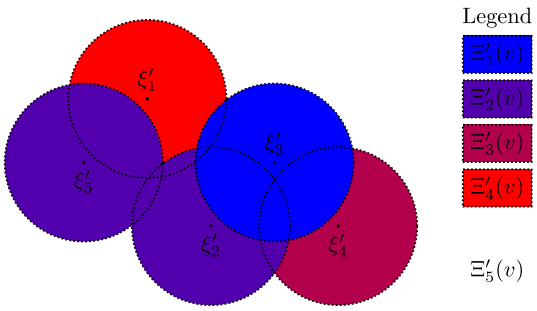

Fig. 1. The associated partition of the event set $\Xi'_N = \cup_{j=1}^{J(v)+1} \Xi'_j(v)$ associated with the vector v given in Example 3.4. (For an interpretation of the colors the reader is referred to the web version of this article.)

First note that the dual characterization of the prediction function in Theorem 3.3 allows for the convex saddle-point formulation (17) to be solved as an ordinary convex minimization problem as the dual formulation (26) is jointly convex in $z \in Z$ and the dual variables. This may be desirable in practice as saddle-point optimization solvers are typically not as mature as solvers addressing standard optimization problems. This dual formulation counts $O(|\Xi'_N| + |\Xi'|)$ variables for $O(|\Xi'||\Xi'_N|)$ constraints.

We now show that this can be further reduced to $O(|\Xi'_N|)$ variables for $O(1)$ constraints and in fact allows for a finite optimization characterization independent of the cardinality of the event set Ξ'. Consider $v : \Xi'_N \to \mathrm{R}$ and define $J(v) := |\{v(\xi') \ : \ \xi' \in \Xi'_N\}|$. Let $\xi'_{[j]}(v)$ for $j \in [1, \ldots, J(v)]$ denote any partition of the observed data points in Ξ'_N such that the dual variables $v(\xi'_{[j]}(v))$ are non-increasing. That is, we have $\Xi'_N = [\xi'_{[1]}(v), \ldots, \xi'_{[J(v)]}(v)]$ and $v(\tilde{\xi}') \geq v(\hat{\xi}')$ for all $\tilde{\xi}' \in \xi'_{[j]}(v), \hat{\xi}' \in \xi'_{[j']}(v)$ with $1 \leq j < j' \leq J(v)$ as well as $v(\tilde{\xi}') = v(\hat{\xi}')$ for all $\tilde{\xi}', \hat{\xi}' \in \xi'_{[j]}(v)$ with $1 \leq j \leq J(v)$. We may now partition the event set $\Xi' = \cup_{j=1}^{J(v)+1} \Xi'_j(v)$ using the sets $\Xi'_0(v) = \emptyset$, $\Xi'_j(v) = \{\xi' \in \Xi' \ : \ \min\{||\xi' - \tilde{\xi}'|| \ : \ \tilde{\xi}' \in \xi'_{[j]}(v)\} \leq \delta\} \setminus \cup_{l \in [1,\ldots,j-1]} \Xi'_l(v)$ for $j \in [1, \ldots, J(v)]$, and $\Xi'_{J(v)+1}(v) = \Xi' \setminus \cup_{l \in [1,\ldots,J(v)]} \Xi'_l(v)$.

Example 3.4. Let $\Xi'_N = [\xi'_1, \ldots, \xi'_5]$ and consider $v : \Xi'_N \to \mathrm{R}$ given as $v(\xi'_1) = 1$, $v(\xi'_2) = v(\xi'_5) = 3$, $v(\xi'_3) = 4$ and $v(\xi'_4) = 2$. Then, $J(v) = 4$ and $\xi'_{[1]}(v) = \{\xi'_3\}$, $\xi'_{[2]}(v) = \{\xi'_2, \xi'_5\}$, $\xi'_{[3]}(v) = \{\xi'_4\}$, $\xi'_{[4]}(v) = \{\xi'_1\}$. The associated partition of the event set Ξ' is given in Fig. 1.

Lemma 3.5 *(Finite convex dual representation). The dual problem (26) admits the following convex reformulation*

$$\begin{array}{ll} \inf & \beta r - \sum_{j=1}^{J(v)} v(\xi'_{[j]}) P'_N(\xi'_{[j]}) + \gamma(\delta - 1) \\ & \quad + \max_{\xi \in \Xi}\left[\ell(z,\xi) + O_\xi(\Xi'_{J(v)+1}(v)) \frac{\beta}{e} \exp\left(\frac{v(\xi'_{[1]}(v))}{\beta}\right)\right. \\ & \quad \left. + \sum_{j=1}^{J(v)} O_\xi(\Xi'_j(v)) \frac{\beta}{e} \exp\left(\frac{\max(v(\xi'_{[1]}(v)), v(\xi'_{[j]}(v)) + \gamma)}{\beta}\right)\right] \\ \text{s.t.} & \beta \geq 0,\ v : \Xi'_N \to \mathrm{R},\ \gamma \geq 0. \end{array} \tag{27}$$

The dual representation (27) can be solved as a finite convex optimization problem using a (stochastic) black-box optimization method [26] as long as we have a (stochastic) oracle which can evaluate the probabilities $O_\xi(\Xi'_j(v))$ for all $\xi \in \Xi$ and $j \in$

$[1, \ldots, J(v) + 1]$. The complexity of our efficient formulation (23) is hence reduced to the complexity of integration of the noise distribution over certain intersections and unions of norm balls; see also Fig. 1.

Data availability

No data was used for the research described in the article.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.orl.2024.107089.

References

[1] J.-P. Aubin, H. Frankowska, Set-Valued Analysis, Springer Science & Business Media, 2009.
[2] B. Bank, J. Guddat, D. Klatte, B. Kummer, K. Tammer, Non-linear Parametric Optimization, Springer, 1982.
[3] G. Bayraksan, D. Love, Data-driven stochastic programming using phi-divergences, in: The Operations Research Revolution, in: INFORMS, 2015, pp. 1–19.
[4] C. Berge, Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity, Courier Corporation, 1997.
[5] D. Bertsimas, I. Popescu, Optimal inequalities in probability theory: a convex optimization approach, SIAM J. Optim. 15 (3) (2005) 780–804.
[6] D. Bertsimas, V. Gupta, N. Kallus, Data-driven robust optimization, Math. Program. 167 (2) (2018) 235–292.
[7] J. Blanchet, Y. Kang, K. Murthy, Robust Wasserstein profile inference and applications to machine learning, J. Appl. Probab. 56 (3) (2019) 830–857.
[8] R. Chen, I.C. Paschalidis, A robust learning approach for regression models based on distributionally robust optimization, J. Mach. Learn. Res. 19 (13) (2018) 1–48.
[9] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, Adv. Neural Inf. Process. Syst. 26 (2013) 2292–2300.
[10] J. Dahl, E. Andersen, A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization, Math. Program. (2021) 1–30.
[11] E. Delage, Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, Oper. Res. 58 (3) (2010) 595–612.
[12] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, vol. 38, Springer Science & Business Media, 2009.
[13] J. Duchi, P. Glynn, H. Namkoong, Statistics of robust optimization: a generalized empirical likelihood approach, Math. Oper. Res. (2021).
[14] R.M. Dudley, Real Analysis and Probability, CRC Press, 2018.
[15] P.M. Esfahani, D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations, Math. Program. 171 (1–2) (2018) 115–166.
[16] I. Gijbels, Censored data, Wiley Interdiscip. Rev.: Comput. Stat. 2 (2) (2010) 178–188.
[17] J. Goh, M. Sim, Distributionally robust optimization and its tractable approximations, Oper. Res. 58 (4-part-1) (2010) 902–917.
[18] J. Gotoh, M.J. Kim, A.E.B. Lim, Robust empirical optimization is almost the same as mean–variance optimization, Oper. Res. Lett. 46 (4) (2018) 448–452.
[19] P. Groeneboom, J.A. Wellner, Information Bounds and Nonparametric Maximum Likelihood Estimation, vol. 19, Springer Science & Business Media, 1992.
[20] V. Gupta, Near-optimal Bayesian ambiguity sets for distributionally robust optimization, Manag. Sci. 65 (9) (2019) 4242–4260.
[21] Z. Hu, L.J. Hong, Kullback-Leibler divergence constrained distributionally robust optimization, Available at Optim. Online (2013).
[22] A.F. Karr, Probability, Springer, New York, 1993.
[23] D. Kuhn, P.M. Esfahani, V.A. Nguyen, S. Shafieezadeh-Abadeh, Wasserstein distributionally robust optimization: theory and applications in machine learning, in: Operations Research & Management Science in the Age of Analytics, in: INFORMS, 2019, pp. 130–166.
[24] H. Lam, Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization, Oper. Res. 67 (4) (2019) 1090–1105.
[25] R.O. Michaud, The Markowitz optimization enigma: Is "optimized" optimal?, Financ. Anal. J. 45 (1) (1989) 31–42.
[26] Y. Nesterov, Introductory Lectures on Convex Programming Volume i: Basic Course, Springer, 1998.
[27] G. Peyré, M. Cuturi, Computational optimal transport: with applications to data science, Found. Trends Mach. Learn. 11 (5–6) (2019) 355–607.
[28] F. Santambrogio, Optimal Transport for Applied Mathematicians, Birkhäuser, NY, 2015.
[29] H. Scarf, A min max solution of an inventory problem, in: Studies in the Mathematical Theory of Inventory and Production, 1958.
[30] S. Shafieezadeh-Abadeh, D. Kuhn, P.M. Esfahani, Regularization via mass transportation, J. Mach. Learn. Res. 20 (103) (2019) 1–68.
[31] V. Strassen, The existence of probability measures with given marginals, Ann. Math. Stat. 36 (2) (1965) 423–439.
[32] B.P. Van Parys, P.J. Goulart, D. Kuhn, Generalized Gauss inequalities via semidefinite programming, Math. Program. 156 (2016) 271–302.
[33] B.P. Van Parys, P.M. Esfahani, D. Kuhn, From data to decisions: distributionally robust optimization is optimal, Manag. Sci. (2020).
[34] W. Wiesemann, D. Kuhn, M. Sim, Distributionally robust convex optimization, Oper. Res. 62 (6) (2014) 1358–1376.
[35] W. Xie, Tractable reformulations of distributionally robust two-stage stochastic programs with $\infty-$ Wasserstein distance, arXiv preprint arXiv:1908.08454, 2019.