IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

# Distributionally-Robust Optimization with Noisy Data for Discrete Uncertainties Using Total Variation Distance

Farhad Farokhi

*Abstract*— **Stochastic programs, where uncertainty distribution must be inferred from *noisy* data samples, are considered. They are approximated with distributionally-robust optimizations that minimize the worst-case expected cost over ambiguity sets, i.e., sets of distributions that are sufficiently compatible with observed data. The ambiguity sets capture probability distributions whose convolution with the noise distribution is within a ball centered at the empirical noisy distribution of data samples parameterized by total variation distance. Using the prescribed ambiguity set, the solutions of the distributionally-robust optimizations converge to the solutions of the original stochastic programs when the number of the data samples grow to infinity. Therefore, the proposed distributionally-robust optimization problems are *asymptotically consistent*. The distributionally-robust optimization problems can be cast as *tractable optimization problems*.**

## I. Introduction

In this paper, we consider a single-stage stochastic program of the form $\inf_{x\in\mathbb{X}}\mathbb{E}^{\mathbb{P}}[h(x,\xi)]$, where $x\in\mathbb{R}^n$ is a decision variable that must be determined to minimize the expected cost $\mathbb{E}^{\mathbb{P}}[h(x,\xi)]$ while $\xi\in\mathbb{R}^m$ is a random uncertainty with distribution $\mathbb{P}$. Distribution $\mathbb{P}$ is not available and must be inferred from data samples. Merely relying on the empirical distribution evaluated using data samples, instead of the original distribution $\mathbb{P}$, can result in disappointing outcomes. This is known colloquially as the "optimizer's curse" and is caused by over-fitting [1]. One way to avoid this concern is to instead consider a set of distributions that are sufficiently compatible with the observed data, known as an ambiguity set $\mathcal{P}$, and minimizes the worst-case expected cost over the ambiguity set $\sup_{\mathbb{Q}\in\mathcal{P}}\mathbb{E}^{\mathbb{Q}}[h(x,\xi)]$. This approach is known as distributionally-robust optimization [2].

Distributionally-robust optimization dates back to ambiguity-averse or robust news-vendor problem [3]. However, there has been a recent revival in the field due to computationally-favourable reformulations [2], [4]–[10]. These results mostly differ on how they construct the ambiguity sets using, e.g., moment constraints [5], the Prohorov metric [6], the Kullback-Leibler divergence [7], and the Wasserstein metric [2], [9].

Distributionally-robust optimization has however mostly focused on noiseless data [2], [4]–[9], i.e., high-quality independently and identically realized samples from distribution $\mathbb{P}$ are required. Noting that we sometimes only have access to noisy data samples, there has been some recent investigations into situations where the data samples are noisy [11], [12]. For instance, data samples may be intentionally corrupted by privacy-preserving noise [13]. Alternatively, data can be truly noisy because of inherent instrumentation uncertainty or noise. The problem with noisy data is that, even with infinitely-many data points, the empirical distribution does not coverage to the original distribution $\mathbb{P}$. The empirical distribution instead converges to $\mathbb{P}'$, which is the outcome of convolution of the original distribution $\mathbb{P}$ and the noise distribution $\mathbb{O}$ [11], [12]. One way to pose distributionally-robust optimization in the presence of noisy data is to expand the ambiguity sets, by setting the radius of the uncertainty ball around the empirical noisy distribution large enough, to contain the noiseless distribution [12], [14]. This however can create unnecessary conservatism because we have to deal with the worst-case expected cost over a large ambiguity set that contains distributions that may behave differently from the original noiseless distribution. Such ideas effectively dictate that the radius of the ambiguity set must remain non-trivially large even in the big data regime. Therefore, these method can only provide useful solutions and guarantees in the small noise regime, i.e., when the variance or entropy of the noise is relatively small, and thus enlargement of the ambiguity set is minimally conservative. As opposed, in Section III of this paper, we observe that the radius of the ambiguity set converges to zero as more samples are gathered. Therefore, the method of this paper is asymptotically less conservative. Another way is to use other metrics for constructing the ambiguity set [11]. The alternatives however still suffer from conservatism discussed above. None of these studies consider the specific way that the noise distribution changes the original distribution, i.e., through convolution. In fact, in the large data regime, it is reasonable to expect that the effect of the noisy measurements is negligible. This is because we can always remove the effect of the noise by density deconvolution [15], [16] even if the noise is significant, i.e., it has large variance or entropy.

An important aspect of obtaining good results in distributionally-robust optimization is to appropriately select the ambiguity set. The ambiguity set must be large enough to contain the original density $\mathbb{P}$ while it must be small enough to not make the results conservative. The solution of $\sup_{\mathbb{Q}\in\mathcal{P}}\mathbb{E}^{\mathbb{Q}}[h(x,\xi)]$ must remain reasonably close to $\mathbb{E}^{\mathbb{P}}[h(x,\xi)]$. Otherwise, we incentivize overly conservative decisions by considering distributions $\mathbb{Q}\in\mathcal{P}$ that are far from reality $\mathbb{P}$. The ambiguity set must also be easily reconstructable from the data samples to ensure computationally tractable reformulation of the distributionally-robust optimization problem. In this paper, we define the ambiguity

F. Farokhi is with the Department of Electrical and Electronic Engineering, The University of Melbourne, Australia. This work has been financially supported, in part, by a grant from the Australian Research Council (ARC) DP210102454.

set by considering a set of probability distributions whose convolution with the known noise distribution remains appropriately close to the empirical noisy distribution of the data samples in the sense of the total variation distance. We prove that, using the prescribed ambiguity set, the solution of the distributionally-robust optimization converges to the solution of the original stochastic program when the number of the data samples grows to infinity. Therefore, the distributionally-robust optimization problem is asymptotically consistent. To prove this result, we need to assume that the distribution of the noise is uniformly diagonally dominant. Finally, we show that the robust optimization problem can be cast as a tractable convex optimization problem.

The rest of the paper is organized as follows. First, we finish this section by presenting some useful notations. Subsequently, in Section II, we formally define distributionally-robust optimization based on noisy data samples. In Section III, we consider asymptotic properties of the distributionally-robust optimization, such as asymptotic consistency, using concentration bounds for learning discrete distributions. We re-cast the distributionally-robust optimization as a tractable convex optimization problem in Section IV. Finally, we present some numerical results in Section V and conclude the paper in Section VI.

### NOTATION

For any set $\mathcal{A}$, the cardinality of the set is denoted by $|\mathcal{A}|$. For any finite set $\mathcal{A}$, i.e., $|\mathcal{A}|<\infty$, $\Delta(\mathcal{A})$ denotes the probability simplex on $\mathcal{A}$. The product of two probability distributions $\mathbb{P}_1\in\Delta(\Xi_1)$ and $\mathbb{P}_2\in\Delta(\Xi_2)$ is the distribution $\mathbb{P}_1\otimes\mathbb{P}_2\in\Delta(\Xi_1\times\Xi_2)$. The $N$-fold product of a distribution $\mathbb{P}\in\Delta(\Xi)$ is denoted by $\mathbb{P}^N\in\Delta(\Xi^N)$. The total variation distance between any two distributions $\mathbb{P}_1,\mathbb{P}_2\in\Delta(\Xi)$ is

$$\begin{aligned}\mathrm{d}_{\mathrm{TV}}(\mathbb{P}_1,\mathbb{P}_2) &:= \sup_{\mathcal{A}\subseteq\Xi}(\mathbb{P}_1(\mathcal{A})-\mathbb{P}_2(\mathcal{A}))\\ &=\frac{1}{2}\sum_{\xi\in\Xi}|\mathbb{P}_1(\xi)-\mathbb{P}_2(\xi)|\in[0,1].\end{aligned}$$

## II. DATA-DRIVEN PROGRAMMING

Consider the stochastic program

$$J^\star := \inf_{x\in\mathbb{X}}\left\{\mathbb{E}^{\mathbb{P}}[h(x,\xi)]=\sum_{\xi\in\Xi}h(x,\xi)\mathbb{P}(\xi)\right\}, \quad (1)$$

with feasible set $\mathbb{X}\subseteq\mathbb{R}^n$, discrete/finite uncertainty set $\Xi\subseteq\mathbb{R}^m$ (i.e., $|\Xi|<\infty$), and loss function $h:\mathbb{X}\times\Xi\rightarrow\mathbb{R}$. The loss function $h$ depends on the decision vector $x\in\mathbb{R}^n$ and the random variable $\xi\in\mathbb{R}^m$, whose distribution $\mathbb{P}$ is supported on $\Xi$, i.e., $\mathbb{P}\in\Delta(\Xi)$. The distribution $\mathbb{P}$ is *unknown*. Therefore problem (1) cannot be solved exactly. Although $\mathbb{P}$ is unknown, it is partially observable through a finite set of $N$ independently and identically distributed (i.i.d.) *noisy* samples

$$\xi_i'\sim\mathbb{O}(\cdot|\xi_i),\quad \xi_i\sim\mathbb{P},\quad i\in\{1,\dots,N\}, \quad (2)$$

where $\xi_i\sim\mathbb{P}$ are i.i.d. samples from $\mathbb{P}\in\Delta(\Xi)$ while $\xi_i'$ are noisy observations of $\xi_i$ realized according to the conditional distribution $\mathbb{O}(\cdot|\xi_i)\in\Delta(\Xi')$. The marginal distribution of the noisy observations is given by

$$\mathbb{P}'(\xi')=\sum_{\xi\in\Xi}\mathbb{O}(\xi'|\xi)\mathbb{P}(\xi),\quad \forall\xi'\in\Xi'. \quad (3)$$

Note that support set of $\mathbb{P}'$, which is $\Xi'$, may not necessarily be equal to the support set of $\mathbb{P}$, which is $\Xi$. Nonetheless, we assume that $\Xi',\Xi\subseteq\mathbb{R}^m$. This assumption is not strictly necessary, however, it simplifies the narrative without substantial conservatism (up to relabeling the elements in $\Xi'$). For instance, noisy data based on additive noise satisfies this condition. For the sake of brevity, we write that

$$\mathbb{P}'=\mathbb{O}\star\mathbb{P}. \quad (4)$$

Note that, here, we have opted for the convolution notation '$\star$' because the relationship in (3) is visually similar to discrete convolution of the probability mass functions, particularly when the noise is additive[1].

**Remark 2.1 (Known Noise Distribution)** *We assume that the distribution of the noise $\mathbb{O}$ is known. The motivation for this is twofold. First, in privacy-preserving applications (e.g., the example in Section V), the distribution of the noise, which is a function of the privacy budget and privacy-preserving mechanism, is often publicly known to improve transparency and accountability and to also allow for post processing [17]. Furthermore, in sensing and instrumentation, the distribution of the noise is usually publicly shared in datasheets while the distribution of the underlying variable is oft unknown. When the distribution of the noise is not known two approaches can be used. We can either use repeated measurements to estimate the distribution of the noise [18], which is not possible in privacy-preserving applications as repeated queries/responses erode privacy, or follow a worst-case adversarial approach [19], which may be conservative.*

**Remark 2.2 (Finite Uncertainty Set)** *The assumption that the uncertainty set $\Xi$ is finite is natural in some cases. For instance, the data could be inherently discrete, such as categorical attributes [20] and finite state spaces [21]. In other instances, communication constraints or post-processing can render the data discrete [22], [23].*

**Remark 2.3 (Impact of Density Deconvolution)** *Combining density deconvolution [16] and techniques used for developing computationally-efficient distributionally-robust optimization [2] is non-trivial. This is because after deconvolution, the empirical density function is no longer composed of delta functions, which breaks down an important step in the proof of Theorem 4.2 in [2].*

We denote the training dataset composed of the noisy samples by $\Xi_N':=\{\xi_i'\}_{i=1}^N$. A data-driven solution for problem (1) is a feasible decision $\hat{x}_N\in\mathbb{X}$ that is constructed from the training dataset $\Xi_N'$. The *out-of-sample performance*

[1] When dealing with additive noise, i.e., $\xi_i'=\xi_i+n_i$, the marginal distribution of the noisy observations in (3) can be rewritten as $\mathbb{P}'(\xi')=\sum_{\xi\in\Xi}\mathbb{O}(\xi'-\xi)\mathbb{P}(\xi)$ as, by slight abuse of notation, $\mathbb{O}(\xi'|\xi)=\mathbb{O}(\xi'-\xi)$. This implies that $\mathbb{P}'$ is equal to convolution of $\mathbb{O}$ and $\mathbb{P}$.

of $\hat{x}_N$ is defined as $\mathbb{E}^{\mathbb{P}}[h(\hat{x}_N,\xi)]$, which is the expected cost of the data-driven solution $\hat{x}_N \in \mathbb{X}$ for a new sample $\xi$ from $\mathbb{P}$, which is independent of the training dataset. Since $\mathbb{P}$ is unknown, the out-of-sample performance cannot be evaluated explicitly in practice. Therefore, we would like to establish performance guarantees for the out-of-sample performance. By construction, because of the feasibility of $\hat{x}_N \in \mathbb{X}$, we know that

$$\mathbb{E}^{\mathbb{P}}[h(\hat{x}_N,\xi)] \geq J^\star,$$

where $J^\star$ is the optimal solution in (1). In line with the literature on distributionally-robust optimization [2], we are interested in out-of-sample performance bounds:

$$\mathbb{P}'^N\{\Xi'_N : \mathbb{E}^{\mathbb{P}}[h(\hat{x}_N,\xi)] \leq \hat{J}_N\} \geq 1-\beta,$$

where $\hat{J}_N$ is an upper bound that potentially depends on the training dataset and $\beta \in (0,1)$ is a significance or confidence parameter. Note that the dataset $\Xi'_N$ is a random variable governed by the $N$-fold product[2] distribution $\mathbb{P}'^N \in \Delta(\Xi'^N)$.

One way to solve this problem is to compute the discrete empirical probability distribution $\hat{\mathbb{P}}'_N \in \Delta(\Xi')$:

$$\hat{\mathbb{P}}'_N(\xi) = \frac{1}{N}\sum_{i=1}^{N}\delta[\xi'_i - \xi], \tag{5}$$

where $\delta : \mathbb{R}^m \to \mathbb{R}$ is the Kronecker delta function, i.e., $\delta[x] = 1$ if $x = 0$ and $\delta[x] = 0$ otherwise. This amounts to approximating the stochastic program in (1) with the noisy sample-average approximation (NSAA) problem:

$$\hat{J}_{\mathrm{NSAA}} := \inf_{x\in\mathbb{X}}\left\{\mathbb{E}^{\hat{\mathbb{P}}'_N}[h(x,\xi)] = \frac{1}{N}\sum_{i=1}^{N}h(x,\xi'_i)\right\}. \tag{6}$$

However, it should be noted that $\lim_{N\to\infty}\mathbb{E}^{\hat{\mathbb{P}}'_N}[h(x,\xi)] \overset{\text{a.s.}}{=} \mathbb{E}^{\mathbb{P}'}[h(x,\xi)] \neq \mathbb{E}^{\mathbb{P}}[h(x,\xi)]$. This is caused by the noisy nature of the samples. In this paper, we address this problem by explicitly considering the effect of noisy measurements. We particularly use $\Xi'_N$ to create an ambiguity set $\hat{\mathcal{P}}_N \subseteq \Delta(\Xi)$ containing all distributions that could have generated the noiseless samples with high confidence. This ambiguity set enables us to define the distributionally-robust optimization problem:

$$\hat{J}_{\mathrm{DRO}}(\hat{\mathcal{P}}_N) := \inf_{x\in\mathbb{X}}\sup_{\mathbb{Q}\in\hat{\mathcal{P}}_N}\mathbb{E}^{\mathbb{Q}}[h(x,\xi)]. \tag{7}$$

If the optimal solution to (7) exists and is attained for some element of $\mathbb{X}$, we denote the solution with $\hat{x}_{\mathrm{DRO}}(\hat{\mathcal{P}}_N)$. We drop the reference to $\hat{\mathcal{P}}_N$ and use $\hat{J}_{\mathrm{DRO}}$ and $\hat{x}_{\mathrm{DRO}}$ instead of $\hat{J}_{\mathrm{DRO}}(\hat{\mathcal{P}}_N)$ and $\hat{x}_{\mathrm{DRO}}(\hat{\mathcal{P}}_N)$, respectively, when the ambiguity set is clear from the context.

**Remark 2.4 (Existence of Solution)** *The inner optimization problem in* (7), *i.e.,* $\sup_{\mathbb{Q}\in\hat{\mathcal{P}}_N}\mathbb{E}^{\mathbb{Q}}[h(x,\xi)]$, *possesses a finite supremum if the cost function* $h(x,\xi)$ *is bounded and the ambiguity set* $\hat{\mathcal{P}}_N$ *is compact, which holds for the ambiguity sets defined using total variation distance in Section III. Existence of solution to the outer problem, i.e.,* $\inf_{x\in\mathbb{X}}(\sup_{\mathbb{Q}\in\hat{\mathcal{P}}_N}\mathbb{E}^{\mathbb{Q}}[h(x,\xi)])$, *is more subtle. The solution exists and is attained whenever* $\sup_{\mathbb{Q}\in\hat{\mathcal{P}}_N}\mathbb{E}^{\mathbb{Q}}[h(x,\xi)]$ *is continuous in* $x$ *(e.g., if the solution to the inner problem is unique, and the cost function is bounded and continuous) and the feasible set* $\mathbb{X}$ *is compact. Note that existence of solutions does not imply computational feasibility of finding one. The latter requires extra assumptions, e.g., convexity.*

In what follows, we construct $\hat{\mathcal{P}}_N$ as a ball around the empirical distribution (5) with respect to the total variation distance with an additional constraint based on density convolution to account for the noisy nature of the data.

## III. Concentration Bounds on Empirical Discrete Distributions

The following concentration bounds provide the basis for establishing finite sample guarantees that we use to develop the ambiguity set $\hat{\mathcal{P}}_N \subseteq \Delta(\Xi)$ in the distributionally-robust optimization framework of (7).

**Theorem 3.1 ([24, Theorem 1])** *For the empirical distribution in* (5), $\mathbb{P}'^N\{\mathrm{d}_{\mathrm{TV}}(\mathbb{P}',\hat{\mathbb{P}}'_N) \leq \varepsilon\} \geq 1-\alpha$ *if* $N \geq \max\{|\Xi|, 2\ln(2/\alpha)\}/\varepsilon^2$.

Let us define total-variation ball $\mathcal{B}'_{\mathrm{TV},\varepsilon}(\hat{\mathbb{P}}'_N) := \{\mathbb{Q} : \mathrm{d}_{\mathrm{TV}}(\mathbb{Q},\hat{\mathbb{P}}'_N) \leq \varepsilon\}$ and

$$\varepsilon_{\mathrm{TV}}(\alpha) := \sqrt{\frac{\max\{|\Xi|, 2\ln(2/\alpha)\}}{N}}. \tag{8}$$

The following corollary follows from Theorems 3.1.

**Corollary 3.1:** $\mathbb{P}'^N\{\mathbb{P}' \in \mathcal{B}'_{\mathrm{TV},\varepsilon_{\mathrm{TV}}(\alpha)}(\hat{\mathbb{P}}'_N)\} \geq 1-\alpha$.

Note that, so far, we have been focused on concentration bounds for learning $\mathbb{P}'$. However, to solve (1), we must learn $\mathbb{P}$. Let us define sets

$$\mathcal{B}_{\mathrm{TV},\varepsilon}(\hat{\mathbb{P}}'_N) := \{\mathbb{Q} : \mathbb{O}\star\mathbb{Q} \in \mathcal{B}'_{\mathrm{TV},\varepsilon}(\hat{\mathbb{P}}'_N)\}. \tag{9}$$

The following corollary immediately follows from Corollary 3.1 and the definition of the ambiguity set in (9).

**Corollary 3.2:** $\mathbb{P}'^N\{\mathbb{P} \in \mathcal{B}_{\mathrm{TV},\varepsilon_{\mathrm{TV}}(\alpha)}(\hat{\mathbb{P}}'_N)\} \geq 1-\alpha$.

**Remark 3.1 (Total Variation Distance vs. Other Probability Metrics)** *Various probability metrics, such as Kullback–Leibler divergence [7] and Wasserstein distance [2], are used for defining the ambiguity sets in distributionally-robust optimization. The use of total variation distance in this paper gives rise to a linear programming problem for computing the worst-case distribution (see Theorem 4.1), which is computationally favorable. The use of Kullback–Leibler divergence would have resulted in a convex nonlinear program. Also note that the total variation distance is an optimal transportation distance with an indicator cost function [25]. To use other optimal transport distances, such as the Wasserstein distance, we must endow the finite uncertainty sets* $\Xi$ *and* $\Xi'$ *with distances. For categorical sets that are not subsets of metric spaces, e.g.,* $\{$*male*, *female*$\}$, *distances can be artificial.*

[2] Note that $\Xi'^N$ denotes $N$-fold Cartesian product of set $\Xi'$ with itself, i.e., $\Xi'^N = \Xi' \times \cdots \times \Xi'$, while $\Xi'_N$ denotes the set of noisy data samples, i.e., $\Xi'_N = \{\xi'_i\}_{i=1}^N$. Therefore, by definition, $\Xi'_N \subseteq \Xi'^N$. The same distinction also holds for $\Xi^N$ and $\Xi_N$.

**Theorem 3.2 (Out-of-Sample Performance)** *Assume that $\hat{J}_{\mathrm{DRO}}$ and $\hat{x}_{\mathrm{DRO}}$ denote the optimal value and an optimizer of the distributionally-robust optimization problem* (7)*, where the ambiguity set is $\hat{\mathcal{P}}_N = \mathcal{B}_{\mathrm{TV},\varepsilon_{\mathrm{TV}}(\alpha)}(\hat{\mathbb{P}}'_N)$. Then,*

$$\mathbb{P}'^N\{\Xi'_N : \mathbb{E}^{\mathbb{P}}[h(\hat{x}_{\mathrm{DRO}},\xi)] \le \hat{J}_{\mathrm{DRO}}\} \ge 1-\alpha.$$

*Proof:* Corollary 3.2 shows that $\mathbb{P} \in \hat{\mathcal{P}}_N$ with probability of at least $1-\alpha$. Therefore, with probability of at least $1-\alpha$, $\mathbb{E}^{\mathbb{P}}[h(\hat{x}_{\mathrm{DRO}},\xi)] \le \sup_{\mathbb{Q}\in\hat{\mathcal{P}}_N}\mathbb{E}^{\mathbb{Q}}[h(\hat{x}_{\mathrm{DRO}},\xi)] = \hat{J}_{\mathrm{DRO}}$. ■

An important property for a stochastic estimator with access to random samples is consistency, i.e., the property that, as the number of samples increases towards infinity, the resulting sequence of estimates converges in some sense (convergence in probability or almost sure convergence) to the true solution [26, § 2.3]. This has been a particularly sought-after property in distributionally-robust optimization [2]. To prove consistency, we make the following assumption regarding the conditional probability of the noisy measurements.

**Assumption 3.1:** For $\Xi' = \Xi$, $\mathbb{O}$ is uniformly diagonally dominant if $\min_{\xi\in\Xi}\mathbb{O}(\xi|\xi) > |\Xi|\max_{\xi,\xi'\in\Xi,\xi\neq\xi'}\mathbb{O}(\xi'|\xi)$.

Assumption 3.1 focuses on conditional distributions that are uniformly diagonally dominant. This is a slightly stronger notion than diagonal dominance, i.e., $\mathbb{O}(\xi|\xi) > \sum_{\xi'\neq\xi}\mathbb{O}(\xi'|\xi)$. Diagonal dominance is a powerful tool for analysis in linear algebra [27, § 2] and probability [28]. Assumption 3.1 essentially requires that the data is not heavily perturbed by the noise.

**Theorem 3.3:** Assume that $\alpha_N \in (0,1)$, for all $N\in\mathbb{N}$, be such that $\sum_{N=1}^{\infty}\alpha_N < \infty$ while $\lim_{N\to\infty}\epsilon_{\mathrm{TV}}(\alpha_N) = 0$. Furthermore, assume that $\hat{J}_{\mathrm{DRO}}$ and $\hat{x}_{\mathrm{DRO}}$ denote the optimal value and an optimizer of the distributionally-robust optimization problem (7), where the ambiguity set is $\hat{\mathcal{P}}_N = \mathcal{B}_{\mathrm{TV},\varepsilon_{\mathrm{TV}}(\alpha)}(\hat{\mathbb{P}}'_N)$. Then, under Assumptions 3.1,

- If there exists $L \ge 0$ such that $|h(x,\xi)| \le L$ for all $(x,\xi)\in\mathbb{X}\times\Xi$, then $\lim_{N\to\infty}\hat{J}_{\mathrm{DRO}} \overset{\text{a.e.}}{=} J^\star$.
- If there exists $L \ge 0$ such that $|h(x,\xi)| \le L$ for all $(x,\xi)\in\mathbb{X}\times\Xi$, $\mathbb{X}$ is closed, and $h(x,\xi)$ is lower semi-continuous in $x$ for every $\xi\in\Xi$, then any accumulation point of $\{\hat{x}_{\mathrm{DRO}}\}_{N\in\mathbb{N}}$ is almost surely an optimal solution for (1).

*Proof:* The proof of this theorem is inspired by [2, Theorem 3.6]. Several aspects are however changed to accommodate noisy measurements, which was not considered in that paper. The detailed proof is moved to an online report [29] due to space constraints. ■

**Remark 3.2 (Finite-Sample Convergence)** *Under Assumption 3.1, the proof of Theorem 3.3 demonstrates that $\hat{J}_{\mathrm{DRO}} - J^\star = \mathcal{O}(\varepsilon_{\mathrm{TV}}(\alpha_N)) = \mathcal{O}(\ln^{1/2}(1/\alpha_N)N^{-1/2})$ when the ambiguity set is determined by the total variation distance, i.e., $\hat{\mathcal{P}}_N := \mathcal{B}_{\mathrm{TV},\varepsilon_{\mathrm{TV}}(\alpha)}(\hat{\mathbb{P}}'_N)$. Therefore, for fixed $\alpha_N = \alpha$, $\hat{J}_{\mathrm{DRO}} - J^\star = \mathcal{O}(N^{-1/2})$. The dependency on $N$ seems to be order optimal, i.e., there seems to exist cost functions and distributions for which $\hat{J}_{\mathrm{DRO}} - J^\star = \Omega(N^{-1/2})$ even when the samples are not noisy; see, e.g., [30, Proposition 1] for continuous distributions. An interesting direction for future research remains to prove the lower bound $\hat{J}_{\mathrm{DRO}} - J^\star = \Omega(N^{-1/2})$ for the exact problem formulation in this paper, i.e., noisy samples and discrete random variables.*

**Remark 3.3 (Deconvolution)** *The importance of Assumption 3.1 remains to be fully investigated. In our proofs, this assumption is used to show that $\lim_{N\to\infty}\mathcal{B}_{\mathrm{TV},\varepsilon_{\mathrm{TV}}(\alpha_{\mathrm{N}})}(\hat{\mathbb{P}}'_N) = \{\mathbb{P}\}$. Therefore, under appropriate conditions that ensure the uniqueness of the deconvolution, i.e., conditions under which it is guaranteed that $\{\mathbb{Q}\in\Delta(\Xi) : \mathbb{O}\star\mathbb{Q} = \mathbb{P}'\} = \{\mathbb{P}\}$, the uniform diagonal dominance in Assumption 3.1 may not be necessary to ensure consistency. Asserting appropriate conditions and proving this statement remains an important direction for future research.*

## IV. Worst-Case Distributions

Motivated by the results of the previous section, we examine a generic worst-case expectation problem:

$$\sup_{\mathbb{Q}\in\mathcal{B}_{\mathrm{TV},\varepsilon}(\hat{\mathbb{P}}'_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]. \tag{10}$$

For instance, $\ell(\xi) = h(x,\xi)$ for a fixed $x\in\mathbb{X}$.

**Theorem 4.1:** The worst-case expectation problem in (10) equals

$$\left\{\begin{array}{ll} \displaystyle\inf_{\substack{(\lambda(\xi'))_{\xi'\in\Xi'},\\ (\mu(\xi'))_{\xi'\in\Xi'},\\ r,t}} & \displaystyle r + 2\varepsilon t + \sum_{\xi'\in\Xi}(\mu(\xi') - \lambda(\xi'))\hat{\mathbb{P}}'_N(\xi'),\\ \text{s.t.} & \displaystyle \ell(\xi) + \sum_{\xi'\in\Xi}(\lambda(\xi') - \mu(\xi'))\mathbb{O}(\xi'|\xi) \le r,\\ & \lambda(\xi') + \mu(\xi') \le t,\\ & \lambda(\xi') \ge 0, \forall\xi'\in\Xi',\\ & \mu(\xi') \ge 0, \forall\xi'\in\Xi'. \end{array}\right.$$

*Proof:* The proof is moved to an online report [29] due to space constraints. ■

Now, we are ready to leverage the result of Theorem 4.1 to compute the solution to the distributionally-robust optimization problem in (7) with $\hat{\mathcal{P}}_N = \mathcal{B}_{\mathrm{TV},\varepsilon}(\hat{\mathbb{P}}'_N)$. In the next corollary, a computationally-friendly convex reformulation for this problem is provided.

**Corollary 4.1:** The worst-case expectation problem in (7) with $\hat{\mathcal{P}}_N = \mathcal{B}_{\mathrm{TV},\varepsilon}(\hat{\mathbb{P}}'_N)$ equals

$$\left\{\begin{array}{ll} \displaystyle\inf_{\substack{(\lambda(\xi'))_{\xi'\in\Xi'},\\ (\mu(\xi'))_{\xi'\in\Xi'},\\ r,t,x\in\mathbb{X}}} & \displaystyle r + 2\varepsilon t + \sum_{\xi'\in\Xi}(\mu(\xi') - \lambda(\xi'))\hat{\mathbb{P}}'_N(\xi'),\\ \text{s.t.} & \displaystyle h(x,\xi) + \sum_{\xi'\in\Xi}(\lambda(\xi') - \mu(\xi'))\mathbb{O}(\xi'|\xi) \le r,\\ & \qquad\qquad\qquad\qquad \forall x\in\mathbb{X}\\ & \lambda(\xi') + \mu(\xi') \le t,\\ & \lambda(\xi') \ge 0, \forall\xi'\in\Xi',\\ & \mu(\xi') \ge 0, \forall\xi'\in\Xi'. \end{array}\right.$$

**Remark 4.1:** Note that (7) involves two nested optimization problems (one maximization and one minimization)

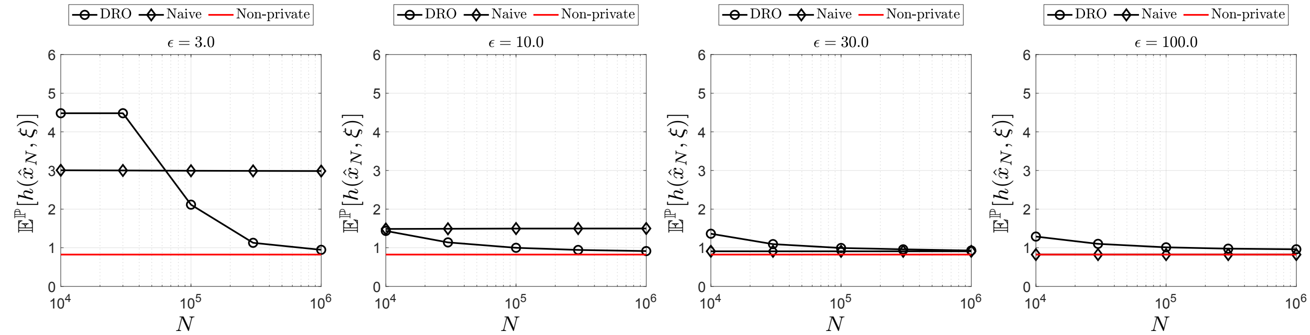

Fig. 1. Out-of-sample performance $\mathbb{E}^{\mathbb{P}}[h(\hat{x}_N,\xi)]$ for linear regression with noise-less non-private data (—), naïve linear regression with noisy data (—◇—), and distributionally-robust linear regression (—○—) versus number of data points $N$.

while the equivalent problem in Corollary 4.1 contains only a single minimization. This can significantly reduce the complexity of solving the problem. Furthermore, if $h(\cdot,\xi):\mathbb{X}\rightarrow\mathbb{R}$ is quasi-convex, the overall optimization problem in Corollary 4.1 is convex because the cost function and the constraints are convex in all decision variables. This problem can be then be solved using the interior point method, which has proved successful in solving generic nonlinear convex problems. In the following section, we solve this optimization problem using SeDuMi [31]. Although the proof of Corollary 4.1 may not require convexity of $h$, solving the optimization problem for non-convex $h$ can be numerically difficult as the constraint set can become the union of disjointed sets.

## V. Numerical Example

In this section, we demonstrate the capabilities of our results on distributionally-robust optimization with noisy data in the context of privacy-preserving linear regression. We use a dataset containing information regarding nearly 2,260,000 loans made on a peer-to-peer lending platform, called the Lending Club, which is available on Kaggle [32]. The dataset contains loan attributes, such as total loan size and interest rates of the loans per annum, and borrower information, such as number of credit lines, state of residence, and age. In our numerical example, we aim to learn a linear regression model for estimating interest rates of the loans based on features of loan size and credit rating. Since our results are for discrete random variables, we discretize the credit rating (mapping scores of 650 to 850 with brackets of 50 to $\{1,\dots,5\}$), the loan amount (mapping \$0 to \$40,000 with brackets of \$10,000 to $\{1,\dots,5\}$), and the interest rate (mapping 5% to 35% with increments of 5% to $\{1,\dots,7\}$). Let $\xi$ be a vector whose entries are, respectively, the discretized credit rating, the discretized loan amount, and the discretized interest rates. To train the linear regression model, we aim to solve (1) with $h(x,\xi)=(\xi_3-[\xi_1\;\xi_2\;1]x)^2$. In what follows, however, we assume that we do not have access to the exact measurements of $\xi$. We are in fact supplied with differentially-private perturbed measurements $\xi'$. The following definition and the subsequent theorem make this more clear. Before stating these results, we would like to define the notation $\mathrm{diam}(\Xi)=\max_{\xi,\bar{\xi}}\|\xi-\bar{\xi}\|$.

**Definition 5.1 (Local Differential Privacy)** *Conditional probability $\mathbb{O}(\xi'|\xi)$ is $\epsilon$-differentially-private if $\mathbb{O}[\xi'\in\mathcal{A}|\xi]\leq\exp(\epsilon)\mathbb{O}[\xi'\in\mathcal{A}|\bar{\xi}]$, for all $\mathcal{A}\subseteq\Xi'$ and all $\xi,\bar{\xi}\in\Xi$.*

**Proposition 5.1:** Assume $\Xi'=\Xi$. The following conditional probability guarantees $\epsilon$-local differential privacy:

$$\mathbb{O}(\xi'|\xi)=\exp\left(\frac{-\epsilon\|\xi-\xi'\|}{2\mathrm{diam}(\Xi)}\right)\Big/\sum_{\xi''\in\Xi}\exp\left(\frac{-\epsilon\|\xi-\xi''\|}{2\mathrm{diam}(\Xi)}\right).$$

*Proof:* The proof is similar to that of exponential mechanisms for differential privacy in [33, §3.4]. Detailed derivations are removed due to space constraints and can be found in an online report [29]. ■

For the sake of comparison, we consider three linear regression models. The baseline for the best achievable performance is given by the optimal linear regression model using noise-less data. Note that, by construction, no linear regression model can beat the baseline. However, according to Theorem 3.3, the performance of the distributionally-robust linear regression converges to the baseline, i.e., the proposed distributionally-robust regression model is asymptotically consistent and optimal, if $\mathbb{O}$ in Proposition 5.1 is uniformly diagonally dominant. Due to the special form of the conditional probability $\mathbb{O}$ in Proposition 5.1,

$$\max_{\xi\neq\xi'}\mathbb{O}(\xi'|\xi)\propto\exp\left(\frac{-\epsilon\min\limits_{\xi\neq\xi'}\|\xi-\xi'\|}{2\mathrm{diam}(\Xi)}\right)=\exp\left(\frac{-\epsilon}{2\mathrm{diam}(\Xi)}\right),$$

where $\propto$ denotes equality up to re-scaling by a constant or proportionality (c.f., Proposition 5.1) and $\min_{\xi\neq\xi'}\|\xi-\xi'\|=1$ in this example. Therefore, Assumption 3.1 is satisfied if

$$\frac{\mathbb{O}(\xi|\xi)}{\max_{\xi\neq\xi'}\mathbb{O}(\xi'|\xi)}=\exp\left(\frac{\epsilon}{2\mathrm{diam}(\Xi)}\right)>|\Xi|,$$

or equivalently if $\epsilon>2\mathrm{diam}(\Xi)\log(|\Xi|)\approx 64.17$. Note that, although we consider privacy budgets $\epsilon$ that are below this bound and thus the conditional density of the privacy-preserving noise is not uniformly diagonally dominant, we can still observe asymptotic consistency numerically. We compare the baseline performance with the performance of two linear regression models in the noisy regime. One of the regression models is naïvely constructed from the noisy data without any processing (i.e., as if the data was noiseless). The other model is the distributionally-robust regression model that can be extracted from Corollary 4.1. This optimization problem is modelled using CVX [34] and solved using SeDuMi [31]. The codes for conducting the experiments in this section can be downloaded from GitHub[3].

Figure 1 illustrates the out-of-sample performance $\mathbb{E}^{\mathbb{P}}[h(\hat{x}_N, \xi)]$ for linear regression with noise-less non-private data, naïve linear regression with noisy data, and distributionally-robust linear regression with noisy data versus the number of data points $N$. For $\epsilon = 3.0$, the out-of-sample performance of the distributionally-robust regression model improves rapidly and surpass the naïve regression model. For $\epsilon = 10.0$, the out-of-sample performance of the distributionally-robust regression model is superior to the naïve regression model for the entire range. Finally, for $\epsilon = 30.0$ $\epsilon = 100.0$, due to the small magnitude of the noise, the out-of-sample performance of the naïve regression model and linear regression with noise-less non-private data are almost identical, and the out-of-sample performance of the distributionally-robust regression model approaches that of the noise-less regression model rapidly. In all cases, as the number of data points $N$ grows, the out-of-sample performance of the distributionally-robust regression model approaches that of the noise-less regression model.

## VI. Conclusions and Future Work

We considered stochastic programs where the uncertainty distribution must be inferred from noisy data samples. We showed that the stochastic programs can be approximated with distributionally-robust optimizations that minimize the worst-case expected cost over an ambiguity set of distributions that are sufficiently compatible with the observed data. Future work can focus on continuous random variables.

## References

[1] R. O. Michaud, "The Markowitz optimization enigma: Is 'optimized' optimal?," *Financial analysts journal*, vol. 45, no. 1, pp. 31–42, 1989.

[2] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[3] H. Scarf, "A min-max solution of an inventory problem," in *Studies in the Mathematical Theory of Inventory and Production* (K. J. Arrow, S. Karlin, and H. E. Scarf, eds.), Stanford University Press, 1958.

[4] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.

[5] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.

[6] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Mathematical Programming*, vol. 107, no. 1, pp. 37–61, 2006.

[7] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, pp. 1695–1724, 2013.

[8] G. Pflug and D. Wozabal, "Ambiguity in portfolio selection," *Quantitative Finance*, vol. 7, no. 4, pp. 435–442, 2007.

[9] D. Wozabal, "A framework for optimization under ambiguity," *Annals of Operations Research*, vol. 193, no. 1, pp. 21–47, 2012.

[10] D. Bartl, S. Drapeau, J. Obłój, and J. Wiesel, "Sensitivity analysis of Wasserstein distributionally robust optimization problems," *Proceedings of the Royal Society A*, vol. 477, no. 2256, p. 20210176, 2021.

[11] B. P. Van Parys, "Efficient data-driven optimization with noisy data," *arXiv preprint arXiv:2102.04363*, 2021.

[12] F. Farokhi, "Distributionally-robust machine learning using locally differentially-private data," *Optimization Letters*, vol. 16, no. 4, pp. 1167–1179, 2022.

[13] J. M. Abowd, "The US Census Bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, 2018.

[14] W. Guo, M. Yin, Y. Wang, and M. Jordan, "Partial identification with noisy covariates: A robust optimization approach," in *Conference on Causal Learning and Reasoning*, pp. 318–335, 2022.

[15] R. J. Carroll and P. Hall, "Optimal rates of convergence for deconvolving a density," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1184–1186, 1988.

[16] F. Farokhi, "Deconvoluting kernel density estimation and regression for locally differentially private data," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[17] C. Dwork, N. Kohli, and D. Mulligan, "Differential privacy in practice: Expose your epsilons!," *Journal of Privacy and Confidentiality*, vol. 9, no. 2, 2019.

[18] A. Delaigle, P. Hall, and A. Meister, "On deconvolution with repeated measurements," *The Annals of Statistics*, vol. 36, no. 2, pp. 665 – 685, 2008.

[19] A. Bennouna and B. Van Parys, "Holistic robust data-driven decisions," *arXiv preprint arXiv:2207.09560*, 2022.

[20] C. Wang, X. Dong, F. Zhou, L. Cao, and C.-H. Chi, "Coupled attribute similarity learning on categorical data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 781–797, 2014.

[21] K. Hoshino and K. Sakurama, "Probability distribution control of finite-state markov chains with wasserstein costs and application to operation of car-sharing services," in *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 6634–6639, IEEE, 2021.

[22] F. Farokhi, "Noiseless privacy: Definition, guarantees, and applications," *IEEE Transactions on Big Data*, vol. 9, no. 1, pp. 51–62, 2023.

[23] C. Wu, X. Zhao, W. Xia, J. Liu, and T. Başar, "L2-gain analysis for dynamic event-triggered networked control systems with packet losses and quantization," *Automatica*, vol. 129, p. 109587, 2021.

[24] C. L. Canonne, "A short note on learning discrete distributions," *arXiv preprint arXiv:2002.11457*, 2020.

[25] C. Villani, *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2008.

[26] M. Akahira and K. Takeuchi, *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. Lecture Notes in Statistics, Springer New York, 2012.

[27] N. Johnston, *Advanced Linear and Matrix Algebra*. Springer International Publishing, 2021.

[28] S. Jiang and S. Tokdar, "Consistent Bayesian community detection," *arXiv preprint arXiv:2101.06531*, 2021.

[29] F. Farokhi, "Distributionally-robust optimization with noisy data for discrete uncertainties using total variation distance," *arXiv preprint arXiv:2302.07454 [math.OC] https://arxiv.org/abs/2302.07454*, 2023.

[30] A. Wibisono, M. J. Wainwright, M. Jordan, and J. C. Duchi, "Finite sample convergence rates of zero-order stochastic optimization methods," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[31] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization methods and software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[32] Kaggle, "All Lending Club loan data: 2007 through current Lending Club accepted and rejected loan data." https://www.kaggle.com/datasets/wordsforthewise/lending-club.

[33] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, p. 211–407, 2014.

[34] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1." http://cvxr.com/cvx, Mar. 2014.

[3] https://github.com/farhadfarokhi/NoisyDRO