Distributionally Robust Optimization and Robust Statistics

Jose Blanchet[*1], Jiajin Li[†1], Sirui Lin[‡1], Xuhui Zhang[§1]

[1]Department of Management Science & Engineering, Stanford University

January 29, 2024

Abstract

We review distributionally robust optimization (DRO), a principled approach for constructing statistical estimators that hedge against the impact of deviations in the expected loss between the training and deployment environments. Many well-known estimators in statistics and machine learning (e.g. AdaBoost, LASSO, ridge regression, dropout training, etc.) are distributionally robust in a precise sense. We hope that by discussing the DRO interpretation of well-known estimators, statisticians who may not be too familiar with DRO may find a way to access the DRO literature through the bridge between classical results and their DRO equivalent formulation. On the other hand, the topic of robustness in statistics has a rich tradition associated with removing the impact of contamination. Thus, another objective of this paper is to clarify the difference between DRO and classical statistical robustness. As we will see, these are two fundamentally different philosophies leading to completely different types of estimators. In DRO, the statistician hedges against an environment shift that occurs *after* the decision is made; thus DRO estimators tend to be pessimistic in an adversarial setting, leading to a min-max type formulation. In classical robust statistics, the statistician seeks to correct contamination that occurred *before* a decision is made; thus robust statistical estimators tend to be optimistic leading to a min-min type formulation.

Keywords: Distributionally Robust Optimization, Robust Statistics

Distributionally Robust Optimization (DRO)

The task of DRO is to estimate a parameter that will perform well on an unseen population from samples generated from a given population, which may or may not be similar to the unseen population.

Robust Statistics

The task of robust statistics is to estimate a parameter that depends on a given population from samples that may be contaminated with outliers or errors.

---

[*]jose.blanchet@stanford.edu

[†]jiajinli@stanford.edu

[‡]siruilin@stanford.edu

[§]xzhang98@stanford.edu

Contents

1 Introduction

	Contaminated data-generating distribution		Data-driven model		Out-of-sample environment

$$\mathbb{P}_\star \quad \longrightarrow \quad \overline{\mathbb{P}} \quad \longrightarrow \quad \hat{\mathbb{P}}_n \quad \longrightarrow \quad \widetilde{\mathbb{P}}$$

$$\widetilde{\mathbb{P}} = \mathbb{P}_\star = \overline{\mathbb{P}} : \text{ Conventional Assumption}$$

$$\widetilde{\mathbb{P}} \neq \hat{\mathbb{P}}_n : \text{ Overfitting}$$

$$\widetilde{\mathbb{P}} \neq \mathbb{P}_\star = \overline{\mathbb{P}} : \text{ Distributional Shift}$$

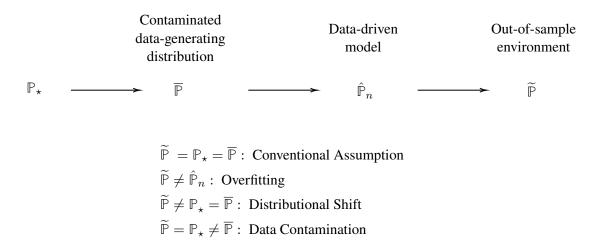$$\widetilde{\mathbb{P}} = \mathbb{P}_\star \neq \overline{\mathbb{P}} : \text{ Data Contamination}$$

Figure 1: Data-Driven Decision Making Cycle

In the conventional data-driven decision-making cycle shown in Figure 1, we typically observe n i.i.d. samples generated from the unknown data-generating distribution $\mathbb{P}_\star$, and make a decision (e.g., parameter estimation) based on a model, $\hat{\mathbb{P}}_n$, built from these samples (such a model could be parametric or non-parametric). These decisions are then deployed in the out-of-sample environment $\tilde{\mathbb{P}}$, which may or may not follow the distribution $\mathbb{P}_\star$. During this cycle, several factors can contribute to suboptimal decision-making.

(i) *Potential Overfitting* ($\widetilde{\mathbb{P}} \neq \hat{\mathbb{P}}_n$): when the sample size n is not large enough, the model learned from the samples may achieve good in-sample performance but fail to generalize its predictive power to the out-of-sample environment, which is commonly termed as overfitting in statistical and machine learning tasks. Specifically, when referring to the out-of-sample environment in this scenario, people usually focus on the data-generating distribution $\mathbb{P}_\star$.

(ii) *Distributional shift* ($\widetilde{\mathbb{P}} \neq \mathbb{P}_\star = \overline{\mathbb{P}}$): in many real-world scenarios, the out-of-sample environment $\tilde{\mathbb{P}}$ may deviate from the data-generating distribution $\mathbb{P}_\star$. This discrepancy, known as distributional shift, can arise in various circumstances. For instance, in adversarial deployment settings, malicious actors can intentionally manipulate the data distribution to undermine the performance of trained models. Additionally, in transfer learning settings, models may be expected to effectively generalize to target datasets that differ slightly from the source datasets used for training.

(iii) *Data Contamination* ($\widetilde{\mathbb{P}} = \mathbb{P}_\star \neq \overline{\mathbb{P}}$): many real data sets contain outliers or have measurement error throughout the steps of data generation and collection. Thus, the observed samples are actually generated by a contaminated distribution $\overline{\mathbb{P}}$, posing challenges to the inference of the underlying uncontaminated distribution $\mathbb{P}_\star$.

The first two cases, (i) and (ii), arise from errors that occur in the post-decision stage, where the trained model or decision rule is applied to the out-of-sample data. The third case, (iii), differs from these two in that the error occurs in the pre-decision stage, during the steps of data generation and collection. Now we discuss two principled approaches to deal with cases (i)-(iii).

Distributionally robust optimization (DRO) is a data-driven decision-making framework that is designed to minimize the potential discrepancies between the in-sample expected loss and the out-of-sample expected loss. In particular, the goal of DRO is to address cases (i) and (ii). DRO takes an adversarial formulation aiming to minimize the expected loss of a model (e.g. the squared loss in linear regression) incurred by a

parameter selection (e.g. the regression parameter), uniformly across a set of possible data distributions. The set of possible data distributions is characterized by a family of models that describe deviations from the training data distribution. To establish a concrete mathematical formulation, we begin by examining a generic stochastic optimization problem. Here, we assume that ξ is a random vector in space Ξ (e.g., $\mathbb{R}^d$) that follows the distribution $\mathbb{P}_\star$. The set of feasible model parameters is denoted Θ (assumed to be finite-dimensional to simplify). Given a realization ξ and a model parameter $\theta \in \Theta$ the corresponding loss is $\ell(\theta, \xi)$. A standard expected loss minimization decision rule is obtained by solving

$$\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta, \xi)] = \int_\Xi \ell(\theta, \xi) \, d\mathbb{P}_\star(\xi). \tag{1.1}$$

Since $\mathbb{P}_\star$ is generally unknown, to approximate the objective function in (1.1), we often i.i.d. samples $\xi_1, \ldots, \xi_n$ each following distribution $\mathbb{P}_\star$ and consider the empirical risk minimization counterpart,

$$\min_{\theta \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)] = \frac{1}{n} \sum_{i=1}^n \ell(\theta, \xi_i), \tag{1.2}$$

where $\hat{\mathbb{P}}_n$ denotes the empirical measure $\frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ and δ_ξ is the Dirac measure centered at ξ. To guarantee a good performance in the sense of finding a bound on the optimal expected population loss with high probability when deployed out-of-sample, the DRO framework introduces an uncertainty set $\mathcal{B}(\hat{\mathbb{P}}_n)$ to capture variations between the in-sample distribution $\hat{\mathbb{P}}_n$ and the out-of-sample distribution (including environment shifts). Then, the DRO formulation minimizes the worst-case loss within this uncertainty set, i.e.,

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)], \tag{1.3}$$

where $\mathbb{E}_Q$ denotes the expectation operator assuming that ξ follows distribution $\mathbb{Q}$. In Section 2, we provide a comprehensive discussion of DRO and its effectiveness in addressing cases (i) and (ii). We examine its theoretical foundations and present evidence from diverse applications in statistical and machine learning communities. In particular, we discuss how DRO recovers a wide range of successful estimators and regularization methods, which are often applied in statistical inference. We also discuss novel uses of DRO strategies that combine historical data with additional domain-knowledge information.

Case (iii) is thoroughly examined in the field of robust statistics, which seeks to address the challenges in inference posed by this case. Our goal here is to discuss a novel perspective that motivates the types of estimators that are obtained in robust statistics and explain their qualitative differences relative to DRO-based estimators. Traditional robust statistics techniques aim to estimate a parameter that depends on an unknown population from its contaminated samples. Formally, given samples $\xi_1, \ldots, \xi_n$ in a metric space Ξ that are i.i.d. generated from some distribution $\bar{\mathbb{P}}$ that may be a contaminated version of $\mathbb{P}_\star$, we aim to learn a mapping $\hat{\theta} : \mathcal{P}(\Xi) \to \Theta$ that maps the empirical distribution $\hat{\mathbb{P}}_n$ of the samples to an estimator $\hat{\theta}(\hat{\mathbb{P}}_n)$ of the underlying parameter under $\mathbb{P}_\star$. The contamination model can be formally represented as $\bar{\mathbb{P}} \in \mathcal{A}(\mathbb{P}_\star)$, where $\mathcal{A}(\mathbb{P}_\star)$ is a set of possible contaminated data generating distributions that contains $\mathbb{P}_\star$. Therefore, for a loss function $\ell : \Theta \times \Xi \to \mathbb{R}$, the out-of-sample risk given the robust learning procedure $\hat{\theta}(\hat{\mathbb{P}}_n)$ is thus

$$\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)].$$

The "adversary" that injects contamination is attempting to maximize this out-of-sample risk, while the decision-maker's goal is to minimize it. It is thus natural to introduce a max-min game to formalize this process as

$$\sup_{\bar{\mathbb{P}} \in \mathcal{A}(\mathbb{P}_\star)} \inf_{\hat{\theta}(\cdot) \in \Psi} \mathbb{E}_{\bar{\mathbb{P}}} \left[\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)] \right].$$

4

Here, we use $\bar{\mathbb{P}}$ in short for the product measure $\bar{\mathbb{P}}^{\otimes n}$. Ψ is a class of policies that the decision-maker can employ, where each policy is a mapping $\hat{\theta} : \mathcal{P}(\Xi) \to \Theta$. Note that in this formulation, the inner randomness comes from ξ that follows the distribution $\mathbb{P}_\star$, and the outer randomness lies in the empirical measure $\hat{\mathbb{P}}_n$, where each sample i.i.d. follows the distribution $\bar{\mathbb{P}}$. Given that the statistician knows that the data has been contaminated, a natural policy class to consider involves rectifying/correcting the contamination, and, for this, we introduce a rectification set $\mathcal{R}(\hat{\mathbb{P}}_n)$ which models a set of possible pre-contamination distributions based on the knowledge of the empirical measure $\hat{\mathbb{P}}_n$. The rectification/decontamination approach naturally induces the following min-min strategy to address the error in case (iii), thus

$$\hat{\theta}(\hat{\mathbb{P}}_n) = \arg\inf_{\theta \in \Theta} \min_{\mathbb{Q} \in \mathcal{R}(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}\left[\ell(\theta, \xi)\right].$$

In Section 3, we furnish a detailed overview of robust statistics, making explicit connections with the min-min approach that we present here while drawing comparisons with the DRO framework.

It is important to highlight the distinctive features of the DRO and the robust statistics formulations induced by varying the order of decision-making relative to the distributional mismatch in the three cases. In both cases (i) and (ii), the distributional mismatch occurs in the post-decision stage. Consequently, the DRO approach employs a min-max game strategy to control the worst-case loss over potential post-decision distributional shifts. In contrast, for case (iii), the robust estimator acts after the pre-decision distributional contamination materializes. Thus the approach of robust statistics can be motivated as being closer to a max-min game against nature. As a consequence, in robust statistics, the adversary moves first, and therefore the statistician can be more optimistic that they can rectify the contamination applied by the nature thus motivating the min-min strategy suggested above.

As we present our discussion, we will often summarize key results in the form of theorems which are stated in a summarized form for ease of exposition. We refer the reader to the references for precise assumptions and proofs.

The DRO literature is rapidly growing, so it is virtually impossible to cover every new application in this review. However, in the conclusion section, we will briefly discuss trending topics on DRO in areas such as dynamic decision-making problems (Xu and Mannor, 2010; Osogami, 2012; Lim et al., 2013; Zhou et al., 2021; Backhoff et al., 2022; Si et al., 2023; Wang et al., 2023) and causal inference (Bertsimas et al., 2022; Rothenhäusler and Bühlmann, 2023; Bennett et al., 2023; Duchi et al., 2023).

Notation. To summarize the notation that we use, it is useful to keep in mind Figure 1. We use $O(\delta)$ to denote a quantity that is bounded by $\delta \times$ some constant as δ goes to zero; we use $\mathbb{1}_A$ to denote the indicator function of the set A; we use δ_ξ to denote the Dirac measure at ξ and let $\hat{\mathbb{P}}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$ be the empirical measure constructed from observed samples $\{\xi_1, \ldots, \xi_n\}$; we use $\mathbb{P}_\star$ to denote the underlying uncontaminated distribution, $\bar{\mathbb{P}}$ to denote the (possibly contaminated) data-generating distribution, $\widetilde{\mathbb{P}}$ to denote the out-of-sample distribution; in Section 3, we also denote by $\bar{\mathbb{P}}_n$ the contaminated version of $\hat{\mathbb{P}}_n$; $\mathbb{E}_\mathbb{P}$ is the expectation over the probability distribution $\mathbb{P}$; for a joint distribution π for (ξ, η), π_ξ denotes the marginal distribution of ξ; $\xrightarrow{d}$ denotes the convergence in distribution; $\mathcal{L}^2(\mathcal{D})$ denotes the L^2-integrable functions defined on domain $\mathcal{D} \subset \mathbb{R}^d$ under the Lebesgue measure on the d-dimensional Euclidean space; $\mathbb{R}_+$ denotes the space of non-negative real numbers; $\mathbb{D}$ denotes discrepancies between probability models; $\mathcal{B}$ denotes the uncertainty set of distributions; $\mathcal{R}$ denotes the rectification set of distributions; $\mathcal{A}$ denotes the class of possible contaminated data generating distributions; Ψ (resp., Φ) denotes the class of policies that the decision maker can employ in the setting of robust statistics (resp., DRO).

Table 1: Statistical tasks related to DRO. (OT is short for optimal transport, ϕ-div is short for ϕ-divergence, CR is short for confidence region, and MMD is short for maximum mean discrepancy.)

Statistical tasks	Uncertainty construction	Reference
Norm regularization	OT	Blanchet et al. (2019a); Li et al. (2022); Gao et al. (2022)
Variance regularization	ϕ-div	Lam (2016, 2018); Duchi et al. (2021)
Adaptive boosting	ϕ-div	Blanchet et al. (2019b)
Domain adaptation	OT/ϕ-div	Taskesen et al. (2021); Zhang et al. (2022a)
Group regularization	OT/ϕ-div	Blanchet and Kang (2017); Hu et al. (2018); Sagawa et al. (2020)
Bayesian estimation	OT/ϕ-div	Zhang et al. (2022b); Nguyen et al. (2023); Lotidis et al. (2023)
CR construction	OT/ϕ-div	Duchi et al. (2021); He and Lam (2021); Blanchet et al. (2022a)
Hypothesis testing	OT/ϕ-div	Gül and Zoubir (2017); Gao et al. (2018); Sun and Zou (2021)
Dropout regularization	Multiplication auxiliary	Blanchet et al. (2023a)

2 Distributionally Robust Optimization

The DRO framework provides a principled approach to understand and analyze various regularization methods from a probabilistic perspective. In this section, we will present a comprehensive overview of how the DRO problem (1.3) connects with some well-known regularization methods commonly used in statistical and machine learning tasks. Throughout this overview, we also discuss how the DRO framework helps extend existing methods for these tasks by modeling distributional uncertainty in an interpretable adversarial way. Following this discussion, we review some of the statistical guarantees obtained by DRO-based estimators and the tools for statistical inference that are induced by the DRO framework, including new statistical objects such as associated worst-case distributions often corresponding to a Nash equilibrium.

Note that, in this section, $\mathbb{P}_\star$ denotes the (uncontaminated) data-generating distribution because the error is assumed to occur in the post-decision stages.

2.1 DRO Formulations and Related Statistical Tasks

To be able to control the model's conservativeness, most of existing literature define the uncertainty set $\mathcal{B}$ as a neighborhood ball that contains the nominal distribution $\hat{\mathbb{P}}_n$ (e.g., empirical distribution). In this subsection, we review various probabilistic characterizations used in the DRO literature to construct the uncertainty set $\mathcal{B}$. Then, we connect the resulting DRO formulation with various statistical and machine learning tasks, which are summarized in Table 1.

A natural approach for modeling the distributional uncertainty set in DRO is given by moment constraints; in fact, this is one of the earlier approaches followed by Scarf (1958). For example, a distributionally robust moment constraint problem involving means and variances was introduced in Delage and Ye (2010) and it is

formulated via

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)],$$

where the set $\mathcal{B}$ is defined as follows. Specify δ_1, δ_2, non-negative constants and estimate nominal mean and covariance matrix $(\hat{\mu}, \hat{\Sigma})$ from the empirical distribution $\hat{\mathbb{P}}_n$. Then,

$$\mathcal{B} = \left\{ \mathbb{Q} : (\mathbb{E}_{\mathbb{Q}}[\xi] - \hat{\mu})^\top \hat{\Sigma}^{-1} (\mathbb{E}_{\mathbb{Q}}[\xi] - \hat{\mu}) \leq \delta_1, \ \mathbb{E}_{\mathbb{Q}}[(\xi - \hat{\mu})^\top (\xi - \hat{\mu})] \preceq \delta_2 \hat{\Sigma} \right\}.$$

This formulation has significant computational advantages discussed in Delage and Ye (2010) because it often can be formulated in terms of semi-definite programming. However, the problem with this formulation from a statistical standpoint is that $\mathcal{B}$ contains excessive distributions that are not in the local neighborhood of $\hat{\mathbb{P}}_n$; even when δ_1, δ_2 are close to zero. As n grows to infinity, under mild assumptions (certainly under i.i.d. assumptions) the nominal distribution $\hat{\mathbb{P}}_n$ converges weakly with probability one to the data-generating distribution $\mathbb{P}_\star$. Nevertheless, the uncertainty set contains distributions that may be far from $\mathbb{P}_\star$ potentially deteriorating the performance of the DRO-based estimator in the sense of being over-conservative (except in cases in which the optimal parameter choice only depends on means and variances).

Alternatively, most of the existing literature adopts uncertainty sets that are induced by probability metrics/discrepancies, that is,

$$\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \left\{ \mathbb{Q} : \mathbb{D}(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta \right\}.$$

This ball of radius $\delta \geq 0$ is defined with respect to a discrepancy measure $\mathbb{D}$ on the probability distribution space $\mathcal{P}(\Xi)$. In the remainder of this section, we will focus on typical uncertainty sets $\mathcal{B}_\delta(\hat{\mathbb{P}}_n)$ that are induced by probability metrics/discrepancies. This focus aims to serve our discussion on the connection of DRO with statistical and machine learning tasks. For a more in-depth discussion of general DRO formulations, interested readers are referred to Rahimian and Mehrotra (2022).

2.1.1 ϕ-Divergence-Based DRO

The ϕ-divergence approach corresponds to methods that penalize deviations from a baseline model in terms of the likelihood ratio; see, for example, Csiszár (1975); Ruszczyński and Shapiro (2006); Rockafellar (2023).

Definition 2.1 (ϕ-divergence)**.** Assume that $\phi : \mathbb{R}_+ \to (-\infty, +\infty]$ is a convex function with $\phi(0) = \lim_{t \to 0^+} \phi(t)$, then the ϕ-divergence between $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$ is

$$\mathbb{D}_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) = \begin{cases} \int_\Xi \phi\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\hat{\mathbb{P}}_n}\right) \mathrm{d}\hat{\mathbb{P}}_n(\xi) & \mathbb{Q} \ll \hat{\mathbb{P}}_n \\ +\infty & \text{otherwise,} \end{cases}$$

where $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\hat{\mathbb{P}}_n}$ is the likelihood ratio between $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$, and $\mathbb{Q} \ll \hat{\mathbb{P}}_n$ indicates that $\mathbb{Q}$ is absolutely continuous with respect to $\hat{\mathbb{P}}_n$. $\qquad\qquad\square$

In the context of data-driven estimation using empirical measures, the approach is closely related to the empirical likelihood (Owen, 2001) method. In the data-driven setting, the works of Hu and Hong (2013); Lam (2016, 2018); Duchi and Namkoong (2018, 2021) provide example showing the utilization of ϕ-divergence to define the uncertainty set in various statistical tasks.

The intuition using ϕ-divergence is that the adversary can re-weight the relative importance of each sample with a budget constraint. So, the adversary systematically explores how re-weighting can potentially impact the performance of an estimator as measured by a given expected loss.

In Duchi and Namkoong (2021), the authors argue that this choice of uncertainty set (with a well-chosen ϕ) can be used to hedge against the potentially low performance of statistical loss in minority subpopulations. Intuitively, if a minority population is severely affected by a decision choice, the adversary will exploit this by increasing the importance of this minority population, thus encouraging the decision maker to make a more equitable decision rule.

When the function $\phi(\cdot) \geq 0$ is twice differentiable and locally strongly convex around 1 (in particular, $\phi(1) = 0, \phi'(1) = 0, \phi''(1) > 0$), then ϕ-divergence-based DRO approach is asymptotically equivalent to variance regularization. This equivalence is formally established by the following theorem:

Theorem 2.1 (Variance regularization (Lam, 2016, Theorem 3.1), (Duchi and Namkoong, 2018, Theorem 1)). Suppose that ϕ is twice differentiable around 1, and for simplicity assume that loss function ℓ is bounded, then we have

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)] = \min_{\theta \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)] + \sqrt{\frac{2\delta}{\phi''(1)} \mathrm{Var}_{\hat{\mathbb{P}}_n}(\ell(\theta, \xi))} + O(\delta),$$

where $\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{Q} : \mathbb{D}_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta\}$, $\mathrm{Var}_{\hat{\mathbb{P}}_n}(\ell(\theta, \xi)) = \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)^2] - (\mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)])^2$ is the empirical variance of $\ell(\theta, \xi)$ under the empirical distribution $\hat{\mathbb{P}}_n$. $\qquad\square$

This result can be intuitively expected with a back-of-the-envelope calculation as follows. Suppose that the center of the distributional uncertainty set is $\mathbb{P}$ (this is more general than the empirical measure $\hat{\mathbb{P}}_n$ in the data-driven setting illustrated in the theorem) and assume that Z is the likelihood ratio of $\frac{d\mathbb{Q}}{d\mathbb{P}}$, such that $\mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)] = \mathbb{E}_{\mathbb{P}}[\ell(\theta, \xi) \cdot Z]$ and $\mathbb{D}_\phi(\mathbb{Q}, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\phi(Z)]$.

Under rather mild assumptions, the constraint $\mathbb{D}_\phi(\mathbb{Q}, \mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\phi(Z)] \leq \delta$ will be active when δ is small. Thus, with $\phi(1) = 0, \phi'(1) = 0, \phi''(1) > 0$, upon the Taylor expansion on $\phi(\cdot)$ around 1, the constraint will basically correspond to $\phi''(1)\mathbb{E}_{\mathbb{P}}[(Z-1)^2]/2 = \delta$. We thus write $Z = 1 + \delta^{1/2} \times \Delta$, where $\mathbb{E}_{\mathbb{P}}[|\Delta|^2] = 2\delta/\phi''(1)$, and $\mathbb{E}_{\mathbb{P}}[\Delta] = 0$ to preserve that Z is a likelihood ratio. We can ignore the positivity constraint on Z since δ will go to zero. As a result, the adversary is essentially maximizing $\mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)] = \mathbb{E}_{\mathbb{P}}[\ell(\theta, \xi) \cdot Z] = \mathbb{E}_{\mathbb{P}}[\ell(\theta, \xi)(1 + \delta^{1/2}\Delta)]$, subject to the indicated constraints on Δ. Then we get the optimal choice for Δ is of the form $\Delta = c\delta^{1/2} \times (\ell(\theta, \xi) - \mathbb{E}_{\mathbb{P}}[\ell(\theta, \xi)])$ due to centering implied by $\mathbb{E}_{\mathbb{P}}[\Delta] = 0$, where the normalizing constant c can be directly computed from $\mathbb{E}_{\mathbb{P}}[|\Delta|^2] = 2\delta/\phi''(1)$. Plugging in this choice of Δ, we obtain the aforementioned result.

The most important insight in this analysis is the form of the reweight employed by the adversary. In particular, the adversary increases the importance of samples for which the loss is large compared to the mean loss and decreases the importance of samples for which the loss is low compared to the mean loss.

Further, as we can see from the previous result, ϕ-divergence DRO implicitly considers the bias-variance trade-off, where we identify the bias with the empirical loss $\mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)]$ and the variance with the empirical variance $\mathrm{Var}_{\hat{\mathbb{P}}_n}(\ell(\theta, \xi))$. While directly minimizing the bias with variance regularization may result in a non-convex optimization problem, the DRO formulation keeps the convexity and thus enjoys the tractability (Duchi and Namkoong, 2018).

The view of adversarial reweighting sheds light on the adaptive reweighting strategy when we use the gradient descent algorithm to solve the DRO problem (1.3). At each step t, the gradient with respect to the

parameter θ at θ_t is computed as:

$$\frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_t} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)] = \sum_{i=1}^n \omega_i^\star \frac{\partial \ell}{\partial \theta}(\theta_t, \xi_i),$$

where

$$\boldsymbol{\omega}^\star = (\omega_1^\star, \ldots, \omega_n^\star)^\top = \arg\max_{\boldsymbol{\omega} \in C_\delta} \sum_{i=1}^n \omega_i \ell(\theta_t, \xi_i),$$

$$C_\delta = \left\{ \boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top : \frac{1}{n}\sum_{i=1}^n \phi(n\omega_i) \leq \delta, \ \sum_{i=1}^n \omega_i = 1, \ \omega_i \geq 0, \ \forall i \in [n] \right\}.$$

Here, we assume the optimal solution of $\omega^\star$ is unique and ℓ is differentiable in θ for simplicity, and in general, the subgradient can be computed instead of the aforementioned gradient. This type of adaptive reweighting is well-known in popular machine learning algorithms, like AdaBoost (Freund and Schapire, 1997). Actually, the AdaBoost algorithm can be recovered in the DRO framework (Blanchet et al., 2019b).

2.1.2   Optimal Transport-Based DRO

Another natural approach in modeling discrepancies in probability distributions is based on perturbing the actual random outcomes (as opposed to the likelihood of the outcome, as in the ϕ-divergence case). This approach can be made operational using the optimal transport discrepancies, the corresponding duality theory has been developed in, for example, studies such as Mohajerin Esfahani and Kuhn (2018); Blanchet et al. (2019a); Gao and Kleywegt (2023). This line of research has provided a probabilistic interpretation of various forms of regularization, more precisely, so-called norm regularization which is a widely adopted strategy to effectively address the issue of overfitting in machine learning models (Tibshirani, 1996; Ng, 2004). We now provide the definition of an optimal transport discrepancy.

Definition 2.2 (Optimal transport discrepancy). Assume that $c : \Xi \times \Xi \to [0, +\infty]$ is a nonnegative lower semi-continuous function, the optimal transport discrepancy between $\mathbb{Q}$ and $\mathbb{P}$ is defined as

$$\mathbb{D}_c(\mathbb{Q}, \mathbb{P}) = \min_\pi \left\{ \int_{\Xi \times \Xi} c(\xi, \xi') \mathrm{d}\pi(\xi, \xi') : \ \pi_\xi = \mathbb{Q}, \ \pi_{\xi'} = \mathbb{P} \right\},$$

where $\pi \in \mathcal{P}(\Xi \times \Xi)$ is a coupling of $\mathbb{Q}$ and $\mathbb{P}$, and π_ξ (resp. $\pi_{\xi'}$) denotes the marginal distribution of π on ξ (resp. ξ'). □

The optimal transport discrepancy has a rich tradition in a wide range of areas in engineering and applied mathematics; see, for example, Villani et al. (2009); Santambrogio (2015); Peyré et al. (2019). When the cost function is a metric, the optimal transport discrepancy recovers the Wasserstein distance. The choice of the cost function can be used to recover both the weak convergence topology (e.g., $c(\xi, \xi') = \|\xi - \xi'\|_2 / (1 + \|\xi - \xi'\|_2)$ when Ξ is a subspace of the Euclidean space) and the total variation topology (e.g., $c(\xi, \xi') = \mathbb{I}_{\xi \neq \xi'}$). So, statistically speaking, optimal transport provides a flexible approach to compare statistical distributions.

The intuition in the definition of the optimal transport discrepancy is that there is a source of mass, say, a pile of sand, and a target, say, a sinkhole, which are described by two distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively. The cost per unit of mass transferred from location ξ in the pile of sand to location ξ' at the sinkhole is given by $c(\xi, \xi')$. The objective function reflects the total cost incurred by transporting all of the mass and the
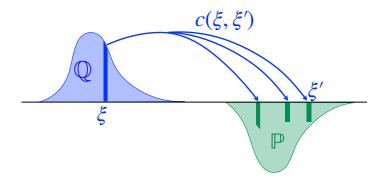
Figure 2: Illustration of optimal transport. The source of mass is denoted by $\mathbb{Q}$ and the target is denoted by $\mathbb{P}$. The optimal transport discrepancy between $\mathbb{Q}$ and $\mathbb{P}$ is the minimal total cost incurred by transporting all of the mass from $\mathbb{Q}$ (source) to $\mathbb{P}$ (target).

constraints reflect that the profile of the pile of sand is modeled by $\mathbb{P}$ and the profile of the sinkhole is modeled by $\mathbb{Q}$, this is illustrated in the Figure 2.

By employing optimal transport as a way to explore the impact of distributional uncertainty, we are intuitively modeling an environment in which the adversary is allowed to act as a transporter of mass, moving the points around subject to budget constraint in the expected cost of transporting mass to achieve maximum impact in the expected loss for the resulting mass configuration.

As we now illustrate, optimal transport discrepancy-based DRO is known to relate to norm regularization in many scenarios. For example, the next result shows that this approach can *exactly* recover the square-root LASSO estimator introduced in Belloni et al. (2011).

Theorem 2.2 (Squre-root LASSO (Blanchet et al., 2019a, Theorem 1))**.** We assume the random input takes the form $\xi_i = (x_i, y_i) \in \mathbb{R}^{d+1}$, the loss function is $\ell(\theta, \xi) = (y - \theta^\top x)^2$, and the cost function admits

$$c((x,y),(x',y')) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y', \\ +\infty & \text{otherwise.} \end{cases}$$

Then, we have

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)] = \min_{\theta \in \Theta} \left(\sqrt{\mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)]} + \sqrt{\delta}\|\theta\|_p \right)^2,$$

where $\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{Q} : \mathbb{D}_c(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta\}$ and $\frac{1}{p} + \frac{1}{q} = 1$. $\qquad\square$

The optimal transport discrepancy-based DRO extend beyond regression and encompass norm regularization for generalized linear classification models, such as logistic regression and support vector machines with hinge loss.

Theorem 2.3 (ℓ_p-norm regularization (Shafieezadeh-Abadeh et al., 2019, Theorem 4, Theorem 14))**.** We assume the random input takes the form $\xi_i = (x_i, y_i) \in \mathbb{R}^{d+1}$. The loss function is set to be: 1. $\ell(\theta, \xi) = L(y \cdot \theta^\top x)$, or 2. $\ell(\theta, \xi) = L(y - \theta^\top x)$, for a continuous function L with Lipschitz constant 1, i.e.,

$\sup_{x \neq y} \frac{|f(x)-f(y)|}{\|x-y\|_2} = 1$, and

$$c((x,y),(x',y')) = \begin{cases} \|x-x'\|_q & y = y' \\ +\infty & y \neq y'. \end{cases}$$

Then, we have

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)] = \min_{\theta \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta,\xi)] + \delta\|\theta\|_p,$$

where $\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{Q} : \mathbb{D}_c(\mathbb{Q},\hat{\mathbb{P}}_n) \leq \delta\}$ and $\frac{1}{p} + \frac{1}{q} = 1$. $\qquad\square$

Similar norm regularization on the model parameter θ can also be recovered in the group LASSO (Blanchet and Kang, 2017). Beyond recovering the well-established techniques in statistics and machine learning, the optimal transport-based DRO also provides a principled approach to potentially enhance the existing learning methods (see, e.g., Sinha et al. (2018)). For a general loss function, the optimal transport-based DRO asymptotically recovers the so-called variation regularization (Gao et al., 2022).

The asymptotic connection to variation regularization can be developed using an intuitive approach similar to the reasoning employed in the ϕ-divergence setting. It is useful to carry out this analysis because it uncovers the structure of the worst-case adversarial strategy and the appearance of the dual norms as regularization terms. We follow the strategy in the perturbation analysis in Bartl et al. (2021), which provides a rather complete study not only of the optimal adversarial perturbation but also of the optimal parameter selection (see also the analysis in Blanchet et al. (2022b)). Consider a general formulation in which the adversary maximizes $\mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)]$ over $\mathbb{Q}$ such that $\mathbb{D}_c(\mathbb{Q},\mathbb{P}) \leq \delta$. Assume, for example, that $c(\xi,\xi') = \|\xi - \xi'\|_q^2$ when Ξ is a subspace of the Euclidean space. The problem for the adversary is equivalent to $\max_\Delta \mathbb{E}_{\mathbb{P}}[\ell(\theta,\xi+\Delta)]$ over random variables Δ satisfying $\mathbb{E}_{\mathbb{P}}[\|\Delta\|_q^2] \leq \delta$. By applying a Taylor expansion on the objective function, we see that when δ is small, the adversary is effectively maximizing $\mathbb{E}_{\mathbb{P}}[\nabla\ell(\theta,\xi)\Delta]$, where ∇ is the gradient with respect to ξ (in linear models, the parameter will naturally appear by the chain rule thus leading to norm regularization). It is then direct from this observation that Δ will be chosen by the adversary "parallel" or "aligned" to $\nabla\ell(\theta,\xi)$ in the corresponding geometry induced by the cost function with the intent of maximizing the loss. This alignment condition is given the dual pair which achieves equality in Holder's inequality. This explains why when the cost is chosen based on the l_q norm, the regularization involves the dual l_p norm. Thus, the perturbation direction employed by the worst-case adversary is fully dictated by the norm dual vector of the loss's gradient with respect to the source of randomness. The size of the norm is fully dictated by the budget constraint $\mathbb{E}_{\mathbb{P}}[\|\Delta\|_q^2] \leq \delta$. For example, if the cost function $c(\xi,\xi')$ is locally quadratic around the diagonal, then $\|\Delta\|_q = O(\delta^{1/2})$. As an illustrative example, Blanchet et al. (2022b, Section 4.1.2) computes the worst-case adversarial distribution in the setting of logistic regression. Figure 3 shows the decision boundaries of binary classification based on the DRO solutions with increasing uncertainty budget δ, and the worst-case adversarial distribution after perturbing the empirical distribution. The perturbation trajectory of one sample point from each class is marked by $+$'s. We note that some sample points that are close to the decision boundary (see the markers $+$) will flip their sign of label after adversarial perturbation. As a concrete application, Shafieezadeh-Abadeh et al. (2023) proposes to generate adversarial examples from the worst-case adversarial distribution.

It is useful to contrast the variance regularization of ϕ-divergence-based DRO compared to the variation regularization which is defined as the norm of the gradient of the loss function with respect to the source of randomness.
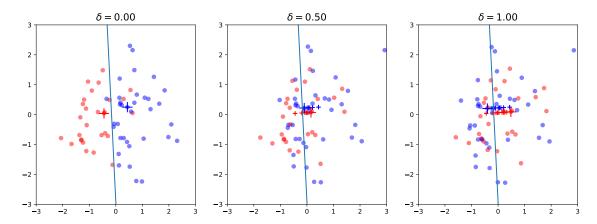
Figure 3: Decision boundary and worst-case distribution. One point from each class is selected, and we use a big + to mark its position. We also use a small + to mark its previous positions when δ is smaller to visualize its trajectory.

Not surprisingly, given the above analysis, the idea of transporting the sample to its neighborhood has direct connections to adversarial training methods (Goodfellow et al., 2015; Madry et al., 2017). This connection has been directly exposed in Sinha et al. (2018). Interested readers are referred to the tutorials Kuhn et al. (2019); Blanchet et al. (2021b) for extensive discussion on the optimal transport-based approach.

Some other versions of the optimal transport-based DRO are explored. For example, the martingale constraints are added to the optimal transport-based uncertainty set (Li et al., 2022; Lotidis et al., 2023) - this means imposing the constraint that the adversary and the baseline models form a martingale pair coupling. By Strassen's theorem (Strassen, 1965), it implies that the worst-case adversary dominates the baseline model in convex ordering, thus is a constraint that is sensible when the loss is convex as a function of the random noise. Adding sensible constraints to DRO formulations, in turn, is useful to control over-conservative estimators. In the particular case of martingale constraints, it is shown in Li et al. (2022) that these constraints recover the classical ridge regression and Tikhonov regularization. A popular variant of the optimal transport discrepancy, the Sinkhorn divergence, is also extensively used in machine learning tasks. This notion can also be used to construct distributional uncertainty sets as demonstrated in Wang et al. (2021); Dapogny et al. (2023); Azizian et al. (2023).

As we shall discuss next, together with ϕ-divergence and optimal transport can be unified under the lens of optimal transport by adding moment constraints related to martingale constraints. A recent unification of these DRO frameworks is presented in Blanchet et al. (2023b), which proposes to lift the sample space by considering the "likelihood" itself as a random variable and therefore is amenable to perturbations based on optimal transport, subject to the constraint that a likelihood ratio must have expectation equal to unity. The constraint that likelihood is non-negative can also be handled using optimal transport because the value of ground transportation cost function can be taken as positive infinity in part of its domain (note that only lower semi-continuity is required in the definition). More precisely, the authors consider the lifted empirical measure in the space of $\Xi \times \mathbb{R}_+$: $\hat{\mathbb{P}}_n \times \delta_1 = ((\xi_1, 1), \ldots, (\xi_n, 1))$, the uncertainty set is formulated as $\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{Q} \in \mathcal{P}(\Xi \times \mathbb{R}_+) : \mathbb{ID}_M(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta\}$, where, for a function $\phi(\cdot)$ defined in Definition 2.1 and a cost function $c(\cdot, \cdot)$ defined in Definition 2.2,

$$\mathbb{ID}_M(\mathbb{Q}, \hat{\mathbb{P}}_n) = \min_\pi \left\{ \int_{(\Xi \times \mathbb{R}_+) \times (\Xi \times \mathbb{R}_+)} \bar{c}((\xi, u), (\xi', u')) \mathrm{d}\pi((\xi, u), (\xi', u')) : \right.$$

$$\pi_{(\xi,u)} = \mathbb{Q}, \ \pi_{(\xi',u')} = \hat{\mathbb{P}}_n \times \delta_1, \mathbb{E}_\pi[u] = 1 \Big\},$$

$$\bar{c}((\xi,u),(\xi',u')) = u \cdot c(\xi,\xi') + (\phi(u) - \phi(u'))^+.$$

This particular choice of a cost function interpolates the aforementioned optimal transport-based and ϕ-divergence-based methods. Moreover, the resulting unified DRO problem is formulated as

$$\min_{\theta \in \Theta} \ \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \ \mathbb{E}_{\mathbb{Q}}[u \cdot \ell(\theta,\xi)],$$

where the modified expected loss $\mathbb{E}_{\mathbb{Q}}[u \cdot \ell(\theta,\xi)]$ illustrates the lifting technique that introduces the variable u as a "likelihood".

This approach provides a principled way to interpolate the norm regularization and the variance regularization. Moreover, the worst-case adversarial strategies simultaneously re-weight and perturb actual likelihoods based on dual gradient directions as discussed earlier in the settings of pure ϕ-divergence-based and optimal transport-based DRO, respectively. Moreover, if one alternatively chooses the nominal distribution with a kernel density, then this framework can also recover the Sinkhorn divergence-based DRO.

2.1.3 Integral Probability Metric-Based DRO

In statistics, energy distances and maximum mean discrepancies (MMD) are also widely considered (Székely, 1989). They also arise in the framework of kernel-based DRO (Staib and Jegelka, 2019; Zhu et al., 2021).

Definition 2.3 (Maximum mean discrepancy)**.** For a reproducing kernel Hilbert space $\mathcal{H}$ (see, e.g. Berlinet and Thomas-Agnan (2011)), the maximum mean discrepancy between distributions $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$ is defined as

$$\mathbb{D}_{\mathcal{H}}(\mathbb{Q},\hat{\mathbb{P}}_n) = \sup_{f:\|f\|_{\mathcal{H}} \leq 1} \ \int_\Xi f \mathrm{d}\mathbb{Q}(\xi) - \int_\Xi f \mathrm{d}\hat{\mathbb{P}}_n(\xi).$$

$\square$

When choosing different spaces $\mathcal{H}$, the kernel-based DRO can recover various DRO formulations, for example, moment-constraint DRO (see Zhu et al. (2021, Example 3.4)). Thus, this type of formulation provides a unified approach to studying a class of uncertainty sets in the DRO. A more general DRO framework is to consider the integral probability metrics (IPM) as the statistical distance (see, e.g., Zhu et al. (2021, Corollary 3.1.1)).

Definition 2.4 (Integral probability metrics)**.** For a function class $\mathcal{F}$, the integral probability metric between distributions $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$ is defined as

$$\mathbb{D}_{\mathcal{F}}(\mathbb{Q},\hat{\mathbb{P}}_n) = \sup_{f \in \mathcal{F}} \ \int_\Xi f \mathrm{d}\mathbb{Q}(\xi) - \int_\Xi f \mathrm{d}\hat{\mathbb{P}}_n(\xi).$$

$\square$

Specifically, assume that $\Xi = \mathbb{R}^d$, then

- if $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ for a Hilbert space $\mathcal{H}$, $\mathbb{D}_{\mathcal{F}} = \mathbb{D}_{\mathcal{H}}$;

- if $\mathcal{F} = \{f : \|f\|_{Lip} \leq 1\}$, where

$$\|f\|_{Lip} = \sup_{\xi,\xi'} \frac{|f(\xi) - f(\xi')|}{\|\xi - \xi'\|_q},$$

  then $\mathbb{D}_{\mathcal{F}} = \mathbb{D}_c$ for $c(\xi,\xi') = \|\xi - \xi'\|_q, q \geq 1$, which results from the Kantorovich-Rubenstein duality (Kantorovich and Rubinshtein, 1958);

- if $\mathcal{F} = \{f : \|f\|_\infty \le \frac{1}{2}\}$, then $\mathbb{D}_\mathcal{F} = \mathbb{D}_\phi$ for $\phi(z) = \frac{1}{2}|z - 1|$, which results from the variational representation of the ϕ-divergence.

2.1.4 Calibration of DRO with Side Information

In practical scenarios, available side information often includes: (i) the structure of the data-generating distribution; (ii) the characteristics of the out-of-sample environment; and (iii) domain knowledge. This information is especially helpful in mitigating the conservatism of the DRO while robustifying the DRO solution against the unknown out-of-sample environment. In the subsequent discussion, we showcase several DRO formulations that leverage diverse types of information, illustrating potential integrations of DRO with specific side information—all grounded in the aforementioned DRO formulations.

(i) *Correlation inside data.* When the observed data have internal correlation within its structure, the DRO framework can be applied with corresponding structure of uncertainty sets. For example, given i.i.d. samples $(x_i, y_i), i = 1, \dots, n$ generated from $\mathbb{P}_\star$, where $y_i \in \mathbb{R}^m$ represents the response variable and $x_i \in \mathbb{R}^n$ represents the predictor or covariate. Assume that the practitioner wants to solve

$$\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\star} [\ell(\theta, y) | x = x_0] \quad \text{for } x_0 \in \mathbb{R}^n,$$

which provides an estimator of y exploiting the side information carried by x. This problem presents several challenges. Firstly, the probability distribution $\mathbb{P}_\star$ is typically unknown, and only i.i.d. samples are available, which may be at risk of contamination. Additionally, there may be a scarcity of observations in the samples with covariate value $x = x_0$. To address these challenges without relying on additional parametric assumptions, Nguyen et al. (2020b, 2021) consider the DRO formulation

$$\min_{\theta \in \Theta} \sup_{\substack{\bar{\mathbb{P}} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n), \\ \bar{\mathbb{P}}(x \in \mathcal{N}(x_0)) > 0}} \mathbb{E}_{\bar{\mathbb{P}}} [\ell(\theta, y) | x \in \mathcal{N}(x_0)],$$

where $\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \left\{ \mathbb{P} : \mathbb{D}_c(\mathbb{P}, \hat{\mathbb{P}}_n) \le \delta \right\}$ and $\hat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$. Here, the side information is specifically modelled as a neighborhood $\mathcal{N}(x_0)$ around a covariate x_0 and the distribution $\bar{\mathbb{P}}$ is constraint such that $\bar{\mathbb{P}}(x \in \mathcal{N}(x_0)) > 0$ to avoid conditioning on a set of measure zero. When $\delta = 0$ and $\mathcal{N}(x_0)$ exactly contains the k nearest samples in $(X_i, 1 \le i \le n)$ from x_0, then this formulation recovers the k-nearest neighbor regression estimator, which is consistent (Stone, 1977) when k is suitably chosen with respect to n. Thus, this formulation can also be considered as a robustification of the k-nearest neighbor estimator.

As another example, Hu et al. (2018) and Sagawa et al. (2020) leverage the prior knowledge of correlations between samples in the training data to robustify their decision against potential group shifts, formally assuming that the data-generating distribution is a mixture of m latent groups, i.e.,

$$\mathbb{P}_\star = \sum_{\eta=1}^m q_\eta \mathbb{P}_{\xi|\eta}.$$

Correspondingly, the uncertainty set of the group DRO is built on the latent probability q_η using ϕ-divergence, i.e.,

$$\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{Q} : \mathbb{D}_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) \le \delta, \mathbb{Q}_{\xi|\eta} = (\hat{\mathbb{P}}_n)_{\xi|\eta} \, \forall \eta = 1, \dots, m\},$$

where the distributions $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$ are assumed to have the same structure with m latent groups.

Additional formulations motivated by group regularization in the context of optimal transport-based DRO are studied in Blanchet and Kang (2017).

(ii) *Out-of-sample environmental shift.* The uncertainty set characterizes the potential shifts in the out-of-sample environment or, put differently, envisions how an adversary might perturb the environment upon deployment of practitioners' decisions. Contrasting with examples in the preceding paragraphs, where adversarial perturbations are assumed to be linked to the internal structure of the data-generating distribution, practitioners can also directly leverage their knowledge of potential distributional shifts to construct uncertainty sets.

An instance of reshaping or informing the uncertainty set with prior knowledge of potential distributional shifts is illustrated in Blanchet et al. (2022b). In the portfolio selection problem, the return vector ξ is the random input of the optimization problem. The authors construct an uncertainty set on the distribution of ξ using an optimal transport cost (Definition 2.2) which is informed by current market information, in particular, the implied volatility derived by calibrating option prices based on the Black-Scholes formula with the market option prices. The implied volatility provides insight into the market participant's collective future belief about the volatility of an asset. Thus, if the implied volatility is large, the transportation cost should be relatively low so that the adversary can more efficiently use the budget to explore the adversarial impact of investing in an asset with, for instance, high implied volatility. Likewise, if the implied volatility is small, it is sensible to impose a relatively high transportation cost because in this way the DRO formulation will discourage the adversary from exploring the impact of future variations on assets that are perceived to be safe collectively by the market, which is captured by the implied volatility. Therefore, the formulation illustrated in Blanchet et al. (2022b) takes the following form.

$$
\begin{aligned}
\mathcal{B}_\delta &= \{\mathbb{Q} : \mathbb{D}_c(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta\}, \\
c(\xi_i, \xi) &= (\xi_i - \xi)^\top A_i (\xi_i - \xi), \\
A_i &= \frac{\bar{V}}{V_i} I_d \quad i = 1, \dots, n.
\end{aligned}
$$

Here, the transport cost of perturbing the sample return ξ_i is defined using the squared Mahalanobis distance with a specified matrix A_i. A_i is the identity matrix scaled by $\frac{\bar{V}}{V_i}$, where $\bar{V} = \frac{1}{n}\sum_{i=1}^n V_i$ and V_i is the implied volatility corresponding to the sample return ξ_i. As a result, for the distributions in the uncertainty set $\mathcal{B}_\delta$, it is cheaper to perturb sample returns with higher implied volatility. The formulation illustrates the intuition discussed earlier, namely, that higher implied volatility suggests larger price uncertainty in future returns by the collective market.

Similarly in Blanchet et al. (2021a), the authors tune the cost function defined in the optimal transport discrepancy to apply in classification problems. The authors first fit a cost function with the property that observations with the same labels are close, while observations with different labels are far apart. Then, this cost function is used as a transportation cost in an optimal transport-based DRO logistic regression formulation. The intuition is that the adversarial budget will be invested more efficiently if the adversary is encouraged to perturb data points that are easier to be flipped in the decision boundary, i.e., moved to the population with the opposite label.

Yet another example arises in the context of unsupervised learning (Blanchet and Kang, 2020, 2021), where the authors unlabeled observations to shape the distributional uncertainty set. The intuition is that the underlying data may lie in a lower dimensional space and only variations along such a lower dimensional space should be explored by the adversary in an optimal transport formulation.

(iii) *Domain knowledge.* The utilization of domain knowledge to enhance the out-of-sample performance of decisions has been extensively investigated in the field of domain adaptation (refer to, e.g., Weiss et al. (2016); Wilson and Cook (2020); Iman et al. (2023)). As a robust formulation aimed at minimizing worst-case risk under potential distributional shifts, the DRO framework can be synergistically integrated with the concept of domain adaptation to further enhance its out-of-sample performance.

For instance, Taskesen et al. (2021) proposes two strategies for applying the DRO in the context of domain adaptation to solve the linear prediction problem. The authors introduce the DRO in a parametric setting such that the uncertainty set is prescribed using only finite parametrization of distributional moments. Here, for simplicity of notation, we present the high-level idea of their work without explicitly delving into the details of their parametric setting. Given nominal distribution from source and target, namely, $\hat{\mathbb{P}}_S$ and $\hat{\mathbb{P}}_T$, the authors propose to create a class of "experts" that integrates the knowledge of source and target and then aggregate the experts. To create robust experts enhancing its out-of-sample performance, they consider two types of uncertainty sets $\mathcal{B}$ and solve the distributionally robust least squares estimation problem in the following form.

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}} \mathbb{E}_{\mathbb{Q}} \left[(\beta^\top x - y)^2 \right], \text{ where}$$

$$(a) \; \mathcal{B} = \mathcal{B}_\delta(\hat{\mathbb{P}}_\lambda) = \{\mathbb{Q} : \mathbb{D}(\mathbb{Q}, \hat{\mathbb{P}}_\lambda) \leq \delta\};$$

$$(b) \; \mathcal{B} = \mathcal{B}_{\delta_S}(\hat{\mathbb{P}}_S) \bigcap \mathcal{B}_{\delta_T}(\hat{\mathbb{P}}_T) = \{\mathbb{Q} : \mathbb{D}(\mathbb{Q}, \hat{\mathbb{P}}_S) \leq \delta_S, \mathbb{D}(\mathbb{Q}, \hat{\mathbb{P}}_T) \leq \delta_T\}.$$

Here, the radii $\delta, \delta_S, \delta_T > 0$, $\hat{\mathbb{P}}_\lambda$ is an interpolation of $\hat{\mathbb{P}}_S$ and $\hat{\mathbb{P}}_T$ parametrized by $\lambda \in [0, 1]$, and the statistical divergence $\mathbb{D}$ can be taken as ϕ-divergence $\mathbb{D}_\phi$ (Definition 2.1) or optimal transport discrepancy $\mathbb{D}_c$ (Definition 2.2). Thus, for each type (i.e. (a) and (b)) of uncertainty and each statistical divergence $\mathbb{D}$, the authors choose various values of radii and parameters to generate a class of "experts" that integrates the knowledge from source and target.

Another example for solving linear regression in the context of domain adaptation is proposed in Zhang et al. (2022a). In their work, the authors evaluate the worst-case risk of the estimator over all pairs of source and target within a specified distance between each other. This formulation can also be viewed as a DRO problem such that the source lies in the uncertainty set centered at the target, which is prescribed as a neighborhood ball in the distribution space. Notably, Zhang et al. (2022a) also discusses the minimax optimality of the proposed estimator with respect to all pairs of source and target within the specified distance in a non-asymptotic sense.

2.2 Statistical Properties of DRO

In this subsection, we focus on the statistical properties of the estimator obtained by solving the DRO problem, referred to as the DRO estimator in the following. Our focus is on the large sample regime, where the sample size approaches infinity while the dimension of individual samples remains fixed. We will discuss the limiting behavior of the DRO estimator to characterize its efficiency in addition to its finite sample properties.

In the scenario when the practitioner wants to defend against potential overfitting, it is natural to shrink the radius of the uncertainty set in the DRO when the number of sample goes to infinity, i.e., $\lim_{n \to \infty} \delta_n = 0$. This is because the empirical distribution $\hat{\mathbb{P}}_n$ converges to the data-generating distribution $\mathbb{P}_\star$ and thus the uncertainty with respect to $\mathbb{P}_\star$ decays to zero. Following this vein, a series of studies explores the appropriate shrinkage rate of the radius of the uncertainty set such that the DRO estimator is consistent and nearly efficient. These studies also give rise to new tools for statistical inference, such as confidence region construction.

Alternatively, if the practitioner aims to provide a robust estimator that performs well not only under the data-generating distribution but also in the face of the unknown distributional shift—a common modeling

challenge—the determination of the radius becomes a nuanced task. In this scenario, the radius may not solely hinge on the sample size but also on the decision maker's risk posture concerning potential variations in the distribution, particularly relative to the empirical distribution derived from past observations. From this perspective, it is non-trivial (and not purely a statistical problem) to optimally choose not only the size of the distributional uncertainty but the distributional uncertainty set. Some studies thus focus on fixing an uncertainty set and consider and compare the DRO solution using empirical data with its population version, which is the solution of the same DRO problem except for centering its uncertainty set at the true data-generating distribution.

In the subsequent discussion, we focus on the uncertainty sets in the form previously described:

$$\mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n) = \left\{ \mathbb{Q} : \mathbb{D}(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta_n \right\},$$

where $\mathbb{D}$ represents a probability metric to be chosen, and the radius δ_n may depend on n. The statistical properties of other DRO variants featuring more general uncertainty sets have, to a large extent, remained underexplored.

2.2.1 Selection of Radius δ_n

The selection of the radius δ_n significantly impacts the statistical property of the DRO solution. When δ_n is too large, the resulting strategy is too conservative to learn useful information from the data. When too small, the strategy is not robust enough to combat against the noise in the data. In many scenarios, the radius δ_n can be directly interpreted as the penalization parameter (see, for example, Theorem 2.1, 2.2). This connection—the penalization parameter is the radius of uncertainty set—provides an interpretable way of selecting the penalization parameter from the view of uncertainty magnitude, other than possibly computation-intensive machine learning methods like cross-validation.

From a theoretical perspective, a sequence of studies suggests ensuring that $\mathbb{P}_\star \in \mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n)$ with high probability (for example, Mohajerin Esfahani and Kuhn (2018, Theorem 3.5)). However, for the likelihood-based ϕ-divergence, $\mathbb{P}_\star$ may never be in the uncertainty set $\mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n)$ when it has continuous density because in this case the likelihood ratio between $\mathbb{P}_\star$ and the empirical distribution $\hat{\mathbb{P}}_n$ is not well-defined. To address this issue, one approach is to consider the parametric distributions that are compatible with $\hat{\mathbb{P}}_n$ and ensure that the uncertainty set contains at least one of them with high probability (see, e.g. Delage and Ye (2010, Corollary 4), Nguyen et al. (2020c, Lemma 4.4)).

For another popular statistical distance, optimal transport discrepancy, the choice of radius may be subject to the curse of dimensionality. In particular, the optimal transport discrepancy between $\mathbb{P}_\star$ and $\hat{\mathbb{P}}_n$ with cost function $c(\xi, \eta) = \|\xi - \eta\|^p, \xi, \eta \in \mathbb{R}^d$ is in the order of $n^{-p/d}$ in expectation (Fournier and Guillin, 2015, Theorem 1), which implies that the radius δ_n, if selected in the aforementioned way, will also grow in the order of $n^{-p/d}$ and result in a too conservative learning strategy in practice. To overcome this issue, instead of letting the uncertainty set contain $\mathbb{P}_\star$ with high probability or fitting $\hat{\mathbb{P}}_n$ to parametric distributions, another radius selection method is to guarantee that the uncertainty set contains at least one distribution of which the learned parameter is compatible with $\mathbb{P}_\star$. In particular, let

$$\theta(\mathbb{Q}) = \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)],$$
$$\Lambda_{\delta_n} = \left\{ \theta(\mathbb{Q}) : \mathbb{Q} \in \mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n) \right\}.$$

Then the radius δ_n is selected such that $\theta(\mathbb{P}_\star) \cap \Lambda_\delta(\hat{\mathbb{P}}_n) \neq \emptyset$ in high probability. Formally, for a specified probability threshold α, the radius δ_n is selected as the smallest number such that

$$\mathbb{P}_\star\left(\theta(\mathbb{P}_\star) \in \Lambda_{\delta_n}(\hat{\mathbb{P}}_n)\right) \geq 1 - \alpha, \tag{2.1}$$

where the first $\mathbb{P}_\star$ is short for the infinite product data-generating distribution $\mathbb{P}_\star^{\otimes\infty}$. While this optimization problem appears daunting, in Section 2.2.5 we will explain how to reduce it to a projection which (in the setting of ϕ-divergence) is closely related to empirical likelihood. A concrete recipe for implementing this selection method is presented in Blanchet et al. (2021b, Algorithm 1). This selection method can be extensively applied to the optimal transport discrepancy-based DRO (Blanchet et al., 2019a,b) and the ϕ-divergence-based DRO (Lam and Zhou, 2017; Blanchet et al., 2019b).

Notably, the selection rule (2.1) finds its connection in the classic LASSO regression literature. In the high dimension regression setting, Theorem 2.2 shows that the radius δ_n of uncertainty set is interpreted as the penalization parameter in front of the norm regularization of the regression parameter. In this case, the rule (2.1) suggests a similar selection of δ_n as in the LASSO literature (Blanchet et al., 2019a, Theorem 7) and it scales similarly in the high dimensional setting and it is also independent of the residual standard errors as in the square-root LASSO setting in Belloni et al. (2011).

2.2.2 Asymptotic Normality of DRO Estimators

In this part, we discuss the limiting behavior the DRO estimator by presenting their asymptotic normality under regular assumptions. We begin by discussing the scenario where δ_n approaches zero as n tends to infinity, and then we explore the case where δ_n remains fixed throughout.

Precisely, we define the DRO estimator with shrinking radius δ_n to be

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta,\xi)].$$

and the parameter to be learned, which is assumed to be unique, is defined as

$$\theta_\star = \arg\min_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta,\xi)].$$

Under regular assumptions, when the radius of the uncertainty set δ_n shrinks to zero as n grows to infinity, the DRO estimator $\hat{\theta}_n$ converges to $\theta_\star$ almost surely. Intuitively, when δ_n approaches zero too fast, the limiting behavior of the DRO estimator will be close to the minimizer of the empirical risk without robust approach; when δ_n grows to slowly, the DRO estimator will be too conservative, leading to poor efficiency when estimating $\theta_\star$. It turns out, when appropriately choosing the shrinking rate of δ_n (as outlined in Section 2.2.1), the error of the DRO estimator with respect to $\theta_\star$ can be characterized by the central limit theorem with an expected convergence rate $1/\sqrt{n}$.

Alternatively, the results in Section 2.1 can also help to decide the shrinking rate of δ_n. For example, when the uncertainty set is based on the ϕ-divergence and $\phi''(1) > 0$, Theorem 2.1 indicates that the error in the estimation of the optimal expected loss is of order $O(\delta_n^{1/2})$. In view of the central limit theorem, this suggests selecting $\delta_n = \frac{\bar{\delta}}{n}$ for some constant $\bar{\delta} > 0$. (If δ_n shrinks faster than $\frac{1}{n}$, the DRO solution cannot provide a satisfying robust certificate as we will discuss in Section 2.2.3). For the constant $\bar{\delta}$, the procedures outlined in Section 2.2.1 and 2.2.5 provide its precise selection. As a result, given $\delta_n = \frac{\bar{\delta}}{n}$, the central limit theorem holds as follows

Theorem 2.4 (Duchi and Namkoong (2018), Theorem 6, informal)**.** Under smoothness assumptions on $\ell(\cdot, \cdot)$,

$$\sqrt{n} \left(\hat{\theta}_n - \theta_\star \right) \xrightarrow{d} N \left(-\sqrt{2\bar{\delta}} b, \Sigma \right),$$

for some vector b and matrix Σ that depends on $\mathbb{P}_\star, \theta_\star, \ell(\cdot, \cdot)$. Here, $\xrightarrow{d}$ denotes the convergence in distribution. □

The asymptotic bias is explicitly characterized in Duchi and Namkoong (2018), but the point is that both the bias and the covariance matrix depend on $\theta_\star$. This dependence is typically continuous in $\theta_\star$, thus the asymptotically valid confidence regions of $\theta_\star$ can be built based on this result with any consistent plug-in estimator.

A completely analogous result can be obtained in the context of the optimal transport discrepancy. Once again, the choice of δ_n can either be guided by the method outlined in the section 2.2.1 to provide its precise selection, or by using the analysis outlined earlier for the structure of the worst case distribution. For example, when the cost function is $c(\xi, \xi') = \|\xi - \xi'\|_q^2$, we saw that the perturbation size is of order $O(\delta_n^{1/2})$, thus leading to an error in the optimal expected loss of order $O(\delta_n^{1/2})$. This implies once again selecting $\delta_n = \frac{\bar{\delta}}{n}$, the central limit theorem results in a completely similar format asymptotically normal limit. It is important, however, that the asymptotic bias is different, but both the asymptotic variances (in the contexts of ϕ-divergence and optimal transport discrepancy) coincide with the case of the empirical risk minimization estimator (i.e. the case $\delta_n = 0$). The asymptotic bias is typically also continuous in θ_* and therefore any consistent plug-in estimator can be used to generate asymptotically valid confidence intervals; see Blanchet et al. (2022a, Theorem 1). The work of Blanchet and Shapiro (2023) provides a comprehensive discussion of both the DRO and the ϕ-divergence asymptotic normality results building from a sensitivity analysis perspective and studying the different asymptotic distributions for different choices of uncertainty radii.

It is important to note that the DRO estimator has a significant bias of order $O(n^{-1/2})$ compared to the bias of the standard empirical risk minimization estimator. Therefore, it is valid for the reader to wonder what is the point of a DRO-based estimator in the face of the above result which indicates that the asymptotic mean squared error.

This is a valid criticism from the standpoint of a purely mean-squared error criterion as a function of the sample size, one that we will address in the conclusion section of this review paper. It suffices to say here, that mean-squared error is not the only criterion of interest and that there are additional parameters of interest and not only sample size (e.g. the complexity of the model to be learned) that are important. The optimality and efficiency of DRO estimators as a function of various statistical parameters of interest (modeling complexity class and equitability or fairness, among other criteria) are topics of significant research interest. In the next section, we will review finite sample results which provide insight into some of these questions.

By contrast, if we keep the radius δ_n fixed as n goes to infinity, and consider its convergence of the DRO estimator to its population counterpart (defined as follows), we have central limit theorem with zero asymptotic bias. To be precise, we define the DRO estimator with fixed radius δ to be

$$\hat{\theta}_{n,\delta} \in \arg \min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)]. \tag{2.2}$$

and its population counterpart, which is assumed to be unique, is defined as

$$\theta_{\star,\delta} = \arg \min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)].$$

When the uncertainty set is based on ϕ-divergence, the central limit theorem of the convergence of $\hat{\theta}_{n,\delta}$ is given by

Theorem 2.5 (Duchi and Namkoong (2021), Theorem 11, informal)**.** Under regular assumptions on ℓ,

$$\sqrt{n}\left(\hat{\theta}_{n,\delta} - \theta_{\star,\delta}\right) \xrightarrow{d} N\left(0, \Sigma^*\right),$$

for some variance matrix Σ^* depending on $\mathbb{P}_\star, \ell(\cdot, \cdot), \delta, \phi, \theta_{\star,\delta}$. $\hfill\square$

The underlying idea of proving this result is to view the DRO estimator as an M-estimator in view of the DRO duality problem. To be precise, by Shapiro (2017, Section 3.2), the DRO duality problem is

$$\sup_{\mathbb{Q}\in\mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta,\xi)] = \inf_{\lambda\geq 0,\mu\in\mathbb{R}}\left\{\mathbb{E}_{\hat{\mathbb{P}}_n}\left[\lambda\phi^*\left(\frac{\ell(\theta,\xi)-\mu}{\lambda}\right)\right] + \lambda\delta + \mu\right\},$$

where $\mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{Q} : \mathbb{D}_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq \delta\}$, $\phi^*(s) = \sup_{t\in\mathbb{R}}\{s \cdot t - \phi(t)\}$. Similarly, in other context such as the optimal transport discrepancy, the central limit theorem of $\hat{\theta}_{n,\delta}$ can be established by solving the corresponding DRO duality problem.

2.2.3 Finite-Sample Guarantees of DRO Estimators

In the studies of statistical properties of DRO, many efforts have been made to characterize the out-of-sample generalization power of the DRO estimator, particularly under the true data-generating distribution. These endeavors involve establishing finite-sample statistical guarantees in the form of

$$\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}_n, \xi)] \leq \inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta,\xi)] \tag{2.3}$$

with high probability (see, for example, Delage and Ye (2010, Theorem 3), Mohajerin Esfahani and Kuhn (2018, Theorem 3.5)). The optimal value of the DRO problem is thus termed as the certificate of the out-of-sample performance of $\hat{\theta}_n$. The reason for emphasizing this one-sided upper bound is probably that in many applications of the DRO, underestimation of the loss is more harmful than overestimation of the loss.

The more recent studies, such as Duchi and Namkoong (2018, Corollary 5), An and Gao (2021, Theorem 1) and Gao (2022, Theorem 1), address the issue of over-conservatism in DRO by selecting a more refined uncertainty set with smaller radii δ_n (e.g. by sending δ_n to zero in a faster rate). This choice often introduces an additional error term of the order $\frac{1}{n}$ to the certificate. These results, however, typically impose strong assumptions on the distributions (e.g. finite support or sub-Gaussianity) or not intended to be used optimally in the high-dimensional setting, for instance.

In addition to building a certificate on the loss under the data-generating distribution, the finite sample guarantees of the DRO estimator's out-of-sample performance are also established through the finite-sample upper bound of:

1. $\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}_n, \xi)] - \inf_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta, \xi)]$: in the context of applying shrinking radius δ_n when the out-of-sample environment is $\mathbb{P}_\star$ (Duchi and Namkoong (2018, Corollary 5), Gao (2022, Remark 1)).

2. $\sup_{\mathbb{Q}\in\mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_\mathbb{Q}[\ell(\hat{\theta}_{n,\delta}, \xi)] - \inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)]$: in the context of applying fixed radius δ when the out-of-sample environment may have distributional shift from $\mathbb{P}_\star$ (Lee and Raginsky (2018, Theorem 3), Duchi and Namkoong (2021, Corollary 2)).

2.2.4 Optimality of DRO Estimators

In addition to its success in empirical studies, theoretical investigations into the DRO aim to establish a statistical optimality argument, further demonstrating its advantages. Several endeavors have provided potential avenues.

In a recent study, Van Parys et al. (2021) justifies the optimality of DRO formulation with the help of the large deviation principles. The authors concentrate on a specific class of certificates, stipulating that the probability of underestimation of the expected loss under $\mathbb{P}_\star$ decays exponentially fast as the sample size n grows to infinity (similar to equation (2.3)). The optimal certificate (in the sense of being the least conservative) is shown to be equivalent to a DRO formulation based on ϕ-divergence (in particular corresponding to inverse KL). For those readers familiar with large deviations theory, the result may be natural given that it formally corresponds to an application of Sanov's theorem. This result is interesting because it provides an interpretation of DRO estimators as being optimal in some sense. However, the optimality criterion hinges on stipulating a large deviation gap in expected loss estimation (which traduces to a fixed uncertainty radius), which is too pessimistic a criterion as an inferential tool.

In another study on the DRO with fixed radius δ in the context of ϕ-divergence-based uncertainty set, Duchi and Namkoong (2021) proves that the DRO estimator can achieve a sharp minimax bound for the loss

$$\sup_{\mathbb{Q} \in \mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_\mathbb{Q}[\ell(\hat{\theta}_{n,\delta}, \xi)] - \inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)],$$

where $\hat{\theta}_{n,\delta}$ is defined in (2.2). This result demonstrates the ability of DRO estimator to learn with uniformly good performance. However, the loss function $\sup_{\mathbb{Q} \in \mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_\mathbb{Q}[\ell(\cdot, \xi)]$ for the DRO estimators to achieve optimal rate still appears too conservative in practice. Further, whether the DRO estimator can achieve a sharp minimax bound for other context such as optimal transport discrepancy is open.

2.2.5 Statistical Inference of DRO

The uncertainty set on the distribution implemented in the DRO inspires new confidence regions of the statistics under the true data-generating distribution, such as

$$\theta_\star = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta, \xi)] \quad \text{and} \quad \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta_\star, \xi)] = \min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta, \xi)].$$

For the confidence region of $\theta_\star$, the radius selection rule (2.1) suggests a suitable candidate

$$\Lambda_{\delta_n} = \bigcup \left\{ \theta(\mathbb{Q}) : \mathbb{Q} \in \mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n) \right\},$$

where $\theta(\mathbb{P}) = \arg\min_{\theta \in \Theta} \mathbb{E}_\mathbb{P}[\ell(\theta, \xi)]$. To select δ_n, note that

$$\mathbb{P}_\star \left(\theta(\mathbb{P}_\star) \in \Lambda_{\delta_n} \right)$$

$$= \mathbb{P}_\star \left(\exists \mathbb{Q} \in \mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n) : \theta(\mathbb{P}_\star) \in \operatorname*{arg\,min}_{\theta \in \Theta} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)] \right)$$

$$= \mathbb{P}_\star \left(\min_{\mathbb{Q} : \theta(\mathbb{P}_\star) \in \arg\min_{\theta \in \Theta} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)]} \mathbb{D}(\hat{\mathbb{P}}_n, \mathbb{Q}) \le \delta_n \right).$$

Then the problem is reduced to compute the quantile of the projection distance

$$\min_{\mathbb{Q} : \theta(\mathbb{P}_\star) \in \arg\min_{\theta \in \Theta} \mathbb{E}_\mathbb{Q}[\ell(\theta, \xi)]} \mathbb{D}(\hat{\mathbb{P}}_n, \mathbb{Q}).$$

For $\mathbb{D} = \mathbb{D}_\phi$, the quantile is related to Owen's theory of the empirical likelihood (Owen, 2001). As a result, by Owen (2001, Theorem 3.4),

$$n \times \min_{\mathbb{Q}:\theta(\mathbb{P}_\star)\in\arg\min_{\theta\in\Theta}\mathbb{E}_\mathbb{Q}[\ell(\theta,\xi)]} \mathbb{D}_\phi(\hat{\mathbb{P}}_n, \mathbb{Q}) \xrightarrow{d} \frac{\phi''(1)}{2}\chi^2_{(q)},$$

where $\chi^2_{(q)}$ is the chi-square distribution with degree of freedom q and q depends on $\ell(\cdot, \cdot), \theta_\star, \mathbb{P}_\star$. Therefore, given the confidence level $1 - \alpha$, δ_n is taken as $\frac{\bar{\delta}}{n}$, where $\bar{\delta}$ is the $(1-\alpha)$-quantile of the distribution $\frac{\phi''(1)}{2}\chi^2_{(q)}$.

For $\mathbb{D} = \mathbb{D}_c$, Blanchet et al. (2019a) provides the quantile when the cost function is defined by the Euclidean distances. For cost functions of the form $c(\xi, \xi') = \|\xi - \xi'\|_q^2$, similar to the context of the ϕ-divergence, δ_n is taken as $\frac{\bar{\delta}}{n}$, where $\bar{\delta}$ depends on the quantile of limiting distribution of the projection distance given a confidence level. Specifically, Blanchet et al. (2019a) considers cost functions $c(\xi, \xi') = \|\xi - \xi'\|_q^2$ for various values of q, implying different geometry used when constructing the confidence region. Figure 4 shows the confidence regions of the linear regression estimator computed in Blanchet et al. (2019a, Section 4.2).
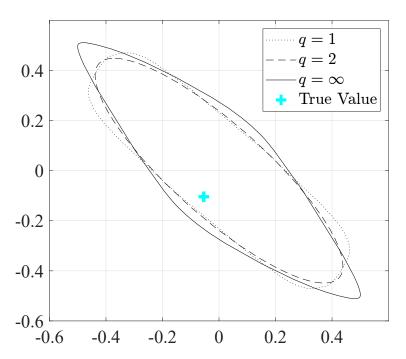

Figure 4: Confidence regions induced by optimal transport discrepancy-based DRO using different Euclidean distances as cost functions.

For the confidence interval of $\mathbb{E}_{\mathbb{P}_\star}[\ell(\theta_\star, \xi)]$, it is straightforward to use the optimal value of the corresponding data-driven DRO problem as an upper bound. To get a lower bound, we consider an alternative problem with the sup inside of DRO replaced by inf (Lam and Zhou, 2017). Let

$$u_n = \min_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta,\xi)],$$

$$l_n = \min_{\theta\in\Theta} \inf_{\mathbb{Q}\in\mathcal{B}_{\delta_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_\mathbb{Q}[\ell(\theta,\xi)],$$

with δ_n chosen similar to (2.1) for a specified threshold $\alpha \in (0, 1)$, then

$$\liminf_{n \to \infty} \mathbb{P}_\star \left(\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta, \xi)] \in [l_n, u_n] \right) \geq 1 - \alpha.$$

If additional smoothness of ℓ is assumed, then the asymptotic coverage can be exactly $1 - \alpha$ with a smaller δ_n (Duchi et al., 2021, Theorem 3).

2.3 Tractability of DRO

In this subsection, we discuss the tractability of the DRO formulations. To be more specific, the computation complexity of DRO problems depends on both the loss function and the uncertainty set, and thus we list some existing algorithms case by case. It's worth noting that the development of fast algorithms for solving different DRO models is still in its initial stages.

Starting from the moment-based DRO (Delage and Ye, 2010), the authors have paid attention to the tractable algorithms to solve the DRO problems, and specifically for the moment-based DRO problem, a reformulation as semidefinite programming is available. For the DRO based on optimal transport discrepancy, if the loss function is piecewise concave with respect to the random input, then the DRO can be reformulated as a conic programming and can be solved using general nonlinear optimization solver, e.g., MOSEK or Gurobi; if the loss function for linear decision rule is considered, Li et al. (2019, 2020); Blanchet et al. (2022b) propose some fast first-order methods; if the loss function is parameterized as a neural network, some certificate is tractable as shown in Sinha et al. (2018). For the ϕ-divergence-based DRO and convex loss function with respect to the parameter, Levy et al. (2020) proposes the mini-batch gradient descent method with complexity that is independent of sample size and dimension, and optimally dependent on the uncertainty set size.

Some evidence implies that the DRO can even bring benefits to the computation. For example, when $\ell(\theta, \xi)$ is strongly convex in θ, for a suitable selection of the radius δ, it is shown that the corresponding DRO problem is also strongly convex in θ (Blanchet et al., 2022b, Theorem 4), while the original non-DRO problem (i.e. $\delta = 0$) may not be strongly convex, for example, in the high dimensional regression problem.

2.4 DRO in the Bayesian Framework

A famous quote, often attributed to the statistician George Box (Box, 1976) indicates that "every model is wrong, but some are useful." Our previous discussion of the DRO framework shows its flexibility in addressing issues related to model misspecification, emphasizing the deviation of out-of-sample distribution from the samples' empirical distribution employed by data-driven methods. However, in a model-driven setting, the issue of misspecification may be even more severe. As an extension to enhancing the robustness of the data-driven decision-making cycle, this subsection illustrates the application of DRO to mitigate the impact of model misspecification within the framework of Bayesian statistics. We emphasize that this area is significantly less investigated, but still, there is a substantial and growing literature that studies this setting.

In the Bayesian framework, the sensitivity of the model performance to the perturbations of the prior and likelihood is an important topic. While there is a rich literature in Bayesian statistics to model robustness, the vast majority of which are related to the specification of prior. Interested readers are referred to another review paper Watson and Holmes (2016). Actually, the robustness with respect to the specification of likelihood is no less important than the prior. However, modeling and hedging against the distributional misspecification of the likelihood is less studied in the Bayesian framework, possibly because robust formulations of the likelihood naturally lead to infinite dimensional (i.e. non-parametric) formulations, and this results in significantly higher

complexity. In contrast, the DRO framework routinely manages non-parametric distributional uncertainty sets, making it conducive for integration into the Bayesian framework. This strategy will be illustrated through examples in the following discussion.

In this subsection, we assume that $\eta \in \mathbb{R}^p$ denotes a prior parameter, $\xi \in \mathbb{R}^d$ denotes the observed sample following a distribution parameterized by η, and $\hat{\theta}(\cdot) \in \Phi$ denotes a decision policy to be taken based on the full Bayesian model. For example, in an investment problem, η denotes the (unknown) vector of mean of the distribution of returns, the returns are denoted as ξ, and $\hat{\theta}(\xi)$ is the vector of portfolio allocations based on the observed returns. The decision maker is interested in minimizing the risk, which can be taken as the negative of a utility.

Precisely, assume that $\ell : \Phi \times \Xi \times \mathcal{H} \to \mathbb{R}$ is the loss function, $\mathbb{Q}_{\xi|\eta}$ is the likelihood and $\mathbb{Q}_\eta$ is the prior, we consider the stochastic optimization problem

$$\min_{\hat{\theta}(\cdot) \in \Phi} \mathbb{E}_{\mathbb{Q}}\left[\ell(\hat{\theta}(\cdot), \xi, \eta)\right] = \mathbb{E}_{\mathbb{Q}_\eta}\left[\mathbb{E}_{\mathbb{Q}_{\xi|\eta}}\left[\ell(\hat{\theta}(\cdot), \xi, \eta)\right]\right].$$

Given $\xi_0 \in \Xi$, we also consider the problem with a slightly different objective

$$\min_{\hat{\theta}(\cdot) \in \Phi} \mathbb{E}_{\mathbb{Q}}\left[\ell(\hat{\theta}(\cdot), \xi, \eta)|\xi = \xi_0\right].$$

For example, when considering the Bayesian minimum mean square estimation problem, we can set $\ell(\hat{\theta}(\cdot), \xi, \eta) = \|\eta - \hat{\theta}(\xi)\|_2^2$.

We now review recent work on the DRO frameworks built on the aforementioned objectives and applied to typical problems in robust Bayesian inference (Levy and Nikoukhah, 2012; Zorzi, 2016; Shafieezadeh Abadeh et al., 2018; Nguyen et al., 2020a; Zhang et al., 2022b; Nguyen et al., 2023; Lotidis et al., 2023). The DRO framework introduced in these studies includes the uncertainty on both prior and likelihood, which extends some of the classic studies on robust prior involving both prior and likelihood misspecification.

(i). Levy and Nikoukhah (2012); Zorzi (2016); Shafieezadeh Abadeh et al. (2018); Lotidis et al. (2023) consider the Kalman filtering: for $i = 1, \ldots, T$,

$$\eta_i = D_i \eta_{i-1} + v_i, \quad \xi_i = B_i \eta_i + u_i,$$

where $v = (v_1, \ldots, v_T)$ denotes the innovation process and $u = (u_1, \ldots, u_T)$ denotes the noise process. Let $\xi = (\xi_i, 1 \le i \le T)$ and $\eta = (\eta_i, 1 \le i \le T)$, the loss function is set to be

$$\ell(\hat{\theta}(\cdot), \xi, \eta) = \sum_{i=1}^T \left\|\eta_i - \hat{\theta}_i(\xi_1, \ldots, \xi_i)\right\|^2.$$

where $\hat{\theta}(\cdot) = (\hat{\theta}_i(\cdot), 1 \le i \le T)$ are the linear functionals to be learned. In the corresponding DRO formulation, the authors consider an uncertainty set on the joint distribution of $(\eta_1, \ldots, \eta_T)$ and $(\varepsilon_1, \ldots, \varepsilon_T)$, that is, given $\xi = \xi_0$,

$$\min_{\hat{\theta}(\cdot) \in \Phi} \sup_{\mathbb{Q} \in \mathcal{B}} \mathbb{E}_{\mathbb{Q}}\left[\ell(\hat{\theta}(\cdot), \xi, \eta)|\xi = \xi_0\right],$$

where $\mathcal{B}$ is an uncertainty set that is defined by a neighborhood ball centered at some nominal distribution, measured by the ϕ-divergence (Levy and Nikoukhah, 2012; Zorzi, 2016) or the optimal transport

discrepancy (Shafieezadeh Abadeh et al., 2018; Lotidis et al., 2023). In the case of ϕ-divergence-based uncertainty set, as it turns out, the decision rule often remains unchanged compared to the non-DRO case, but the mean squared error naturally increases. In contrast, in the case of optimal transport-based uncertainty set, the decision rule undergoes a change. In both cases, if the baseline distribution is Gaussian, the worst-case distribution remains Gaussian and, therefore, affine decision rules are optimal when minimizing over the class arbitrary prediction functions. This result is also shown in Nguyen et al. (2023), where the authors consider the general Bayesian minimum mean square error estimation problem with the optimal transport-based uncertainty set.

(ii). Nguyen et al. (2020a) considers the Bayesian classification problem with the classification error to be

$$\ell(\hat{\theta}(\cdot), \xi, \eta) = \hat{\theta}(\xi)\mathbb{I}_{\eta=0} + (1 - \hat{\theta}(\xi))\mathbb{I}_{\eta=1},$$

where $\eta \in \{0, 1\}$ denotes the unobserved label, $\xi \in \Xi$ is the observed sample to be classified, and $\hat{\theta} : \Xi \to \{0, 1\}$ is the (randomized) classifier to be learned. The authors consider the DRO formulation of an objective, that given $\xi = \xi_0$,

$$\min_{\hat{\theta}(\cdot) \in \Phi} \sup_{\mathbb{Q} \in \mathcal{B}} \mathbb{E}_{\mathbb{Q}}\left[\ell(\hat{\theta}(\cdot), \xi, \eta)|\xi = \xi_0\right].$$

Specifically, the uncertainty set $\mathcal{B}$ is built on the joint distribution of prior and likelihood and is set to be

$$\mathcal{B} = \left\{ \mathbb{Q} = \pi_0 \mathbb{Q}_{\xi|\eta=0} + \pi_1 \mathbb{Q}_{\xi|\eta=1} : \begin{array}{ll} \mathbb{Q}_{\xi|\eta=0} \in \mathcal{P}, & \mathbb{D}_\phi(\mathbb{Q}_{\xi|\eta=0}, \mathbb{Q}_0) \leq \delta_0 \\ \mathbb{Q}_{\xi|\eta=1} \in \mathcal{P}, & \mathbb{D}_\phi(\mathbb{Q}_{\xi|\eta=1}, \mathbb{Q}_1) \leq \delta_1 \end{array} \right\},$$

where $\delta = (\delta_0, \delta_1)$, (π_0, π_1) is a specified prior weight on η, $(\mathbb{Q}_0, \mathbb{Q}_1)$ are two nominal distributions, and $\mathcal{P}$ is a specified parametric family.

(iii). Zhang et al. (2022b) considers Gaussian process regression and linear inverse problems. This is one of the few DRO results that involve non-parametric decision rules, non-parametric distributional uncertainty, and an infinite dimensional outcome space. In particular, the formulation is as follows,

$$\xi_i = \eta(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n.$$

where $\xi = (\xi_i, 1 \leq i \leq n)$ are observed values in $\mathbb{R}$, η is the parameter of interest, which is modeled as a real-valued random process (endowed with prior) with continuous sample paths, $(t_i, 1 \leq i \leq n)$ are some specified evaluation points in a domain $\mathcal{D} \subset \mathbb{R}^d$, and $(\varepsilon_i, 1 \leq i \leq n)$ denote the observational noise. The loss function is

$$\ell(\hat{\theta}(\cdot), \xi, \eta) = \left\| \eta - \hat{\theta}(\xi_1, \ldots, \xi_n) \right\|_{\mathcal{L}^2(\mathcal{D})}^2,$$

where the estimator $\hat{\theta} : \mathbb{R}^n \to \mathcal{L}^2(\mathcal{D})$ maps the observed values $(\xi_i, 1 \leq i \leq n)$ to a continuous path on the domain $\mathcal{D}$. In the corresponding DRO formulation, the authors consider the uncertainty set on the joint distribution of the sample path η (prior) and ξ (likelihood) using the optimal transport discrepancy. In particular, they consider the reproducing kernel Hilbert space (RKHS) to characterize the neighborhood of a sample path of η and further the neighborhood of the nominal distribution of η. Remarkably, the authors show that if the ground transportation cost function c is a squared Hilbert norm and the nomial distribution at the center of the uncertainty set is a Gaussian process, then the worst-case distribution is also a Gaussian process, and therefore the optimal prediction function is affine in the observations. Moreover, the authors show that a Nash equilibrium exists and it is unique for sufficiently small uncertainty budgets.

3 Robust Statistics

$$\mathbb{P}_n = \tfrac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i^\star}, \quad \hat{\mathbb{P}}_n = \tfrac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$
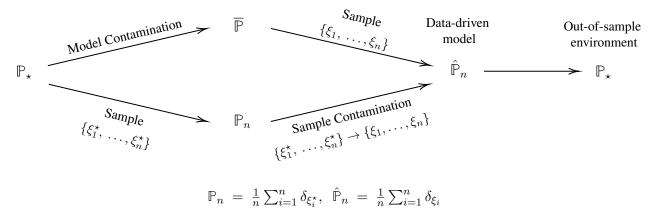
Figure 5: Decision Making Cycle in the Context of Robust Statistics

In the context of robust statistics, robust estimators act after the occurrence of pre-decision distributional contamination. For example, in scenarios involving noisy observations, samples are i.i.d. generated from a contaminated data-generating distribution. In another agnostic setting, a more potent adversary is assumed to contaminate the samples after they are generated from the true data-generating distribution. To distinguish between these two types of pre-decision contamination, we refer to Figure 5, which, in contrast to Figure 1, excludes post-decision contamination and focuses on estimating the true data-generating distribution. A more detailed discussion of types of pre-decision contamination will be presented in Section 3.1.1.

To clarify our notations in this section, we use the following symbols: $\overline{\mathbb{P}}$ for the contaminated data-generating distribution, $\mathbb{P}_n$ for the empirical distribution composed of uncontaminated samples, and $\hat{\mathbb{P}}_n$ for the empirical distribution composed of contaminated samples (also representing the observed samples, consistent with the notation used in the discussion of DRO). We will use $\hat{\mathbb{P}}_n$ interchangeably to denote both the contaminated samples and their empirical distribution, and $\mathbb{P}_n$ interchangeably to denote both the uncontaminated samples and their empirical distribution.

In the subsequent sections, we embark on a comprehensive review of the field of robust statistics. Following this, we draw comparisons between robust statistics and DRO. For ease of exposition, our focus will be on the fundamental task of robust estimation for the location parameter, specifically robust mean estimation, serving as a running example in our review.

3.1 Literature Review

The field of robust statistics emerged from statistical procedures tailored to confront two practical challenges: the detection and rejection of "outliers", and the analysis of data when the underlying distribution deviates from normality. Historically, there was ambiguity associated with the concept of robustness (Huber, 1972). The term "robustness" appeared to have first been introduced by Box (Box, 1953), who used the term to refer to the property of a procedure insensitive to departures from ideal assumptions (Box, 1979). Tukey also pioneered the recognition of sensitivity of some conventional statistical procedures to minor deviations from the assumptions (Tukey, 1960, 1962), and developed a series of work in nonparametric methods and rank-based procedures (Scheffe and Tukey, 1944, 1945). The early foundations of robust estimation were further developed by Huber (Huber, 1964, 1968) and Hampel (Hampel, 1968, 1971), among others. To specify

the setup of robust estimation considered here, we again assume that ξ is a random vector in the space Ξ that follows a data-generating distribution $\mathbb{P}_\star$ (the canonical example being a normal distribution), and consider $\theta \in \Theta$ as the parameter of the model to be learned. In particular, we consider $\theta = \theta(\mathbb{P}_\star)$ coming from a statistical functional $\theta(\cdot) : \mathcal{P}(\Xi) \to \Theta$. In the framework of M-estimators (Vaart, 1998), the functional $\theta(\mathbb{P}_\star)$ would be a minimizer of the criterion function

$$\min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\star}[\ell(\theta, \xi)] = \int_{\mathbb{R}^d} \ell(\theta, \xi) \mathrm{d}\mathbb{P}_\star(\xi), \tag{3.1}$$

where $\ell(\theta, \xi)$ is a measurable extended real-valued function. For example, to obtain the mean parameter it is typical to consider the squared loss $\ell(\theta, \xi) = \|\theta - \xi\|_2^2$. Given observations $\xi_1, \dots, \xi_n$ each of which follows marginal law $\mathbb{P}_\star$, and are assumed to be independent, the M-estimator is the solution to the empirical criterion function

$$\min_{\theta \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta, \xi)] = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, \xi_i), \tag{3.2}$$

where $\hat{\mathbb{P}}_n$ denotes the empirical measure $\frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$. In the example of mean estimation, the solution turns out to be the sample mean. In practice, the in-sample empirical measure $\hat{\mathbb{P}}_n$ may deviate from the ideal assumptions imposed, consequently compromising the accuracy of naive M-estimators. Regarding sample mean, the presence of even a single outlier within the data points can substantially undermine the statistical performance of this estimator. There are several different reasons for the in-sample empirical measure $\hat{\mathbb{P}}_n$ to deviate from the ideal assumptions. To be precise, now we review the type of deviations typically considered in the literature. These mechanisms of deviations are also called contamination models (or corruption models).

3.1.1 Types of Contamination Models

Conceptually, we consider the existence of an adversary, who generates the in-sample data set in an adversarial fashion. As shown in Figure 5, the adversary can perturb the true data-generating distribution $\mathbb{P}_\star$ to a contaminated data-generating distribution $\overline{\mathbb{P}}$, after which the samples $\hat{\mathbb{P}}_n$ are drawn i.i.d. from $\overline{\mathbb{P}}$. A stronger form of contamination is that the adversary can perturb the in-sample data set *adaptively*, which can contaminate the samples $\mathbb{P}_n$ that are directly generated from the uncontaminated data-generating distribution $\mathbb{P}_\star$, resulting in the contaminated samples $\hat{\mathbb{P}}_n$. Specifically, we mention

(i) *Huber's ϵ-contamination model*: In a seminal paper Huber (1964), Huber considered the contamination for the data-generating distribution as $\overline{\mathbb{P}} = (1-\epsilon)\mathbb{P}_\star + \epsilon\mathbb{H}$. In his original paper $\mathbb{P}_\star$ is a normal distribution, $\mathbb{H}$ is an unknown contamination distribution, and ϵ is a known constant representing the level of contamination. Hence, the adversary is allowed to add contamination to (but not subtract from) the population distribution that generates the samples.

(ii) *Full-neighborhood contamination*: Instead of the restricted neighborhood in Huber's contamination model, the adversary perturbs the data-generating distribution in a full-neighborhood by the use of a statistical distance $\mathbb{D}(\overline{\mathbb{P}}, \mathbb{P}_\star) \leq \epsilon$. When $\mathbb{D}$ is chosen as the total variation distance, this neighborhood is strictly larger than Huber's contamination. The choice of total variation distance is typically for the study of gross errors in the dataset (Donoho and Liu, 1988; Zhu et al., 2022). The statistical distance can be modified to study other types of natural model errors, such as rounding errors. For instance, refer to Zhu et al. (2022); Liu and Loh (2022) for the use of Wasserstein distance of order 1.

(iii) *Adaptive contamination model*: A more powerful contamination works as follows. Once the samples $\mathbb{P}_n$ are drawn i.i.d. from $\mathbb{P}_\star$, the adversary inspects the samples, and at their disposal, remove up to ϵn samples and replace them with arbitrary points, resulting in the contaminated in-sample data set $\hat{\mathbb{P}}_n$ (Diakonikolas et al., 2019a; Zhu et al., 2022). This is equivalent to the constraint that $\mathbb{D}(\mathbb{P}_n, \hat{\mathbb{P}}_n) \leq \epsilon$, where $\mathbb{D}$ is chosen as the total variation distance. To model the scenario where every sample can be slightly perturbed, Zhu et al. (2022); Liu and Loh (2022) also consider the use of Wasserstein distance of order 1.

3.1.2 Types of Robustness Criteria

The objective of robust statistics is to develop estimators that exhibit certain "robustness" properties in relation to the aforementioned forms of contamination. There are multiple criteria to quantify robustness. In this discussion, we examine some of the robustness criteria.

To restate our notations corresponding to the three contamination models previously described, we generically denote $\mathcal{A}_\epsilon(\mathbb{P}_\star) = \{\mathbb{P} : \mathbb{D}(\mathbb{P}, \mathbb{P}_\star) \leq \varepsilon\}$ to be a set of contamination models with probability metric $\mathbb{D}$ to be further specified, $\overline{\mathbb{P}} \in \mathcal{A}_\epsilon(\mathbb{P}_\star)$ to represent the population distribution from which the contaminated samples $\hat{\mathbb{P}}_n$ are drawn. This draw is i.i.d for the first two types of contamination models aforementioned. For the adaptive contamination model, we denote $\hat{\mathbb{P}}_n \in \mathcal{A}_\epsilon(\mathbb{P}_n)$ as the empirical distribution of samples modified from $\mathbb{P}_n$, which is the empirical distribution of samples i.i.d. generated from $\mathbb{P}_\star$. Further, we generically denote $\hat{\theta} = \hat{\theta}(\hat{\mathbb{P}}_n)$ as the estimator of interest.

Now we start to review the robustness criteria, which include:

(i) *Efficiency*: A small contamination level should cause a small degradation on the statistical performance. An intuitive formal criterion to minimize would be

$$err(\epsilon, \mathbb{P}_\star) = \sup_{\overline{\mathbb{P}} \in \mathcal{A}_\epsilon(\mathbb{P}_\star)} \mathbb{E}_{\overline{\mathbb{P}}} \left[\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)] \right], \tag{3.3}$$

where the outer expectation $\mathbb{E}_{\overline{\mathbb{P}}}$ averages the randomness over the draw of the contaminated samples $\hat{\mathbb{P}}_n$ and the inner expectation $\mathbb{E}_{\mathbb{P}_\star}$ denotes averaging over the draw of ξ following the true data-generating distribution $\mathbb{P}_\star$. The corresponding criterion to minimize for the adaptive contamination model would be

$$err(\epsilon, \mathbb{P}_n) = \sup_{\hat{\mathbb{P}}_n \in \mathcal{A}_\epsilon(\mathbb{P}_n)} \mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)],$$

where $\mathbb{E}_{\mathbb{P}_\star}$ denotes averaging over the draw of ξ following the true data-generating distribution $\mathbb{P}_\star$.

Choosing the squared loss $\ell(\theta, \xi) = \|\theta - \xi\|_2^2$ and the mean functional $\theta(\mathbb{P}_\star) = \mathbb{E}_{\mathbb{P}_\star}[\xi]$, the above criteria are equivalent to

$$err(\epsilon, \mathbb{P}_\star) = \sup_{\bar{\mathbb{P}} \in \mathcal{A}_\epsilon(\mathbb{P}_\star)} \mathbb{E}_{\bar{\mathbb{P}}} \left[\left\| \hat{\theta}(\hat{\mathbb{P}}_n) - \theta(\mathbb{P}_\star) \right\|_2^2 \right],$$

and

$$err(\epsilon, \mathbb{P}_n) = \sup_{\hat{\mathbb{P}}_n \in \mathcal{A}_\epsilon(\mathbb{P}_n)} \left\| \hat{\theta}(\hat{\mathbb{P}}_n) - \theta(\mathbb{P}_\star) \right\|_2^2,$$

respectively. For technical reasons, instead of the expected risk in (3.3), prior works focused on establishing high probability bounds for the quantity

$$\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)],$$

where besides assumptions on the contamination models, usually some assumptions on $\mathbb{P}_\star$ are imposed. For instance, the class of normal or elliptical distributions was considered in Chen et al. (2018); Chao et al.

(2019); Gao (2020), and the class of distributions having subgaussian tails or bounded moments was studied in Diakonikolas et al. (2017a); Steinhardt et al. (2017, 2018); Diakonikolas et al. (2018a, 2019b); Bateni and Dalalyan (2019); Liu et al. (2020); Depersin and Lecué (2022).

(ii) *Break-down point*: The notion of break-down point was introduced by Hampel (Hampel, 1968, 1971), and later developed by Donoho and Huber (Donoho and Huber, 1983; Huber, 2004) (among others) to quantify the influence of outliers on a given estimator $\hat{\theta}$. The finite-sample break-down point, following Donoho and Huber (1983), can be formulated as

$$\min\{\epsilon : \|\hat{\theta}(\mathbb{P}_n) - \hat{\theta}(\hat{\mathbb{P}}_n)\|_2 = \infty, \ \ \hat{\mathbb{P}}_n \in \mathcal{A}_\epsilon(\mathbb{P}_n)\},$$

the smallest fraction of corruption to the samples that causes the estimator to be arbitrarily bad. For example, the break-down point for the sample mean is easy to be seen as $\frac{1}{n}$. A variant of the break-down point in the population sense can be similarly defined as

$$\min\{\epsilon : err(\epsilon, \mathbb{P}_\star) = \infty\},$$

the smallest ϵ that causes the expected risk in (3.3) to be arbitrarily large.

Simultaneously achieving efficiency in terms of statistical convergence rate and a high break-down point is an on-going area of research. In Chen et al. (2018), the authors propose a population variant of the breakdown point which they term as "δ-breakdown point", and show that for a given estimator that has convergence rate δ under the Huber's ϵ-contamination model, its δ-breakdown point is at least ϵ. This suggests that efficiency under Huber's ϵ-contamination model may be a more general notion of robustness than the breakdown point.

3.1.3 Early Work on Robust Mean Estimation

In the seminal paper Huber (1964), for one-dimensional estimation, Huber considered an M-estimator by replacing the squared loss $\ell(\theta, \xi) = (\theta - \xi)^2$ with the loss function

$$\ell(\theta, \xi) = \rho(\theta - \xi),$$

where

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & \text{for } |t| < k \\ k|t| - \frac{1}{2}k^2 & \text{for } |t| \geq k, \end{cases}$$

and k is related to ϵ by a non-linear equation. Huber demonstrated the optimality of this M-estimator among all translation-invariant estimators in terms of robustness. However, the measure of robustness considered therein is defined as the worst-case (suprema) of the asymptotic variance of the estimator over Huber's ϵ-contamination model $\overline{\mathbb{P}} = (1 - \epsilon)\mathbb{P}_\star + \epsilon\mathbb{H}$, with normal $\mathbb{P}_\star$ and symmetric $\mathbb{H}$.

Tukey's median (Tukey, 1975) is a robust mean estimator that can be defined in any dimension d. First, Tukey's depth function of any $\eta \in \mathbb{R}^d$ with respect to any distribution $\mathbb{P}$ on $\mathbb{R}^d$ is defined as

$$\mathcal{D}(\eta, \mathbb{P}) = \inf_{u \in \mathcal{S}^{d-1}} \mathbb{P}(u^\top \xi \leq u^\top \eta), \quad \xi \sim \mathbb{P},$$

where $\mathcal{S}^{d-1}$ is the d-dimensional unit sphere in $\mathbb{R}^d$. Tukey's median is defined to be the deepest point with respect to the empirical distribution $\hat{\mathbb{P}}_n$

$$\hat{\theta}(\hat{\mathbb{P}}_n) = \arg\max_\eta \mathcal{D}(\eta, \hat{\mathbb{P}}_n).$$

The convergence rate of Tukey's median under Huber's ϵ-contamination model is $\|\hat{\theta}(\hat{\mathbb{P}}_n) - \theta(\mathbb{P}_\star)\|_2^2 \lesssim O\left(\frac{d}{n} \vee \epsilon^2\right)$, and such a finite-sample rate is also optimal in a minimax sense, as shown in Chen et al. (2018). Unfortunately, it is well-known that Tukey's median is NP-hard to compute in high dimensions (Johnson and Preparata, 1978; Bernholt, 2006). A natural alternative is the componentwise median. Despite having a high break-down point (Donoho and Gasko, 1992), this estimator suffers from a convergence rate $O\left(d\left(\frac{1}{n} \vee \epsilon^2\right)\right)$, which is inferior in high dimensions (Chen et al., 2018).

3.1.4 Recent Work on Computationally Efficient High-Dimensional Robust Mean Estimation

Until recently, computationally efficient methods capable of achieving statistically optimal convergence rates were elusive. Naive polynomial time approaches typically lead to an error rate of $O(d\epsilon^2)$ (assuming enough number of samples), which scales linearly with the dimension d. The recent works (Diakonikolas et al., 2017a, 2019a) obtained the first dimension-independent error rate for computationally efficient robust mean estimation of an isotropic normal distribution under the adaptive contamination model, and simultaneously, Lai et al. (2016) obtained an error rate that scales with $O(\log(d))$ under a weaker bounded fourth moment assumption. To obtain these dimension independent errors, fairly strong assumptions on the uncorrupted sample $\mathbb{P}_n$ are imposed.

Definition 3.1 (Stability (Diakonikolas and Kane, 2023, Definition 2.1)). Fix $0 < \epsilon < 1/2$ and $\delta \geq \epsilon$. A finite set $S \subseteq \mathbb{R}^d$ is (ϵ, δ)-stable with respect to a vector μ if for every unit vector $v \in \mathcal{S}^{d-1}$ and every $S' \subseteq S$ with $|S'| \geq (1 - \varepsilon)|S|$, the following conditions hold:

1. $\left|\frac{1}{|S'|} \sum_{\xi \in S'} v \cdot (\xi - \mu)\right| \leq \delta$, and

2. $\left|\frac{1}{|S'|} \sum_{\xi \in S'} (v \cdot (\xi - \mu))^2 - 1\right| \leq \frac{\delta^2}{\epsilon}$.

Similarly, a distribution $\mathbb{P}$ on $\mathbb{R}^d$ is (ϵ, δ)-stable with respect to a vector μ if for every unit vector $v \in \mathcal{S}^{d-1}$ and distribution $\bar{\mathbb{P}}$ obtained from $\mathbb{P}$ by conditioning on an event of probability at least $1 - \epsilon$, the following conditions hold:

1. $\mathbb{E}_{\bar{\mathbb{P}}}\left[|v \cdot (\xi - \mu)|\right] \leq \delta$, and

2. $\left|\mathbb{E}_{\bar{\mathbb{P}}}\left[(v \cdot (\xi - \mu))^2\right] - 1\right| \leq \frac{\delta^2}{\varepsilon}$.

$\square$

In simple terms, these conditions imply that the removal of any ϵ-fraction of the sample points will not change the mean by more than δ nor the variance in any direction by more than $\frac{\delta^2}{\epsilon}$. It can be shown that these conditions are satisfied by many distributions (and with high probability their empirical samples) with appropriate concentration properties (e.g., normal distributions). The first condition in Definition 3.1 is also studied under the name of *resilience* by Steinhardt et al. (2018).

Under these stability conditions, Diakonikolas et al. (2019a) proposed an iterative greedy algorithm aimed at purifying a dataset by progressively removing corrupted samples. Starting with an initial data set S that includes both corrupted and uncorrupted samples, the algorithm in each iteration either calculates the sample mean of the current set of samples or employs a filter to refine S into a subset S' that is substantially closer to the uncontaminated set.

Another approach proposed in Diakonikolas et al. (2019a) is a convex programming based algorithm. Weights w_i are computed for each sample ξ_i, so the weighted empirical average $\hat{\mu}_{\boldsymbol{\omega}} = \sum_{i=1}^N \omega_i \xi_i$ approximates

the true mean μ. These weights are constrained within a convex set C_δ, defined as:

$$C_\delta = \left\{ \boldsymbol{\omega} : \begin{array}{c} \sum_{i=1}^n \omega_i = 1,\ 0 \le \omega_i \le \frac{1}{(1-\varepsilon)n}\ \forall i \\ \left\| \sum_{i=1}^n \omega_i (\xi_i - \mu)(\xi_i - \mu)^\top - I \right\|_2 \le \delta \end{array} \right\}.$$

Since μ is unknown, the algorithm substitutes μ by $\hat{\mu}_{\boldsymbol{\omega}}$. Using spectral techniques to approximate a separation oracle for C_δ, the algorithm is shown to achieve computational efficiency in estimating μ.

Since the works of Lai et al. (2016); Diakonikolas et al. (2017a, 2019a), the field has seen a proliferation of research. For a comprehensive overview, one can refer to the survey by Diakonikolas and Kane (2019). In a different thread of works, Chao et al. (2019) established a connection between generative adversarial networks (Goodfellow et al., 2014) and classical depth-based robust estimators (e.g. Tukey's median), leading them to study robust mean estimators using GANs and establish minimax optimal error bounds. This connection also enables for the computation of robust estimators utilizing techniques originally developed for training GANs. The follow-up works extended the technique for f-GAN (Wu et al., 2020) and Wasserstein-GAN (Liu and Loh, 2022), under various contamination models.

3.1.5 Information-Theoretical Lower Bound

A unified expression of the minimax rates for robust estimation was developed by Chen et al. (2018). The minimax rate is defined as the quantity $M(\epsilon)$ that satisfies, for a constant $c > 0$,

$$\inf_{\hat{\theta}} \sup_{\overline{\mathbb{P}} \in \mathcal{A}_\epsilon(\mathbb{P}_\star)} \overline{\mathbb{P}} \left(\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)] \ge M(\epsilon) \right) \ge c \tag{3.4}$$

holds. Here the outer probability $\overline{\mathbb{P}}$ (shorthand for the product measure $\overline{\mathbb{P}}^n$) is over the randomness of the draw of the contaminated samples $\hat{\mathbb{P}}_n$ and the inner expectation $\mathbb{E}_{\mathbb{P}_\star}$ denotes averaging over the draw of ξ following the true data-generating distribution $\mathbb{P}_\star$. This minimax rate was shown to have the form of

$$M(\epsilon) = \max\{M(0), \omega(\epsilon, \mathcal{F})\},$$

where $M(0)$ is the classical minimax rate for uncontaminated distributions, and $\omega(\epsilon, \mathcal{F})$ represents the modulus of continuity over a family $\mathcal{F}$ of probability distributions. For example, $\mathcal{F}$ represents the convex combination of normal distributions and an ϵ fraction of contamination in Huber's contamination model.

This concept of modulus of continuity, tracing back to the foundational work of Donoho and Liu (Donoho and Liu, 1991; Donoho, 1994), represents the fact that in the worst-case contamination scenario, it is theoretically impossible to distinguish between parameters within $\omega(\epsilon, \mathcal{F})$ for a given loss.

In the existence of densities, Liu and Gao (2019) studied density estimation under pointwise loss in the presence of contamination, and derived the minimax optimal rate.

Besides the task of robust mean estimation, recent works have also focused on robust covariance estimation (Lai et al., 2016; Diakonikolas et al., 2019a; Zhu et al., 2022), learning mixtures of spherical Gaussians (Kothari and Steinhardt, 2017; Hopkins and Li, 2018), lower bounds against statistical query algorithms (Diakonikolas et al., 2017b), list-decodable learning (Diakonikolas et al., 2018b; Karmalkar et al., 2019; Raghavendra and Yau, 2020), robust linear regression (Bhatia et al., 2015, 2017; Klivans et al., 2018; Suggala et al., 2019) and robust stochastic optimization (Charikar et al., 2017; Diakonikolas et al., 2019b; Prasad et al., 2020). It is worth noting that an expanding body of research on robust estimation is focusing on robustifying estimators to heavy-tailed distributions, see Audibert and Catoni (2011); Minsker (2015); Donoho and Montanari (2016); Devroye et al. (2016); Joly et al. (2017); Lugosi and Mendelson (2019). The results of which are of a different nature comparing to the setting of contamination models.

3.2 Comparing DRO and Robust Statistics

The attentive reader may recognize that the formulations (3.1) and (3.2) presented above is analogous to equations (1.1) and (1.2). This parallelism is an intentional choice to facilitate a comparison between the frameworks of robust statistics and DRO. We now integrate robust statistics in a lense that is reminiscent of the DRO framework.

Consider a two-player zero-sum game where the nature's action is to generate the contaminated in-sample data set $\hat{\mathbb{P}}_n$, and, having observed $\hat{\mathbb{P}}_n$, the statistician's action space is a class of policies Ψ for constructing an estimator $\hat{\theta}(\cdot) : \mathcal{P}(\Xi) \to \Theta$. The loss incurred to the statistician (or equivalently the gain incurred to the nature) is thus

$$\mathbb{E}_{\widetilde{\mathbb{P}}}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)],$$

which measures the risk of the chosen decision $\hat{\theta}(\hat{\mathbb{P}}_n)$ in the out-of-sample environment $\widetilde{\mathbb{P}}$. In the context of robust statistics, this environment is the same as the original data-generating distribution, i.e. $\widetilde{\mathbb{P}} = \mathbb{P}_\star$. The important characterization of this game is that the nature first generates $\hat{\mathbb{P}}_n$, after which the statistician chooses the decision which takes advantage of the realizations $\hat{\mathbb{P}}_n$. Hence, we have a *max-min* game

$$\sup_{\overline{\mathbb{P}} \in \mathcal{A}_\epsilon(\mathbb{P}_\star)} \inf_{\hat{\theta}(\cdot) \in \Psi} \mathbb{E}_{\overline{\mathbb{P}}}\left[\mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)]\right], \tag{3.5}$$

where the outer expectation $\mathbb{E}_{\overline{\mathbb{P}}}$ averages the randomness over the draw of the contaminated samples $\hat{\mathbb{P}}_n$ and the inner expectation $\mathbb{E}_{\mathbb{P}_\star}$ denotes averaging over the draw of ξ following the true data-generating distribution $\mathbb{P}_\star$.

This is a harder game to play for the nature given the ordering of the actions, as the nature needs to foresee the choice of the statistician. When $\hat{\mathbb{P}}_n$ is generated according to the adaptive contamination model, the game (3.5) is framed as the sample-wise max-min game

$$\sup_{\hat{\mathbb{P}}_n \in \mathcal{A}_\epsilon(\mathbb{P}_n)} \inf_{\hat{\theta}(\cdot) \in \Psi} \mathbb{E}_{\mathbb{P}_\star}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)], \tag{3.6}$$

whose value is dependent on the realization $\mathbb{P}_n$ of the empirical distribution of uncontaminated samples.

Comparing this to the DRO formulation (1.3), which is a *min-max* game repeated below with a slightly different presentation :

$$\inf_{\hat{\theta}(\cdot) \in \Phi} \sup_{\widetilde{\mathbb{P}} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_{\widetilde{\mathbb{P}}}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)]. \tag{3.7}$$

In this game, the statistician first observes the clean in-sample data set $\hat{\mathbb{P}}_n$, and makes a decision $\hat{\theta}(\hat{\mathbb{P}}_n)$ according to a class of policies Φ. After that, nature enters into the game by choosing an out-of-sample environment $\widetilde{\mathbb{P}} \in \mathcal{B}_\delta(\hat{\mathbb{P}}_n) = \{\mathbb{P} : \mathbb{D}(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \delta\}$ with some probability metric $\mathbb{D}$, in an adversarial fashion relative to the statistician's decision. In particular, nature's action can depart from the in-sample empirical measure $\hat{\mathbb{P}}_n$, and is constrained by the ambiguity set $\mathcal{B}(\hat{\mathbb{P}}_n)$. The loss incurred to the statistician (or equivalently the gain incurred to the nature) is thus

$$\mathbb{E}_{\widetilde{\mathbb{P}}}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)],$$

which measures the risk of the decision $\hat{\theta}(\hat{\mathbb{P}}_n)$ in the out-of-sample environment $\widetilde{\mathbb{P}}$. This is a harder game to play for the statistician given the ordering of the actions.

The value of the game (3.7) is also sample dependent, which is comparable to the game (3.6) under the adaptive contamination model. A slight adjustment to this DRO formulation would bear a closer resemblance with (3.5). Here we consider the formulation

$$\inf_{\hat{\theta}(\cdot) \in \Phi} \sup_{\widetilde{\mathbb{P}} \in \mathcal{B}_\delta(\mathbb{P}_\star)} \mathbb{E}_{\mathbb{P}_\star}\left[\mathbb{E}_{\widetilde{\mathbb{P}}}[\ell(\hat{\theta}(\hat{\mathbb{P}}_n), \xi)]\right],$$

where $\mathcal{B}_\delta(\mathbb{P}_\star)$ is the set of adversarial distributions. The outer expectation $\mathbb{E}_{\mathbb{P}_\star}$ is averaging over the randomness of the empirical measure $\hat{\mathbb{P}}_n$, upon which the statistician's action $\hat{\theta}(\cdot) \in \Phi$ is based, and the inner expectation $\mathbb{E}_{\widetilde{\mathbb{P}}}$ averages the out-of-sample environment chosen by the nature.

3.3 Connection of Robust Statistics to Rockafellian relaxations

In the context of robust statistics, once the statistician receives the contaminated sample $\hat{\mathbb{P}}_n$, a sensible strategy is first to rectify the contamination, possibly in an "optimistic" fashion, and then optimize the loss function based on the rectified data points. This decision-making approach involves a joint minimization of the form

$$\inf_{\theta \in \Theta} \inf_{\mathbb{Q} \in \mathcal{R}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\ell(\theta, \xi)]. \tag{3.8}$$

Recent studies, such as those by Norton et al. (2017); Nguyen et al. (2019); Jiang and Xie (2023); Royset et al. (2023); Gotoh et al. (2023), have considered this optimistic formulation in problem settings where inherent conservativeness of the DRO paradigm could render an actually optimal solution as suboptimal (Royset et al., 2023). For example, Nguyen et al. (2019) considers an optimistic formulation where the goal is to find the best-case measure that maximizes the non-parametric likelihood in the setting of Bayesian likelihood approximation, and the rectification set $\mathcal{R}$ is constrained by KL-divergences, moment constraints, or Wasserstein distances. Also see Gotoh et al. (2023) for a result that the optimistic decision-making can sometimes outperform the empirical risk minimizer in out-of-sample contexts while there is no guarantee for DRO solutions. Comparing to DRO, the optimistic formulation is typically non-convex, see Aravkin and Davis (2020) for a stochastic proximal-gradient algorithm for solving the resulting nonconvex optimization problem in the setting of trimmed M-estimators.

The concept of optimistic decision-making in the operations research literature can be traced back to "Rockafellian Relaxations" (Rockafellar, 1963, 1974, 1985, 1997; Royset, 2021; Royset and Wets, 2022; Royset et al., 2023). Also refer to Zalinescu (2002); Bauschke and Combettes (2011), where optimistic decision making was studied under the name "perturbation functions" or "bivariate functions".

A natural question is how to choose the rectification set $\mathcal{R}(\hat{\mathbb{P}}_n)$ appropriately to clean up the contamination. In recent works, it is shown that there are some interesting connections between (3.8) and well-known estimator in robust statistics. For example, in the setting of univariate mean estimation, Jiang and Xie (2023) considers the simple formulation

$$\inf_{\theta} \inf_{\boldsymbol{\omega} \in \mathcal{R}} \sum_{i=1}^{n} \omega_i (\theta - \xi_i)^2,$$

where the rectification set is defined as a re-weighting of the samples

$$\mathcal{R} = \left\{ \boldsymbol{\omega} \in \mathbb{R}_+^n : \sum_{i=1}^{n} \frac{1}{\omega_i} = n^2 \right\}.$$

They show that the resulting estimator corresponds to the sample median. This rectification set is slightly more general than formulation (3.8) in that the weights ω_i do not necessarily sum to one. Nevertheless, under this rectification set, the authors Jiang and Xie (2023) also show that they are able to recover more similar robust statistics, such as median absolute deviation, least absolute deviation, and least median of squares.

We conclude this section by demonstrating that the solution to the min-min formulation (3.8) recovers an instance of statistical minimax optimality in the literature.

Theorem 3.1 (Informal corollary of (Zhu et al., 2022, Theorem H.1)). Let $\mathbb{P}_\star$ be σ^2-subgaussian, let $\overline{\mathbb{P}} \in \mathcal{A}_\epsilon(\mathbb{P}_\star)$ follow the ϵ-total variation contamination model, from which the contaminated sample $\hat{\mathbb{P}}_n$ are drawn i.i.d, and let $\mathcal{R}_\delta(\hat{\mathbb{P}}_n)$ be a δ-total variation neighborhood of $\hat{\mathbb{P}}_n$ comprising "resilient" distributions (Zhu et al., 2022, equation (441)). Then, choosing $\delta = 2\left(\sqrt{\epsilon} + \sqrt{\frac{\log(1/\eta)}{2n}}\right)^2$, we define the estimator $\hat{\theta}(\hat{\mathbb{P}}_n)$ as the solution to

$$\arg\min_{\theta \in \mathbb{R}^d} \min_{\mathbb{Q} \in \mathcal{R}_\delta(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}\left[\|\theta - \xi\|_2^2\right].$$

With probability at least $1 - 3\eta$, it holds that

$$\left\|\theta(\mathbb{P}_\star) - \hat{\theta}(\hat{\mathbb{P}}_n)\right\|_2 \leq C\sigma\left(\epsilon\sqrt{\log(\epsilon)} + \sqrt{\frac{d + \log(1/\eta)}{n}}\right),$$

where C is a universal constant. The dependence on ϵ, d and n are information-theoretically optimal (cf. equation (3.4)). $\qquad\qquad\square$

We remark that minimax optimality of the min-min formulation (3.8) in a wider context (e.g., Wasserstein contamination) is a prospective area of interest for future research.

4 Conclusions and Discussion

The goal of this section is to briefly discuss various areas of potential research interest in connection with DRO and statistics. The discussion is not exhaustive, but rather we want to expose the fact that DRO as a statistical tool offers a wide range of opportunities for the statistical community.

We mentioned in Section 2.2.2, simply in terms of asymptotic mean squared error as the sample size increases, it is sensible to ponder on the benefits of DRO estimators. The work of Lam (2021) shows that in the presence of sufficient regularity (e.g. smoothness of the loss) DRO estimators tend to dominate in second-order stochastic dominance of the empirical risk minimization estimator of the optimal loss. This observation is consistent with our discussion in Section 2.2.2. Nevertheless, the situation may be different if these regularity conditions are not satisfied. For instance, Duchi and Namkoong (2018, Section 3.3) shows that in settings involving non-smooth losses, DRO estimators may enjoy superior rates compared to their empirical risk minimization counterparts.

Continuing in the context of classical statistical analysis involving large sample properties. There are objects that the DRO estimation approach offers that are interesting statistically speaking. The most natural such object is the worst-case distribution, which is a by-product of the DRO approach and possesses rich interpretations. Even in the context of the estimators that DRO recovers exactly and that are well-known in statistics, the DRO approach furnishes additional insight to these classical estimators using the associated worst-case distribution.

Another example of an interesting statistical object to study offered by DRO formulations is the natural confidence region induced by the DRO and discussed in Subsection 2.2.5. Using the duality between confidence regions and hypothesis testing we can compare the efficiency of various confidence regions implied by standard notions of efficiency in hypothesis testing.

Statistical efficiency is also of interest in connection to important parameters, such as the dimension, for example. We have seen that a suitably chosen distributional uncertainty region can be used to show the

equivalence between a DRO estimator and a well-known estimator. An example of this situation is square-root LASSO and the DRO-motivated choice of uncertainty size recovers regularization prescriptions studied in the high-dimensional statistics literature. Likewise, in the context of ϕ-divergence, the DRO-based estimator is used to re-weight samples in order to hedge against significant inference errors in subpopulations. In summary, while the DRO estimator may have a higher asymptotic mean squared error compared to the empirical risk minimization estimator when used in situations in which the statistical problem is ill-posed (i.e. the sample size is relatively small compared to the information required to estimate the parameter) or when the goal is not purely based on mean-squared error but we are interested in hedging a different type of risk, then DRO based estimation offers enough flexibility and interpretability, not only through their formulation but also through the associated worst-case distribution.

In general, we also note that adding constraints or exploring other types of distributional uncertainty sets that can be used to better inform the attributes of the adversary to reduce conservativeness is a significant topic of research interest. For example, the work of Olea et al. (2022) explores different DRO uncertainty sets based on the sliced Wasserstein distance. The advantage of this formulation is that it does not suffer from the statistical course of dimensionality for comparing distributions in high dimensions (as is the case of the Wasserstein distance); see also the approaches recently advocated by Bennouna and Van Parys (2022); Liu et al. (2023).

Another area of significant interest which we touched only superficially is the issue of fairness. We mentioned that ϕ-divergence has been utilized to try to improve the inference quality in estimated statistics involving minority sub-populations. Other DRO-based ideas have been recently applied in the context of fairness. For example, Taskesen et al. (2020); Si et al. (2021) propose a projection-based hypothesis test closely related to the one discussed in Section 2.2.5 for algorithmic fairness. This is a setting in which the associated distribution induced by DRO-type mechanisms deserves significantly more statistical investigation.

Next, we comment on dynamic DRO settings. This is an area that closely connects with what is known as distributionally robust reinforcement learning and it is in its infancy (see, e.g., Xu and Mannor (2010); Osogami (2012); Lim et al. (2013); Zhou et al. (2021); Backhoff et al. (2022); Si et al. (2023)). Even fundamental problems involving how to formulate associated distributionally robust Markov decision processes based on the sequentially available information for the agent and the adversary are significantly non-trivial (see Wang et al. (2023)). This area opens up a wide range of interesting questions for the statistics community. To give a sense of why DRO naturally offers a meaningful approach to estimation and optimization in these settings, note that in many situations of interest in stochastic control, there is a real possibility of facing unobserved (i.e. confounding) variables. This type of formulation is naturally posed as a so-called Partially Observed Markov Decision Process, which is challenging to study since it requires a history-dependent specification at every point in time. In these settings, the statistician can introduce a Markovian model (thus reducing the problem to a standard reinforcement learning environment) and instead use DRO to hedge against the model misspecification which has been introduced for tractability.

Finally, we finish our discussion by noting that the robust statistics perspective offered in this paper provides a useful point of view to connect and contrast DRO estimators and classical robust estimators. This perspective, characterized by the order in which the statistician and the adversary make their decision, was introduced in this paper primarily to motivate the fundamental differences in the nature of these types of robust estimators, The DRO estimator is pessimistic in nature because the statistician is at the mercy of an adversary that will change the out-of-sample environment. In robust statistics, hidden in the data lies useful information about the actual out-of-sample distribution - the adversary has made its move. Therefore, the statistician naturally could try to clean or rectify the contamination employed by the adversary thus leading to an optimistic approach.

Acknowledgments

References

An, Y. and Gao, R. (2021). Generalization bounds for (Wasserstein) robust optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 10382–10392. 20

Aravkin, A. and Davis, D. (2020). Trimmed statistical estimation via variance reduction. *Mathematics of Operations Research*, 45(1):292–322. 33

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Annals of Statistics*, 39(5):2766 – 2794. 31

Azizian, W., Iutzeler, F., and Malick, J. (2023). Regularization for wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:33. 12

Backhoff, J., Bartl, D., Beiglböck, M., and Wiesel, J. (2022). Estimating processes in adapted Wasserstein distance. *Annals of Applied Probability*, 32(1):529–550. 5, 35

Bartl, D., Drapeau, S., Ob lój, J., and Wiesel, J. (2021). Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proceedings of the Royal Society A*, 477(2256):20210176. 11

Bateni, A.-H. and Dalalyan, A. S. (2019). Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electronic Journal of Statistics*. 29

Bauschke, H. and Combettes, P. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York. 33

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806. 10, 18

Bennett, A., Kallus, N., Mao, X., Newey, W., Syrgkanis, V., and Uehara, M. (2023). Minimax instrumental variable regression and L_2 convergence guarantees without identification or closedness. *arXiv preprint arXiv:2302.05404*. 5

Bennouna, A. and Van Parys, B. (2022). Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*. 35

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media. 13

Bernholt, T. (2006). Robust estimators are hard to compute. Technical Report 2005,52, Universität Dortmund. 30

Bertsimas, D., Imai, K., and Li, M. L. (2022). Distributionally robust causal inference with observational data. *arXiv preprint arXiv:2210.08326*. 5

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. In *Advances in Neural Information Processing Systems*, volume 30. 31

Bhatia, K., Jain, P., and Kar, P. (2015). Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, volume 28. 31

Blanchet, J. and Kang, Y. (2017). Distributionally robust groupwise regularization estimator. In *Asian Conference on Machine Learning*, pages 97–112. PMLR. 6, 11, 15

Blanchet, J. and Kang, Y. (2020). Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods*, 5:1–33. 15

Blanchet, J. and Kang, Y. (2021). Sample out-of-sample inference based on Wasserstein distance. *Operations Research*, 69(3):985–1013. 15

Blanchet, J., Kang, Y., and Murthy, K. (2019a). Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857. 6, 9, 10, 18, 22

Blanchet, J., Kang, Y., Olea, J. L. M., Nguyen, V. A., and Zhang, X. (2023a). Dropout training is distributionally robust optimal. *Journal of Machine Learning Research*, 24(180):1–60. 6

Blanchet, J., Kang, Y., Zhang, F., He, F., and Hu, Z. (2021a). Doubly robust data-driven distributionally robust optimization. *Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis Methods and Tools*, 7:75–90. 15

Blanchet, J., Kuhn, D., Li, J., and Taskesen, B. (2023b). Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*. 12

Blanchet, J., Murthy, K., and Nguyen, V. A. (2021b). Statistical analysis of wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 227–254. INFORMS. 12, 18

Blanchet, J., Murthy, K., and Si, N. (2022a). Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, 109(2):295–315. 6, 19

Blanchet, J., Murthy, K., and Zhang, F. (2022b). Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research*, 47(2):1500–1529. 11, 15, 23

Blanchet, J. and Shapiro, A. (2023). Statistical limit theorems in distributionally robust optimization. *arXiv preprint arXiv:2303.14867*. 19

Blanchet, J., Zhang, F., Kang, Y., and Hu, Z. (2019b). A distributionally robust boosting algorithm. In *2019 Winter Simulation Conference*, pages 3728–3739. IEEE. 6, 9, 18

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799. 23

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pages 201–236. Elsevier. 26

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335. 26

Chao, G., Yuan, Y., and Weizhi, Z. (2019). Robust estimation via generative adversarial networks. In *International Conference on Learning Representations*. 28, 31

Charikar, M., Steinhardt, J., and Valiant, G. (2017). Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. 31

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics*, 46(5):1932 – 1960. 28, 29, 30, 31

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, pages 146–158. 7

Dapogny, C., Iutzeler, F., Meda, A., and Thibert, B. (2023). Entropy-regularized wasserstein distributionally robust shape and topology optimization. *Structural and Multidisciplinary Optimization*, 66(3):42. 12

Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612. 6, 7, 17, 20, 23

Depersin, J. and Lecué, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Annals of Statistics*, 50(1):511 – 536. 29

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *Annals of Statistics*, 44(6):2695 – 2725. 31

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019a). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864. 28, 30, 31

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., and Stewart, A. (2019b). Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR. 29, 31

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017a). Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR. 29, 30, 31

Diakonikolas, I. and Kane, D. (2023). *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press. 30

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*. 31

Diakonikolas, I., Kane, D. M., and Stewart, A. (2017b). Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 73–84. 31

Diakonikolas, I., Kane, D. M., and Stewart, A. (2018a). Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073. 29

Diakonikolas, I., Kane, D. M., and Stewart, A. (2018b). List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. 31

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969. 31

Donoho, D. L. (1994). Statistical Estimation and Optimal Recovery. *Annals of Statistics*, 22(1):238 – 270. 31

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20(4):1803–1827. 30

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184. 29

Donoho, D. L. and Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *Annals of Statistics*, 16(2):552–586. 27

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, III. *Annals of Statistics*, pages 668–701. 31

Duchi, J., Hashimoto, T., and Namkoong, H. (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664. 5

Duchi, J. and Namkoong, H. (2018). Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 19:1–55. 7, 8, 19, 20, 34

Duchi, J. C., Glynn, P. W., and Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969. 6, 23

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406. 7, 8, 20, 21

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738. 17

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139. 9

Gao, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139 – 1170. 29

Gao, R. (2022). Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*. 20

Gao, R., Chen, X., and Kleywegt, A. J. (2022). Wasserstein distributionally robust optimization and variation regularization. *Operations Research*. 6, 11

Gao, R. and Kleywegt, A. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655. 9

Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, volume 31. 6

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. 31

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*. 12

Gotoh, J.-y., Kim, M. J., and Lim, A. E. (2023). A data-driven approach to beating SAA out of sample. *Operations Research*. 33

Gül, G. and Zoubir, A. M. (2017). Minimax robust hypothesis testing. *IEEE Transactions on Information Theory*, 63(9):5572–5587. 6

Hampel, F. (1968). *Contributions to the Theory of Robust Estimation*. University of California. 26, 29

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42(6):1887 – 1896. 26, 29

He, S. and Lam, H. (2021). Higher-order expansion and bartlett correctability of distributionally robust optimization. *arXiv preprint arXiv:2108.05908*. 6

Hopkins, S. B. and Li, J. (2018). Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. 31

Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR. 6, 14

Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9. 7

Huber, P. (2004). *Robust Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley. 29

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101. 26, 27, 29

Huber, P. J. (1968). Robust confidence limits. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 10(4):269–278. 26

Huber, P. J. (1972). The 1972 Wald lecture robust statistics: A review. *Annals of Mathematical Statistics*, 43(4):1041 – 1067. 26

Iman, M., Arabnia, H. R., and Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40. 16

Jiang, N. and Xie, W. (2023). Distributionally favorable optimization: A framework for data-driven decision-making with endogenous outliers. *Available at Optimization Online*. 33

Johnson, D. and Preparata, F. (1978). The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107. 30

Joly, E., Lugosi, G., and Oliveira, R. I. (2017). On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11(1):440 – 451. 31

Kantorovich, L. V. and Rubinshtein, S. (1958). On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59. 13

Karmalkar, S., Klivans, A., and Kothari, P. (2019). List-decodable linear regression. In *Advances in Neural Information Processing Systems*, volume 32. 31

Klivans, A., Kothari, P. K., and Meka, R. (2018). Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR. 31

Kothari, P. K. and Steinhardt, J. (2017). Better agnostic clustering via relaxed tensor norms. *arXiv preprint arXiv:1711.07465*. 31

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS. 12

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science*, pages 665–674. IEEE Computer Society. 30, 31

Lam, H. (2016). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275. 6, 7, 8

Lam, H. (2018). Sensitivity to serial dependency of input processes: A robust approach. *Management Science*, 64(3):1311–1327. 6, 7

Lam, H. (2021). On the impossibility of statistically improving empirical optimization: A second-order stochastic dominance perspective. *arXiv preprint arXiv:2105.13419*. 34

Lam, H. and Zhou, E. (2017). The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters*, 45(4):301–307. 18, 22

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31. 20

Levy, B. C. and Nikoukhah, R. (2012). Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance. *IEEE Transactions on Automatic Control*, 58(3):682–695. 24

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. (2020). Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 8847–8860. 23

Li, J., Chen, C., and So, A. M.-C. (2020). Fast epigraphical projection-based incremental algorithms for Wasserstein distributionally robust support vector machine. In *Advances in Neural Information Processing Systems*, volume 33, pages 4029–4039. 23

Li, J., Huang, S., and So, A. M.-C. (2019). A first-order algorithmic framework for distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 32. 23

Li, J., Lin, S., Blanchet, J., and Nguyen, V. A. (2022). Tikhonov regularization is optimal transport robust under martingale constraints. In *Advances in Neural Information Processing Systems*, volume 35, pages 17677–17689. 6, 12

Lim, S. H., Xu, H., and Mannor, S. (2013). Reinforcement learning in robust Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 26. 5, 35

Liu, H. and Gao, C. (2019). Density estimation with contamination: minimax rates and theory of adaptation. *Electronic Journal of Statistics*, 13(2):3613 – 3653. 31

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2020). High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 411–421. PMLR. 29

Liu, Z. and Loh, P.-L. (2022). Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference: A Journal of the IMA*, 12(1):312–362. 27, 28, 31

Liu, Z., Van Parys, B. P., and Lam, H. (2023). Smoothed f-divergence distributionally robust optimization: Exponential rate efficiency and complexity-free calibration. *arXiv preprint arXiv:2306.14041*. 35

Lotidis, K., Bambos, N., Blanchet, J., and Li, J. (2023). Wasserstein distributionally robust linear-quadratic estimation under martingale constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 8629–8644. PMLR. 6, 12, 24, 25

Lugosi, G. and Mendelson, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783 – 794. 31

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*. 12

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308 – 2335. 31

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166. 9, 17, 20

Ng, A. Y. (2004). Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 78. 9

Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., and Mohajerin Esfahani, P. (2023). Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1):1–37. 6, 24, 25

Nguyen, V. A., Shafieezadeh Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. (2019). Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems*, volume 32. 33

Nguyen, V. A., Si, N., and Blanchet, J. (2020a). Robust Bayesian classification using an optimistic score ratio. In *International Conference on Machine Learning*, pages 7327–7337. PMLR. 24, 25

Nguyen, V. A., Zhang, F., Blanchet, J., Delage, E., and Ye, Y. (2020b). Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, 33:15232–15242. 14

Nguyen, V. A., Zhang, F., Blanchet, J., Delage, E., and Ye, Y. (2021). Robustifying conditional portfolio decisions via optimal transport. *arXiv preprint arXiv:2103.16451*. 14

Nguyen, V. A., Zhang, X., Blanchet, J., and Georghiou, A. (2020c). Distributionally robust parametric maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 7922–7932. 17

Norton, M., Takeda, A., and Mafusalov, A. (2017). Optimistic robust optimization with applications to machine learning. *arXiv preprint arXiv:1711.07511*. 33

Olea, J. L. M., Rush, C., Velez, A., and Wiesel, J. (2022). On the generalization error of norm penalty linear regression models. *arXiv preprint arXiv:2211.07608*. 35

Osogami, T. (2012). Robustness and risk-sensitivity in Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 25. 5, 35

Owen, A. B. (2001). *Empirical Likelihood*. CRC press. 7, 22

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607. 9

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627. 31

Raghavendra, P. and Yau, M. (2020). List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM. 31

Rahimian, H. and Mehrotra, S. (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85. 7

Rockafellar, R. (1974). *Conjugate Duality and Optimization*. Society for Industrial and Applied Mathematics. 33

Rockafellar, R. (1985). Extensions of subgradient calculus with applications to optimization. *Nonlinear Analysis: Theory, Methods & Applications*, 9(7):665–698. 33

Rockafellar, R. (1997). *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press. 33

Rockafellar, R. T. (1963). *Convex Functions and Dual Extremum Problems*. Phd thesis, University of Washington. 33

Rockafellar, R. T. (2023). Distributional robustness, stochastic divergences, and the quadrangle of risk. 7

Rothenhäusler, D. and Bühlmann, P. (2023). Distributionally robust and generalizable inference. *Statistical Science*, 38(4):527–542. 5

Royset, J. and Wets, R. (2022). *An Optimization Primer*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing. 33

Royset, J. O. (2021). Good and bad optimization models: Insights from Rockafellians. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 131–160. INFORMS. 33

Royset, J. O., Chen, L. L., and Eckstrand, E. (2023). Rockafellian relaxation in optimization under uncertainty: Asymptotically exact formulations. *arXiv preprint arXiv:2204.04762.* 33

Ruszczyński, A. and Shapiro, A. (2006). Optimization of risk measures. *Probabilistic and Randomized Methods for Design under Uncertainty*, pages 119–157. 7

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally robust neural networks. In *International Conference on Learning Representations.* 6, 14

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Birkhäuser. 9

Scarf, H. (1958). A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press. 6

Scheffe, H. and Tukey, J. W. (1944). A formula for sample sizes for population tolerance limits. *Annals of Mathematical Statistics*, 15(2):217. 26

Scheffe, H. and Tukey, J. W. (1945). Non-parametric estimation. I. Validation of order statistics. *Annals of Mathematical Statistics*, 16(2):187 – 192. 26

Shafieezadeh-Abadeh, S., Aolaritei, L., Dörfler, F., and Kuhn, D. (2023). New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv preprint arXiv:2303.03900.* 11

Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. (2019). Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68. 10

Shafieezadeh Abadeh, S., Nguyen, V. A., Kuhn, D., and Mohajerin Esfahani, P. M. (2018). Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems*, volume 31. 24, 25

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275. 20

Si, N., Murthy, K., Blanchet, J., and Nguyen, V. A. (2021). Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, pages 9649–9659. PMLR. 35

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2023). Distributionally robust batch contextual bandits. *Management Science.* 5, 35

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations.* 11, 12, 23

Staib, M. and Jegelka, S. (2019). Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, volume 32. 13

Steinhardt, J., Charikar, M., and Valiant, G. (2018). Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference*, volume 94, pages 45:1–45:21. 29, 30

Steinhardt, J., Koh, P. W., and Liang, P. (2017). Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, volume 30, page 3520–3532. 29

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, pages 595–620. 14

Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36(2):423–439. 12

Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. (2019). Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR. 31

Sun, Z. and Zou, S. (2021). A data-driven approach to robust hypothesis testing using kernel mmd uncertainty sets. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 3056–3061. IEEE. 6

Székely, G. J. (1989). Potential and kinetic energy in statistics. Lecture Notes, Budapest Institute of Technology (Technical University). 13

Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. (2020). A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*. 35

Taskesen, B., Yue, M.-C., Blanchet, J., Kuhn, D., and Nguyen, V. A. (2021). Sequential domain adaptation by synthesizing distributionally robust experts. In *International Conference on Machine Learning*, pages 10162–10172. PMLR. 6, 16

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288. 9

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 448–485. 26

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33(1):1–67. 26

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, page 523–531. 29

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 27

Van Parys, B. P., Esfahani, P. M., and Kuhn, D. (2021). From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402. 21

Villani, C. et al. (2009). *Optimal Transport: Old and New*, volume 338. Springer. 9

Wang, J., Gao, R., and Xie, Y. (2021). Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*. 12

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023). On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*. 5, 35

Watson, J. and Holmes, C. (2016). Approximate models and robust decisions. *Statistical Science*, 31(4):465–489. 23

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1):1–40. 16

Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46. 16

Wu, K., Ding, G. W., Huang, R., and Yu, Y. (2020). On minimax optimality of gans for robust mean estimation. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pages 4541–4551. PMLR. 31

Xu, H. and Mannor, S. (2010). Distributionally robust Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 23. 5, 35

Zalinescu, C. (2002). *Convex Analysis in General Vector Spaces*. G - Reference, Information and Interdisciplinary Subjects Series. World Scientific. 33

Zhang, X., Blanchet, J., Ghosh, S., and Squillante, M. S. (2022a). A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics*, pages 3794–3820. PMLR. 6, 16

Zhang, X., Blanchet, J., Marzouk, Y., Nguyen, V. A., and Wang, S. (2022b). Distributionally robust Gaussian process regression and Bayesian inverse problems. *arXiv preprint arXiv:2205.13111*. 6, 24, 25

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR. 5, 35

Zhu, B., Jiao, J., and Steinhardt, J. (2022). Generalized resilience and robust statistics. *Annals of Statistics*, 50(4):2256 – 2283. 27, 28, 31, 34

Zhu, J.-J., Jitkrittum, W., Diehl, M., and Schölkopf, B. (2021). Kernel distributionally robust optimization: Generalized duality theorem and stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR. 13

Zorzi, M. (2016). Robust Kalman filtering under model perturbations. *IEEE Transactions on Automatic Control*, 62(6):2902–2907. 24