# Outlier-Robust Wasserstein DRO

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Distributionally robust optimization (DRO) is an effective approach for data-driven decision-making in the presence of uncertainty. Geometric uncertainty due to sampling or localized perturbations of data points is captured by Wasserstein DRO (WDRO), which seeks to learn a model that performs uniformly well over a Wasserstein ball centered around the observed data distribution. However, WDRO fails to account for non-geometric perturbations such as adversarial outliers, which can greatly distort the Wasserstein distance measurement and impede the learned model. We address this gap by proposing a novel outlier-robust WDRO framework for decision-making under both geometric (Wasserstein) perturbations and non-geometric (total variation (TV)) contamination that allows an $\varepsilon$-fraction of data to be arbitrarily corrupted. We design an uncertainty set using a certain robust Wasserstein ball that accounts for both perturbation types. We derive minimax optimal excess risk bounds for this procedure that explicitly capture the Wasserstein and TV risks. We prove a strong duality result that enables efficient computation of our outlier-robust WDRO problem. When the loss function depends only on low-dimensional features of the data, we eliminate certain dimension dependencies from the risk bounds that are unavoidable in the general setting. Finally, we present experiments validating our theory on standard regression and classification tasks.

## 1 Introduction

The safety and effectiveness of various operations rely on making informed, data-driven decisions in uncertain environments. Distributionally robust optimization (DRO) has emerged as a powerful framework for decision-making in the presence of uncertainties. In particular, Wasserstein DRO (WDRO) captures uncertainties of geometric nature, e.g., due to sampling or localized (adversarial) perturbations of the data points. The WDRO problem is a two-player, zero-sum game between a learner (decision-maker), who chooses a decision $\theta \in \Theta$, and nature (adversary), who chooses a distribution $\nu$ from an ambiguity set defined as the $p$-Wasserstein ball of a prescribed radius around the observed data distribution $\tilde{\mu}$. Namely, WDRO is given by[1]

$$\inf_{\theta \in \Theta} \sup_{\nu : \mathsf{W}_p(\nu, \tilde{\mu}) \leq \rho} \mathbb{E}_{Z \sim \nu}[\ell(\theta, Z)], \tag{1}$$

whose solution $\hat{\theta} \in \Theta$ performs uniformly well over the Wasserstein ball with respect to (w.r.t.) the loss function $\ell$. WDRO has received considerable attention in many fields, including machine learning [2, 15, 35, 38, 49], estimation and filtering [26, 27, 36], and chance constraint programming [7, 45], among others.

In many practical scenarios, the observed data may be contaminated by non-geometric perturbations, such as adversarial outliers. Unfortunately, the WDRO problem from (1) is not suited for handling this

[1]Here, $\mathsf{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left( \int \|x - y\|^p d\pi(x, y) \right)^{1/p}$ is the $p$-Wasserstein metric between $\mu$ and $\nu$, where $\Pi(\mu, \nu)$ is the set of all their couplings.

Submitted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023. Do not distribute.

issue, as even a small fraction of outliers can greatly distort the $\mathsf{W}_p$ measurement and impede decision-making. In this work, we address this gap by proposing a novel outlier-robust WDRO framework that can learn well-performing decisions even in the presence of outliers. We couple it with a comprehensive theory of excess risk bounds, statistical guarantees, and computationally-tractable reformulations, as well as supporting numerical results.

### 1.1 Contributions

We consider a scenario where the observed data distribution $\tilde{\mu}$ is subject to both geometric (Wasserstein) perturbations and non-geometric (total variation (TV)) contamination, which allows an $\varepsilon$-fraction of data to be arbitrarily corrupted. Namely, if $\mu$ is the true (unknown) data distribution, then the Wasserstein perturbation maps it to some $\mu'$ with $\mathsf{W}_p(\mu', \mu) \leq \rho$, and the TV contamination step further produces $\tilde{\mu}$ with $\|\tilde{\mu} - \mu'\|_{\mathsf{TV}}$ (e.g., in the special case of the Huber model, $\tilde{\mu} = (1-\varepsilon)\mu' + \varepsilon\alpha$ where $\alpha$ is an arbitrary noise distribution). To enable robust decision-making under this model, we replace the Wasserstein ambiguity set in (1) with a ball w.r.t. the recently proposed outlier-robust Wasserstein distance $\mathsf{W}_p^\varepsilon$ [28, 29]. The $\mathsf{W}_p^\varepsilon$ distance is defined via a partial optimal transport (OT) problem (see (2) ahead) that first filters out the $\epsilon$-fraction of mass from the contaminated distribution that contributed most to the transportation cost, and then measures the $\mathsf{W}_p$ distance post-filtering. To obtain well-performing solutions for our WDRO problem, the $\mathsf{W}_p^\varepsilon$ ball is intersected with a set that encodes (necessary) moment assumptions on the uncorrupted data distribution.

We establish minimax optimal excess risk bounds for the decision $\hat{\theta}$ that solves the proposed outlier-robust WDRO problem. The bounds control the gap $\mathbb{E}[\ell(\hat{\theta}, Z)] - \mathbb{E}[\ell(\theta, Z)]$, where $Z \sim \mu$ follows the true data distribution, subject to regularity properties of $\ell(\theta, \cdot)$ for any arbitrary decision $\theta \in \Theta$. In turn, they imply that the learner can make effective decisions using outlier-robust WDRO based on the contaminated observation $\tilde{\mu}$, so long that there exists a (near) optimal $\theta$ with low variational complexity. The bounds capture this complexity using the Lipschitz or Sobolev seminorms of $\ell(\theta, \cdot)$ and clarify the distinct effect of each perturbation (Wasserstein versus TV) on the quality of the learned $\hat{\theta}$ solution. Moreover, they demonstrate notable improvements when the loss function depends only on $k$-dimensional linear features, for $k \ll d$. All of our bounds are shown to be minimax optimal, in that there exists a learning problem for which each is tight.

We then move to study the computational side of the problem, which may initially appear intractable due to non-convexity of the constraint set. We resolve this via a preprocessing step that computes a robust estimate of the mean [9] and replaces the original constraint set (that involves the true mean) with a version centered around the estimate. We adapt our excess risk bounds to this formulation and then prove a strong duality theorem. The dual form is reminiscent of the one for classical WDRO with adaptations reflecting the constraint to the clean distribution family and the partial transportation under $\mathsf{W}_p^\varepsilon$. Under additional convexity conditions on the loss, we further derive an efficiently-computable, finite-dimensional, convex reformulation. Using the developed machinery, we present experiments that validate our theory on simple regression tasks and demonstrate the superiority of the proposed approach over classical WRDO, when the observed data is contaminated.

### 1.2 Related Work

**Distributionally robust optimization.** The Wasserstein distance has emerged as a powerful tool for modeling uncertainty in the data generating distribution. It was first used to construct an ambiguity set around the empirical distribution in [30]. Recent advancements in convex reformulations and approximations of the WDRO problem, as discussed in [4, 14, 25], have brought notable computational advantages. Additionally, WDRO is linked to various forms of variation [1, 5, 12, 33] and Lipschitz [3, 6, 34] regularization, which contribute to its success in practice. Robust generalization guarantees can also be provided by WDRO via measure concentration argument or transportation inequalities [11, 21, 22, 41, 43, 44]. Several works have raised concerns regarding the sensitivity of standard DRO to outliers [16, 19, 48]. An attempt to address this was proposed in [46] using a refined risk function based on a family of $f$-divergences. This formulation aims to prevent DRO from overfitting to potential outliers but is not robust to geometric perturbations. Further, their risk bounds require a moment condition to hold uniformly over $\Theta$, in contrast to our bounds that depend only on a single (near) optimal $\theta$. We are able to address these limitations by setting a WDRO framework based on partial transportation. While partial OT has been previously used in the context of DRO problems, it was introduced to address stochastic programs with side information in [10] rather than to account for outlier robustness.

**Robust statistics.** The problem of learning from corrupted data corruptions dates back to [20]. Over the years, various robust and sample-efficient estimators, particularly for mean and scale parameters, have been developed in the robust statistics community; see [31] for a comprehensive survey. The theoretical computer science community, on the other hand, has focused on developing computationally efficient estimators that achieve optimal estimation rates in high dimensions [8, 9]. Recently, [48] developed a unified robust estimation framework based on minimum distance estimation that gives sharp population-limit and good finite-sample guarantees for mean and covariance estimation. Their analysis centers on a generalized resilience quantity, which will be also essential to our work. Also key to our analysis is the outlier-robust Wasserstein distance from [28, 29], which was shown to yield an optimal minimum distance estimate for robust distribution estimation under $\mathsf{W}_p$ loss.

## 2 Preliminaries

**Notation.** We consider Euclidean space $\mathbb{R}^d$ equipped with the $\ell_2$ norm $\|\cdot\|$. A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called $\alpha$-smooth if $\|\nabla f(z) - \nabla f(z')\| \leq \alpha\|z - z'\|$, for all $z, z' \in \mathbb{R}^d$. The perspective function of a lower semi-continuous (l.s.c.) and convex function $f$ is $P_f(x, \lambda) := \lambda f(x/\lambda)$ for $\lambda > 0$, with $P_f(x, \lambda) = \lim_{\lambda \to 0} \lambda f(x/\lambda)$ when $\lambda = 0$. The convex conjugate of $f$ is $f^*(y) := \sup_{x \in \mathbb{R}^d} y^\top x - f(x)$. The set of integers up to $n \in \mathbb{N}$ is denote by $[n]$; we also use the shorthand $[x]_+ = \max\{x, 0\}$. We write $\lesssim, \gtrsim, \asymp$ for inequalities/equality up to absolute constants.

We use $\mathcal{M}(\mathbb{R}^d)$ for the set of signed Radon measures on $\mathbb{R}^d$ equipped with the TV norm $\|\mu\|_{\mathsf{TV}} := \frac{1}{2}|\mu|(\mathcal{Z})$, and write $\mu \leq \nu$ for set-wise inequality. The class of Borel probability measures on $\mathbb{R}^d$ is denoted by $\mathcal{P}(\mathbb{R}^d)$. Write $\mathbb{E}_\mu[f(Z)]$ for expectation of $f(Z)$ with $Z \sim \mu$; when clear from the context, the random variable is dropped and we write $\mathbb{E}_\mu[f]$. Define $\mathcal{P}_p(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) : \inf_{z_0 \in \mathbb{R}^d} \mathbb{E}_\mu[\|Z - z_0\|^p] < \infty\}$. The push-forward of $f$ through $\mu \in \mathcal{P}(\mathbb{R}^d)$ is $f_\# \mu(\cdot) := \mu(f^{-1}(\cdot))$, and, for $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$, write $f_\# \mathcal{A} := \{f_\# \mu : \mu \in \mathcal{A}\}$. The $p$th order homogeneous Sobolev (semi)norm of continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}$ w.r.t. $\mu$ is $\|f\|_{\dot{H}^{1,p}(\mu)} := \mathbb{E}_\mu[\|\nabla f\|^p]^{1/p}$. Given $Z \sim \mu$ and an even convex, non-decreasing function $\psi : \mathbb{R} \to \mathbb{R}_+$ with $\psi(0) = 0$ and $\psi(x) \to \infty$ as $|x| \to \infty$, we define the Orlicz norm $\|Z\|_\psi = \sup\{\sigma \geq 0 : \sup_{\theta \in \mathbb{S}^{d-1}} \mathbb{E}[\psi(\theta^\top Z/\sigma)] \leq 1\}$.

**Classical and outlier-robust Wasserstein distances.** For $p \in [1, \infty)$, the *$p$-Wasserstein distance* between $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is $\mathsf{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \big(\mathbb{E}_\pi\big[\|X - Y\|^p\big]\big)^{1/p}$, where $\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi(\cdot \times \mathbb{R}^d) = \mu, \pi(\mathbb{R}^d \times \cdot) = \nu\}$ is the set of all their couplings. Some basic properties of $\mathsf{W}_p$ are (see, e.g., [32, 42]): (i) $\mathsf{W}_p$ is a metric on $\mathcal{P}_p(\mathbb{R}^d)$; (ii) the distance is monotone in the order, i.e., $\mathsf{W}_p \leq \mathsf{W}_q$ for $p \leq q$; and (iii) $\mathsf{W}_p$ metrizes weak convergence plus convergence of $p$th moments: $\mathsf{W}_p(\mu_n, \mu) \to 0$ if and only if $\mu_n \stackrel{w}{\to} \mu$ and $\int \|x\|^p d\mu_n(x) \to \int \|x\|^p d\mu(x)$.

To handle corrupted data, we define the *$\varepsilon$-outlier-robust $p$-Wasserstein distance*[2] between $\mu$ and $\nu$ by

$$\mathsf{W}_p^\varepsilon(\mu, \nu) := \inf_{\substack{\mu' \in \mathcal{P}(\mathbb{R}^d) \\ \|\mu' - \mu\|_{\mathsf{TV}} \leq \varepsilon}} \mathsf{W}_p(\mu', \nu) = \inf_{\substack{\nu' \in \mathcal{P}(\mathbb{R}^d) \\ \|\nu' - \nu\|_{\mathsf{TV}} \leq \varepsilon}} \mathsf{W}_p(\mu, \nu'). \tag{2}$$

The second equality is a useful consequence of Lemma 4 in [29].

**Robust statistics.** Resilience is a standard sufficient condition for population-limit robust statistics bounds. The *mean resilience* of a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\tau(\mu, \varepsilon) := \sup_{\mu' \leq \frac{1}{1-\varepsilon}\mu} \big\|\mathbb{E}_\mu[Z] - \mathbb{E}_{\mu'}[Z]\big\|,$$

and that of a family $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R})$ by $\tau(\mathcal{G}, \varepsilon) := \sup_{\mu \in \mathcal{G}} \tau(\mu, \varepsilon)$. The *$p$-Wasserstein resilience* of $\mu$ is given by

$$\tau_p(\mu, \varepsilon) := \sup_{\mu' \leq \frac{1}{1-\varepsilon}\mu} \mathsf{W}_p(\mu', \mu)$$

[2] While not a metric, $\mathsf{W}_p^\varepsilon$ is symmetric and satisfies an approximate triangle inequality ([29], Proposition 3).

and that of a family $\mathcal{G}$ by $\tau_p(\mathcal{G}, \varepsilon) \coloneqq \sup_{\mu \in \mathcal{G}} \tau_p(\mu, \varepsilon)$. When inference depends on $k$-dimensional projections, we use $\tau_{p,k}(\mu, \varepsilon) = \sup_{U \in \mathbb{R}^{k \times d}: UU^\top = I_k} \tau_p(U_\# \mu, \varepsilon)$ and $\tau_{p,k}(\mathcal{G}, \varepsilon) = \sup_{\mu \in \mathcal{G}} \tau_{p,k}(\mu, \varepsilon)$.

The relation between resilience and robust estimation is formalized in the following proposition.

**Proposition 1** (Robust estimation under resilience [29, 39])**.** *For any $\mu \in \mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$ and $\tilde{\mu} \in \mathcal{P}(\mathbb{R}^d)$ such that $\|\tilde{\mu} - \mu\|_{\mathsf{TV}} \le \varepsilon \le 1/2$, the minimum distance estimate $\hat{\mu} = \operatorname{argmin}_{\nu \in \mathcal{G}} \|\nu - \tilde{\mu}\|_{\mathsf{TV}}$ satisfies $\| \mathbb{E}_{\hat{\mu}}[Z] - \mathbb{E}_\mu[Z]\| \le 2\tau(\mathcal{G}, 2\varepsilon)$. Similarly, if $0 \le \varepsilon \le 0.49$ and $\mathsf{W}_p^\varepsilon(\tilde{\mu}, \mu) \le \rho$, then the minimum distance estimate $\hat{\mu} = \operatorname{argmin}_{\nu \in \mathcal{G}} \mathsf{W}_p^\varepsilon(\nu, \tilde{\mu})$ satisfies $\mathsf{W}_p(\hat{\mu}, \mu) \lesssim \rho + \tau_p(\mathcal{G}, 2\varepsilon)$.*[3]

In practice, we consider families $\mathcal{G}$ encoding tail bounds like bounded covariance or sub-Gaussianity:

$$\mathcal{G}_{\mathrm{cov}} \coloneqq \left\{\mu \in \mathcal{P}(\mathbb{R}^d) : \Sigma_\mu \preceq I_d\right\}, \quad \mathcal{G}_{\mathrm{subG}} \coloneqq \{\mu \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_\mu[e^{(\theta^\top Z)^2}] \le 2, \, \forall \theta \in \mathbb{S}^{d-1}\}.$$

**Proposition 2** (Resilience under standard tail bounds)**.** *Fixing $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $0 \le \varepsilon < 1$, we have*

$$\tau(\mathcal{G}_{\mathrm{cov}}, \varepsilon) \lesssim \sqrt{\varepsilon}, \qquad \tau_{p,k}(\mathcal{G}_{\mathrm{cov}}, \varepsilon) \lesssim \sqrt{k}\varepsilon^{\frac{1}{p} - \frac{1}{2}},$$
$$\tau(\mathcal{G}_{\mathrm{subG}}, \varepsilon) \lesssim \varepsilon\sqrt{\log \tfrac{1}{\varepsilon}}, \quad \tau_{p,k}(\mathcal{G}_{\mathrm{subG}}, \varepsilon) \lesssim \sqrt{k + p + \tfrac{1}{\varepsilon}}\varepsilon^{\frac{1}{p}}.$$

These bounds are computed in the proof of Theorem 5 in [29].

## 3 Outlier-robust WDRO

We perform stochastic optimization with respect to an unknown data distribution $\mu$, given access only to a corrupted version $\tilde{\mu}$. We first consider a Wasserstein perturbation mapping $\mu$ to $\mu'$ such that $\mathsf{W}_p(\mu, \mu') \le \rho$. Then we allow a TV $\varepsilon$-corruption taking $\mu'$ to $\tilde{\mu}$ with $\|\tilde{\mu} - \mu'\|_{\mathsf{TV}} \le \varepsilon$. Equivalently, we have $\mathsf{W}_p^\varepsilon(\tilde{\mu}, \mu) \le \rho$. Our full model is as follows.

**Setting A:** Fix a $p$-Wasserstein radius $\rho \ge 0$ and TV contamination level $\varepsilon \in [0, 0.49]$[4]. Let $\mathcal{L}$ be a family of real-valued loss functions on $\mathcal{Z}$, such that each $\ell \in \mathcal{L}$ is l.s.c. with $\sup_{z \in \mathcal{Z}} \frac{\ell(z)}{1 + \|z\|^p} < \infty$, and fix a class $\mathcal{G} \subseteq \mathcal{P}_p(\mathbb{R}^d)$ encoding distributional assumptions. We consider the following game:

(i) Nature selects a distribution $\mu \in \mathcal{G}$, unknown to the learner;

(ii) The learner observes $\tilde{\mu} \in \mathcal{P}(\mathbb{R}^d)$ with $\mathsf{W}_p^\varepsilon(\tilde{\mu}, \mu) \le \rho$ and selects decision $\hat{\ell} \in \mathcal{L}$;

(iii) The learner suffers excess risk $\mathbb{E}_\mu[\hat{\ell}] - \inf_{\ell \in \mathcal{L}} \mathbb{E}_\mu[\ell]$.

We seek a decision-making procedure for the learner which provides strong excess risk guarantees when $\ell_\star \coloneqq \operatorname{argmin}_{\ell \in \mathcal{L}} \mu(\ell)$[5] is appropriately "simple." To learn in this setting, we introduce the *$\varepsilon$-outlier-robust $p$-Wasserstein DRO problem*:

$$\inf_{\ell \in \mathcal{L}} \sup_{\nu \in \mathcal{G}:\, \mathsf{W}_p^\varepsilon(\tilde{\mu}, \nu) \le \rho} \mathbb{E}_\nu[\ell]. \tag{OR-WDRO}$$

Our results are most cleanly stated under the following structural assumptions.

**Assumption 1** (Bounded Orlicz norm)**.** The class $\mathcal{G} = \mathcal{G}_\psi(\sigma)$ consists of all distributions $Z \sim \mu \in \mathcal{P}(\mathbb{R}^d)$ for which $\|Z - \mathbb{E}[Z]\|_\psi \le \sigma$, where $\psi(x) = \sum_{i \ge 1} a_i x^{2i}$ is real analytic and even, with $a_i \ge 0$ for all $i \ge 1$ and $\psi(1) \le 2$.

**Assumption 2** ($\ell_\star$ depends on $k$-dimensional features)**.** The optimal loss function $\ell_\star$ can be decomposed as $\ell_\star = \underline{\ell} \circ A$ for an affine map $A : \mathbb{R}^d \to \mathbb{R}^k$ and some $\underline{\ell} : \mathbb{R}^k \to \mathbb{R}$.

Assumption 1 captures a variety of standard Orlicz norm bounds.

**Example 1.** *Taking $\sigma = 1$ and $\psi(x) = x^2$, we obtain the class $\mathcal{G}_{\mathrm{cov}} = \{\mu \in \mathcal{P}(\mathbb{R}^d) : \Sigma_\mu \preceq I_d\}$ of bounded covariance distributions, while $\psi(x) = e^{x^2} - 1$ gives the class $\mathcal{G}_{\mathrm{subG}}$ of 1-sub-Gaussian distributions.*

[3] If a minimizer does not exist for either problem, an infimizing sequence will achieve the same guarantee.

[4] While the choice of 0.49 is arbitrary, our bounds degrade as $\varepsilon \to 1/2$ (the optimal breakdown point).

[5] While our stated risk bounds will depend on $\ell_\star$, they extend naturally to approximate minimizers.

Assumption 2 is not necessarily restrictive, since one may always take $k = d$ and $A = I_d$. However, in many practical settings, all loss functions exhibit $k$-dimensional affine structure for $k \ll d$ (e.g., multi-linear regression). Our risk bounds are substantially stronger in this regime.

**Example 2** (Supervised learning with low-dimensional structure)**.** *Suppose that $\mathbb{R}^d = \mathbb{R}^{d_f} \times \mathbb{R}^{d_\ell}$ for a $d_f$ dimensional feature space and $d_\ell$ dimensional label space. Fix any hypothesis class $\mathcal{H}$ of $\mathbb{R}^{d_\ell}$-valued functions on $\mathbb{R}^{d_f}$ such that each $h \in \mathcal{H}$ can be written as $h(x) = \underline{h}(A(x))$, where $A : \mathbb{R}^d \to \mathbb{R}^{k-1}$ is affine and $\underline{h} : \mathbb{R}^{k-1} \to \mathbb{R}^{d_\ell}$ is Lipschitz. Let $L : \mathbb{R}^{d_\ell} \to \mathbb{R}$ be a l.s.c. loss function with bounded pth order growth, i.e., $\sup_{w \in \mathbb{R}^{d_\ell}} \frac{|L(w)|}{1+\|w\|^p} < \infty$. For example, we may take $L(w) = \|w\|^p$ or $L(w) = \mathbb{1}\{w \neq 0\}$. Then $\mathcal{L} = \{(x, y) \mapsto L(h(x) - y) : h \in \mathcal{H}\}$ satisfies Assumption 2. Indeed, for each $h = \underline{h} \circ A$ in $\mathcal{H}$, we can write $L(h(x) - y) = \underline{\ell}(B((x, y)))$, where $B : \mathbb{R}^d \to \mathbb{R}^k$ defined by $B((x, y)) = (Ax, y)$ is affine and $\underline{\ell}((Ax, y)) = L(\underline{h}(Ax) - y)$.*

Setting A considers the "population-limit" (i.e. no explicit model for sampling). We examine the performance of outlier-robust WDRO in this regime before turning to finite-sample risk bounds and computation. Proofs are provided in Supplement C.

### 3.1 Population-Limit Excess Risk Bounds

We now quantify the excess risk of decisions made using $\varepsilon$-outlier-robust $p$-WDRO.

**Theorem 1** (Population-limit excess risk bound)**.** *Consider Setting A under Assumptions 1 and 2. Let $\hat{\ell}$ minimize* (OR-WDRO)*. Then, the excess risk $\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star]$ is at most*

$$\begin{cases} 2\|\ell_\star\|_{\mathrm{Lip}}\big(\rho + \tau_{1,k}(\mathcal{G}, 2\varepsilon)\big), & p = 1, \ell_\star \text{ Lipschitz} \\ 2\|\ell_\star\|_{\dot{H}^{1,2}(\mu)}\big(\rho + \tau(\mathcal{G}, 2\varepsilon)\big) + \frac{44\alpha}{1-2\varepsilon}\big(\rho + \tau_{2,k}(\mathcal{G}, 2\varepsilon)\big)^2, & p = 2, \ell_\star\ \alpha\text{-smooth} \end{cases}.$$

Note that $\frac{1}{1-2\varepsilon} = O(1)$ since $\varepsilon \leq 0.49$. These bounds imply that the learner can make effective decisions when the optimal decision $\ell_\star$ has low variational complexity. In contrast, there are simple regression settings with TV corruption that drive the excess risk of standard WDRO to infinity. Moreover, the TV component of the risk is considerably smaller when $k \ll d$. In Table 1, we present tight risk bounds for OR-WDRO in a variety of environments. Each environment corresponds to a set of restrictions on $\mu$, the optimal loss function $\ell_\star$, and the order $p$ of the Wasserstein perturbation. The guarantees of OR-WDRO are minimax optimal for all settings considered (see Appendix C.2).

Our proof controls excess risk via the following two regularizers:

$$\Omega_{\mathsf{W}_p}(\ell_\star; \mu, \rho) := \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}^d) \\ \mathsf{W}_p(\nu, \mu) \leq \rho}} \mathbb{E}_\nu[\ell_\star] - \mathbb{E}_\mu[\ell_\star], \quad \Omega_{\mathsf{TV}}(\ell_\star; \mu, \mathcal{G}, \varepsilon) := \sup_{\substack{\nu \in \mathcal{G} \\ \|\nu - \mu\|_{\mathsf{TV}} \leq \varepsilon}} \mathbb{E}_\nu[\ell_\star] - \mathbb{E}_\mu[\ell_\star].$$

The $\mathsf{W}_p$ regularizer is well-studied and known to control excess risk for WDRO. When $\varepsilon = 0$, our proof recovers the known excess risk bound of $\Omega_{\mathsf{W}_p}(\ell_\star; \mu, \rho)$, and the theorem's bound is a standard upper bound on this quantity. The TV regularizer can similarly be shown to control excess risk for population-limit robust statistics (i.e. when $\rho = 0$), though, to the best of our knowledge, no previous work has derived explicit bounds on this quantity. The risk bound in Theorem 1 is a consequence of the following decomposition,

| | Environment | | | |
|---|---|---|---|---|
| | $\mu \in \mathcal{G}_{\mathrm{cov}}$<br>$\|\ell_\star\|_{\mathrm{Lip}} \leq L$<br>$p = 1$ | $\mu \in \mathcal{G}_{\mathrm{subG}}$<br>$\|\ell_\star\|_{\mathrm{Lip}} \leq L$<br>$p = 1$ | $\mu \in \mathcal{G}_{\mathrm{cov}}$<br>$\|\ell_\star\|_{\dot{H}^{1,2}(\mu)} \leq L$<br>$\ell_\star$ $\alpha$-smooth, $p = 2$ | $\mu \in \mathcal{G}_{\mathrm{subG}}$<br>$\|\ell_\star\|_{\dot{H}^{1,2}(\mu)} \leq L$<br>$\ell_\star$ $\alpha$-smooth, $p = 2$ |
| OR-WDRO<br>excess risk (OPT) | $L(\rho + \sqrt{k\varepsilon})$ | $L(\rho + \sqrt{k}\varepsilon)$ | $L(\rho + \sqrt{\varepsilon})$<br>$+\,\alpha(\rho^2 + k)$ | $L(\rho + \varepsilon)$<br>$+\,\alpha(\rho^2 + k\varepsilon)$ |

**Table 1:** Tight excess risk bounds for OR-WDRO in varied environments. Logarithmic factors omitted for ease of presentation; see Appendix C.2 for details.

$$\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star] \leq \Omega_{\mathsf{W}_p}(\ell_\star; \mu, 2\rho) + \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}^d) \\ \mathsf{W}_p(\nu,\mu) \leq \rho}} \Omega_{\mathsf{TV}}(\ell_\star; \nu, \mathcal{G}, 2\varepsilon),$$

whose components reveal the effect of each perturbation (viz. Wasserstein versus TV) on the quality of the decision. When $p = 1$, we rely on Kantorovich duality for $\mathsf{W}_1$, and, for $p = 2$, we use that $\ell$ can be well-approximated by its Taylor expansion about $Z \sim \mu$. Finally, we show that $\Omega_{\mathsf{TV}}$ depends only on a subproblem in $\mathbb{R}^k$. Notably, WDRO adapts automatically to the intrinsic dimensionality of $\ell_\star$ without requiring knowledge of $k$.

**Remark 1** (Comparison to recentered WDRO)**.** We note that non-trivial guarantees can be obtained by performing classic WDRO recentered around the minimum distance estimate $\hat{\mu} = \operatorname{argmin}_{\nu \in \mathcal{G}} \mathsf{W}_1^\varepsilon(\tilde{\mu}, \nu)$ with an expanded radius. For example, when $p = 1$, this estimate satisfies $\mathsf{W}_1(\mu, \hat{\mu}) \leq 2\rho + 2\tau_1(\mathcal{G}, 2\varepsilon)$, and so WDRO about $\hat{\mu}$ with this expanded radius incurs excess risk at most $O(\|\ell_\star\|_{\mathrm{Lip}}(\rho + \tau_1(\mathcal{G}, 2\varepsilon)))$. Ignoring the computational complexity of finding such a center $\hat{\mu}$ (which to the best of our knowledge, has not been established), the full-dimensional $\mathsf{W}_1$ resilience term $\tau_1(\mathcal{G}, \varepsilon)$ is substantially larger than the optimal $\tau_{1,k}(\mathcal{G}, \varepsilon)$ for $k \ll d$. We defer a comprehensive comparison against this MDE+WDRO approach for future work.

### 3.2 Finite-Sample Excess Risk Bounds

We next formalize a finite-sample model and provide statistical guarantees.

**Setting B:** Fix $\rho$, $\varepsilon$, $\mathcal{L}$, and $\mathcal{G}$ as in Setting A. We consider the following environment:

(i) Nature samples $Z_1, \ldots, Z_n$ i.i.d. from $\mu \in \mathcal{G}$, with empirical measure $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$;

(ii) Nature produces $\tilde{Z}_1, \ldots, \tilde{Z}_n$ with empirical measure $\tilde{\mu}_n$ such that $\mathsf{W}_p^\varepsilon(\tilde{\mu}_n, \hat{\mu}_n) \leq \rho$;

(iii) The learner observes $\tilde{\mu}_n$, selects $\hat{\ell} \in \mathcal{L}$, and suffers excess risk $\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star]$.

The learner is now tasked with selecting $\hat{\ell} \in \mathcal{L}$ given only $\tilde{\mu}_n$. The results from Section 3 apply immediately whenever $\rho \geq \rho_0 + \mathsf{W}_p(\mu, \hat{\mu}_n)$ with high probability.

**Proposition 3** (Choosing $\rho$)**.** *Consider Setting B under Assumption 2 with $\mathcal{G} = \mathcal{G}_{\mathrm{cov}}$. Assume $d \geq 3$. Take any $\hat{\ell} \in \mathcal{L}$ minimizing* (OR-WDRO) *when centered about $\tilde{\mu} = \tilde{\mu}_n$ with $p = 1$. Then the excess risk bounds of Theorem 1 hold with probability at least* 0.99 *so long as $\rho \geq \rho_0 + c\sqrt{d}n^{-\frac{1}{d}}$, where $c > 0$ is an absolute constant. If rather $\mathcal{G} = \mathcal{G}_{\mathrm{subG}}$, we have the same for both $p = 1$ and $p = 2$.*

While beyond the scope of this workshop submission, we note that this $n^{-1/d}$ rate may be improved to $n^{-1/k}$ under a Poincaré-type assumption on $\mu$ and a mild change to (OR-WDRO).

## 4 Tractable Reformulation and Computation

We now turn to computation. Due to space constraints, we focus on $\mathcal{G} = \mathcal{G}_{\mathrm{cov}}$ with $p = 1$ and $k = d$, though the approach below can be significantly extended. Initially, (OR-WDRO) may appear intractable, since $\mathcal{G}_{\mathrm{cov}}$ is non-convex when viewed as a subset of the cone $\mathcal{M}_+(\mathbb{R}^d)$. Moreover, enforcing membership to this class is non-trivial. To remedy these issues, we propose using a cheap preprocessing step to obtain a robust estimate $z_0 \in \mathbb{R}^d$ of the mean $\mathbb{E}_\mu[Z]$ and then optimizing over $\mathcal{G}_2(\sigma, z_0) := \left\{\nu \in \mathcal{P}(\mathbb{R}^d) : \sqrt{\mathbb{E}_\nu[\|Z - z_0\|^2]} \leq \sigma\right\}$, for some $\sigma > 0$. Finally, for technical reasons it is preferable to consider the one-sided robust distance $\mathsf{W}_p^\varepsilon(\mu \| \nu) := \inf_{\mu' \in \mathcal{P}(\mathbb{R}^d) : \mu' \leq \frac{1}{1-\varepsilon}\mu} \mathsf{W}_p(\mu', \nu)$. All together, we propose solving the simplified problem

$$\inf_{\ell \in \mathcal{L}} \sup_{\nu \in \mathcal{G}_2(\sigma, z_0) : \, \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \leq \rho} \mathbb{E}_\nu[\ell], \tag{3}$$

which admits risk bounds matching Theorem 1.

**Proposition 4** (Risk bound for simplified problem)**.** *Consider Setting B with $p = 1$ and $\mathcal{G} = \mathcal{G}_{\mathrm{cov}}$. Fix $z_0 \in \mathcal{Z}$ such that $\|z_0 - \mathbb{E}_\mu[Z]\| \leq \rho_0 + O(\sqrt{d})$, and take $\hat{\ell}$ minimizing* (3) *with $\rho = \rho_0 + \mathsf{W}_1(\hat{\mu}_n, \mu) + O(\sqrt{d\varepsilon})$ and $\sigma = \rho_0 + O(\sqrt{d})$. Then, excess risk is bounded by*

$$\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star] \lesssim \|\ell_\star\|_{\mathrm{Lip}}\big(\rho_0 + \mathsf{W}_1(\hat{\mu}_n, \mu) + \sqrt{d\varepsilon}\,\big).$$

The proof uses the fact that $\mu \in \mathcal{G}_{\text{cov}}$ implies $\mu \in \mathcal{G}_2\big(\sqrt{d} + \|z_0 - \mathbb{E}_\mu[Z]\|, z_0\big)$, along with the resilience bound $\tau_1\big(\mathcal{G}_2(\sigma, z_0), \varepsilon\big) \lesssim \sqrt{d\varepsilon}$. For efficient computation, we must specify a robust mean estimation algorithm to obtain $z_0$ and a procedure for solving (3). For the former, we show that the popular iterative filtering algorithm [9] works even with adversarial Wasserstein perturbations.

**Proposition 5** (Robust mean estimation)**.** *Consider Setting B with $\mathcal{G} = \mathcal{G}_{\text{cov}}$, $p = 1$, and $\varepsilon \le 1/12$. For $n = \Omega(d \log(d)/\varepsilon)$, there exists an iterative filtering algorithm which takes $\tilde{\mu}_n$ as input, runs in time $\tilde{O}(nd^2)$, and outputs $z_0 \in \mathbb{R}^d$ such that $\|z_0 - \mathbb{E}_\mu[Z]\| \lesssim \rho_0 + \sqrt{\varepsilon}$ with probability at least 0.99.*

It is not immediately clear that iterative filtering should still work under $\mathsf{W}_1^\varepsilon$ perturbations (compared the $\mathsf{TV}$ corruptions it was designed for), since the $\mathsf{W}_1$ step can arbitrarily increase the initial covariance bound. Fortunately, we show that trimming a small fraction of samples mitigates this potential increase. With some effort omitted from this submission, we expect that the upper bound on $\varepsilon$ can be replaced with any constant less than $1/2$, and that the running time can be improved to $\tilde{O}(nd)$.

We next show that that the inner maximization problem of (3) can be simplified to a minimization problem involving only three scalars provided the following assumption holds.

**Assumption 3** (Slater condition I)**.** Given the distribution $\tilde{\mu}_n$ and the fixed point $z_0$, there exists $\nu_0 \in \mathcal{P}(\mathcal{Z})$ such that $\mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu_0) < \rho$ and $\mathbb{E}_{\nu_0}[\|Z - z_0\|^2] < \sigma^2$. Additionally, we require $\rho > 0$.

Notice that Assumption 3 indeed holds for $\nu_0 = \mu$ as applied in Proposition 4.

**Proposition 6** (Strong duality)**.** *Under Assumption 3, for any $\ell \in \mathcal{L}$ and $z_0 \in \mathbb{R}^d$, we have*

$$\sup_{\nu \in \mathcal{G}_2(\sigma, z_0) :\, \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \le \rho} \mathbb{E}_\nu[\ell] = \inf_{\substack{\lambda_1, \lambda_2 \in \mathbb{R}_+ \\ \alpha \in \mathbb{R}}} \lambda_1 \sigma^2 + \lambda_2 \rho^p + \alpha + \frac{1}{1 - \varepsilon} \mathbb{E}_{\tilde{\mu}_n} \left[ \overline{\ell}(\cdot\,; \lambda_1, \lambda_2, \alpha) \right], \tag{4}$$

*where $\overline{\ell}(z; \lambda_1, \lambda_2, \alpha) := \sup_{\xi \in \mathbb{R}^d} \left[ \ell(\xi) - \lambda_1 \|\xi - z_0\|^2 - \lambda_2 \|\xi - z\|^p - \alpha \right]_+$.*

The minimization problem over $(\lambda_1, \lambda_2, \alpha)$ is an instance of stochastic convex optimization, where the expectation of the implicit function $\overline{\ell}$ is taken w.r.t. the contaminated empirical measure $\tilde{\mu}_n$. In contrast, the dual reformulation for classical WDRO only involves $\lambda_2$ and takes the expectation of the implicit function $\underline{\ell}(z; \lambda_2) := \sup_{\xi \in \mathbb{R}^d} \ell(\xi) - \lambda_2 \|\xi - z\|^p$ w.r.t. $\tilde{\mu}_n$. The additional $\lambda_1$ variable above is introduced to account for the clean family $\mathcal{G}_2(\sigma, z_0)$, and the use of partial transportation under $\mathsf{W}_p^\varepsilon$ results in the introduction of the operator $[\cdot]_+$ and the decision variable $\alpha$.

**Remark 2** (Connection to conditional value at risk (CVaR))**.** The CVaR of a Borel measurable loss function $\ell$ acting on a random vector $Z \sim \mu \in \mathcal{P}(\mathbb{R}^d)$ with risk level $\varepsilon \in (0, 1)$ is defined as

$$\mathrm{CVaR}_{1-\varepsilon, \mu}[\ell(Z)] = \inf_{\alpha \in \mathbb{R}} \alpha + \frac{1}{1 - \varepsilon} \mathbb{E}_{Z \sim \mu} \left[ [\ell(Z) - \alpha]_+ \right].$$

CVaR is also known as expected shortfall and is equivalent to the conditional expectation of $\ell(Z)$, given that it is above an $\varepsilon$ threshold. This concept is often used in finance to evaluate the market risk of a portfolio. With this definition, the result of Proposition 6 can be written as

$$\sup_{\substack{\nu \in \mathcal{G}_2(\sigma, z_0): \\ \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \le \rho}} \mathbb{E}_\nu[\ell] = \inf_{\lambda_1, \lambda_2 \in \mathbb{R}_+} \lambda_1 \sigma^2 + \lambda_2 \rho^p + \mathrm{CVaR}_{1-\varepsilon, \tilde{\mu}_n} \left[ \sup_{\xi \in \mathbb{R}^d} \ell(\xi) - \lambda_1 \|\xi - z_0\|^2 - \lambda_2 \|\xi - Z\|^p \right].$$

When $\varepsilon \to 0$ and $\sigma \to \infty$, whence CVaR reduces to expected value and the constrained class $\mathcal{G}_2(\sigma, z_0)$ becomes the whole space of distributions $\mathcal{P}(\mathbb{R}^d)$), the dual formulation above reduces to that of classical WDRO [13].

Evaluating $\overline{\ell}$, however, requires solving a maximization problem, which could be in itself challenging. To overcome this, we impose additional convexity assumptions, standard for WDRO [25, 33].

**Assumption 4** (Convexity condition)**.** The loss $\ell$ is a pointwise maximum of finitely many concave functions, i.e., $\ell(\xi) = \max_{j \in [J]} \ell_j(\xi)$, for some $J \in \mathbb{N}$, where $\ell_j$ is real-valued, l.s.c., and concave.

**Theorem 2** (Convex reformulation)**.** *Under Assumption 3, for any $\ell \in \mathcal{L}$ satisfying Assumption 4 and $z_0 \in \mathbb{R}^d$, we have $\sup_{\nu \in \mathcal{G}_q(\sigma, z_0) :\, \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \le \rho} \mathbb{E}_\nu[\ell] = \inf \lambda_1 \sigma^2 + \lambda_2 \rho^p + \alpha + \frac{1}{n(1-\varepsilon)} \sum_{i \in [n]} s_i$, where the right-hand side is optimized over the constraint set*

$$\begin{cases} \lambda_1, \lambda_2 \in \mathbb{R}_+,\ \alpha \in \mathbb{R},\ s, \tau_{ij} \in \mathbb{R}_+^n,\ \zeta_{ij}^\ell, \zeta_{ij}^{\mathcal{G}}, \zeta_{ij}^{\mathsf{W}}, \in \mathbb{R}^d, & \forall i \in [n], \forall j \in [J] \\ s_i \ge (-\ell_j)^*(\zeta_{ij}^\ell) + z_0^\top \zeta_{ij}^{\mathcal{G}} + \tau_{ij} + \tilde{Z}_i^\top \zeta_{ij}^{\mathsf{W}} + P_h(\zeta_{ij}^{\mathsf{W}}, \lambda_2) - \alpha, & \forall i \in [n], \forall j \in [J] \\ \zeta_{ij}^\ell + \zeta_{ij}^{\mathcal{G}} + \zeta_{ij}^{\mathsf{W}} = 0,\ \|\zeta_{ij}^{\mathcal{G}}\|^2 \le \lambda_1 \tau_{ij}, & \forall i \in [n], \forall j \in [J], \end{cases}$$

*and $P_h$ is the perspective function of $h$ defined by*

$$h(\zeta) := \begin{cases} \chi_{\{z \in \mathbb{R}^d :\, \|z\| \leq 1\}}(\zeta), & p = 1 \\ \frac{(p-1)^{p-1}}{p^p} \|\zeta\|^{\frac{p}{p-1}}, & p > 1 \end{cases}. \tag{5}$$

The minimization problem in Theorem 2 is a finite-dimensional convex program.

## 5 Experiments

Lastly, we implement our tractable reformulation and validate our excess risk bounds. Fixing $\mathbb{R}^d = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^{d-1} \times \mathbb{R}$, we focus on linear regression with the mean absolute deviation loss, i.e., $\mathcal{L} = \{\ell_\theta(x, y) = |\theta^\top x - y| : \theta \in \mathbb{R}^d\}$. See Supplement E for additional experiments treating classification and multivariate regression, along with full code and experimental details. The experiments below were run in 30 minutes on an M1 MacBook Air with 16GB RAM.

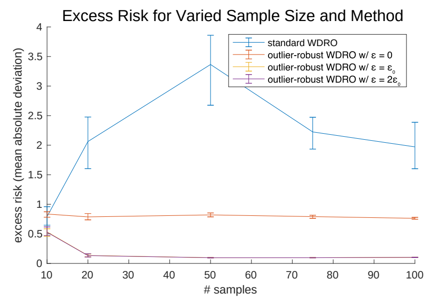


**Figure 1:** Excess risk of standard WDRO and several forms of outlier-robust WDRO for linear regression under $\mathsf{W}_p$ and TV corruptions, with varied sample size.

Let $\mathcal{Z} = (\mathbb{R}^d, \|\cdot\|_2)$ for $d \geq 2$ and fix $\rho = 0.1$, $\varepsilon_0 = 0.05$. We take $\theta_0, \theta_1 \in \mathbb{S}^{d-2}$ with $\|\theta_0 - \theta_1\|_2 \leq \rho d^{-1/2}$. Letting $X \sim \mathcal{N}(0, I_{d-1})$, we consider clean data $(X, \theta_0^\top X) \sim \mu$. The corrupted data $(\tilde{X}, \tilde{Y}) \sim \tilde{\mu}$ satisfies $(\tilde{X}, \tilde{Y}) = (X, \theta_1^\top X)$ with probability $1 - \varepsilon_0$ and $(\tilde{X}, \tilde{Y}) = (20X, -20\theta_1^\top X)$ with probability $\varepsilon_0$, so that $\mathsf{W}_p^{\varepsilon_0}(\tilde{\mu}\|\mu) \leq \rho$. In Figure 1 (top), we fix $d = 10$ and compare the excess risk $\mathbb{E}_\mu[\ell_{\hat{\theta}}] - \mathbb{E}_\mu[\ell_{\theta_0}]$ of standard WDRO ($\varepsilon = 0$, no moment constraints) and OR-WDRO with $\varepsilon \in \{0, \varepsilon_0, 2\varepsilon_0\}$, as described by Proposition 4 and implemented via Theorem 2. The results are averaged over $T = 20$ runs for sample size $n \in \{10, 20, 50, 75, 100\}$. Implementation of the reformulation was performed in MATLAB using the YALMIP toolbox [24] and SeDuMi solver [40].

## 6 Concluding Remarks

In this work, we have introduced a novel framework for outlier-robust WDRO that allows for both geometric and non-geometric perturbations of the observed data distribution, as captured by $\mathsf{W}_p$ and TV, respectively. We provided minimax-optimal excess risk bounds and strong duality results that enable efficient computation via convex reformulation. The full version of this paper will include refined statistical guarantees, tractable convex reformulations for distribution families beyond $\mathcal{G}_{\text{cov}}$ and for $k \ll d$, and a detailed discussion of parameter tuning. Overall, our approach enables principled, data-driven decision-making in realistic scenarios where observations may be subject to adversarial contamination by outliers.

## References


[1] D. Bartl, S. Drapeau, J. Obloj, and J. Wiesel. Robust uncertainty sensitivity analysis. *arXiv preprint arXiv:2006.12022*, 4, 2020.

[2] J. Blanchet, P. W. Glynn, J. Yan, and Z. Zhou. Multivariate distributionally robust convex regression under absolute error loss. In *Advances in Neural Information Processing Systems*, pages 11817–11826, 2019.

[3] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

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

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

[6] R. Chen and I. C. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(1):517–564, 2018.

[7] Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over Wasserstein balls. *Operations Research (Forthcoming)*, 2022.

[8] Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

[9] I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[10] A. Esteban-Pérez and J. M. Morales. Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, 195(1-2):1069–1105, 2022.

[11] R. Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022.

[12] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research (Forthcoming)*, 2022.

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

[14] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research (Forthcoming)*, 2023.

[15] R. Gao, L. Xie, Y. Xie, and H. Xu. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.

[16] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[17] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

[18] S. Hopkins, J. Li, and F. Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912, 2020.

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

[20] P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[21] Y. Kwon, W. Kim, J.-H. Won, and M. C. Paik. Principled learning method for Wasserstein distributionally robust optimization with local perturbations. In *International Conference on Machine Learning*, pages 5567–5576, 2020.

[22] J. Lee and M. Raginsky. Minimax statistical learning with Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 2687–2696, 2018.

[23] J. Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. 2020.

[24] J. Lofberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *IEEE international conference on robotics and automation*, pages 284–289. IEEE, 2004.

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

[26] V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Operations Research*, 70(1):490–515, 2022.

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

[28] S. Nietert, R. Cummings, and Z. Goldfeld. Outlier-robust optimal transport with applications to generative modeling and data privacy. In *Theory and Practice of Differential Privacy Workshop at ICML (TPDP-2021)*, July 2021.

[29] S. Nietert, R. Cummings, and Z. Goldfeld. Robust estimation under the Wasserstein distance. *arXiv preprint arXiv:2302.01237*, 2023.

[30] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[31] E. M. Ronchetti and P. J. Huber. *Robust Statistics*. John Wiley & Sons Hoboken, 2009.

[32] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

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

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

[35] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[36] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems*, pages 8474–8483, 2018.

[37] A. Shapiro. On duality theory of conic linear problems. In M. Á. Goberna and M. A. López, editors, *Semi-Infinite Programming*, pages 135–165. Kluwer Academic Publishers, 2001.

[38] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[39] J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Innovations in Theoretical Computer Science Conference*, volume 94, 2018.

[40] J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11(1-4):625–653, 1999.

[41] Z. Tu, J. Zhang, and D. Tao. Theoretical analysis of adversarial learning: A minimax approach. In *Advances in Neural Information Processing Systems*, pages 12280–12290, 2019.

[42] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

[43] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5339–5349, 2018.

[44] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595, 2019.

[45] W. Xie. On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming*, 186(1):115–155, 2021.

[46] R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar. DORO: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355, 2021.

[47] J. Zhen, D. Kuhn, and W. Wiesemann. Mathematical foundations of robust and distributionally robust optimization. *arXiv preprint arXiv:2105.00760*, 2021.

[48] B. Zhu, J. Jiao, and J. Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256 – 2283, 2022.

[49] S. Zhu, L. Xie, M. Zhang, R. Gao, and Y. Xie. Distributionally robust weighted k-nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 29088–29100, 2022.

# A Preliminary Results

We first recall and prove some basic facts about $\mathsf{W}_p^\varepsilon$, Orlicz norms, projected moment bounds, and resilience. To start, we prove that $\mathsf{W}_p^\varepsilon$ is equivalent to a certain partial OT problem.

**Lemma 1** ($\mathsf{W}_p^\varepsilon$ as partial OT)**.** *For any $\varepsilon \in [0,1]$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we have*

$$\mathsf{W}_p^\varepsilon(\mu,\nu) = (1-\varepsilon)^{1/p} \inf_{\substack{\mu',\nu' \in \mathcal{P}(\mathbb{R}^d) \\ \mu' \le \frac{1}{1-\varepsilon}\mu,\, \nu' \le \frac{1}{1-\varepsilon}\nu}} \mathsf{W}_p(\mu',\nu')$$

*Proof.* Write $\widetilde{\mathsf{W}}_p^\varepsilon(\mu,\nu)$ for the RHS. Rescaling, we have

$$\widetilde{\mathsf{W}}_p^\varepsilon(\mu,\nu) = \inf_{\substack{\mu',\nu' \in (1-\varepsilon)\mathcal{P}(\mathbb{R}^d) \\ \mu' \le \mu,\, \nu' \le \nu}} \mathsf{W}_p(\mu',\nu'), \tag{6}$$

matching the definition for robust OT in [29]. By their triangle inequality (Proposition 3 therein), we have for any $\tilde{\mu} \in \mathcal{P}(\mathbb{R}^d)$ with $\|\tilde{\mu} - \mu\|_{\mathsf{TV}} \le \varepsilon$ that

$$\widetilde{\mathsf{W}}_p^\varepsilon(\mu,\nu) \le \widetilde{\mathsf{W}}_p^\varepsilon(\mu,\tilde{\mu}) + \mathsf{W}_p(\tilde{\mu},\nu) = \mathsf{W}_p(\tilde{\mu},\nu).$$

Infimizing over $\tilde{\mu}$, we find that $\widetilde{\mathsf{W}}_p^\varepsilon(\mu,\nu) \le \mathsf{W}_p^\varepsilon$. For the opposite direction, consider any feasible $\mu', \nu'$ for (6), and let $\tilde{\mu} = \mu' + (\nu - \nu')$. By construction, we have $\|\tilde{\mu} - \mu\|_{\mathsf{TV}} \le \varepsilon$. Moreover, by Lemma 5 of [29], we have $\mathsf{W}_p(\tilde{\mu},\nu) \le \mathsf{W}_p(\mu',\nu')$. Thus, $\mathsf{W}_p^\varepsilon(\mu,\nu) \le \mathsf{W}_p(\mu',\nu')$, and infimizing over $\mu', \nu'$ gives the lemma. □

Next, we address the simple setting of Orlicz norms for constant random variables.

**Lemma 2** (Orlicz norm of constant random variable)**.** *For any constant random variable $Z = z \in \mathbb{R}^d$, and any Orlicz function $\psi$ satisfying the conditions in Assumption 1, we have $\|Z\|_\psi \le 2\|z\|$.*

*Proof.* For each $\theta \in \mathbb{S}^{d-1}$, we bound

$$\begin{aligned}
\mathbb{E}\left[\psi\left(\frac{|\theta^\top Z|}{2\|z\|}\right)\right] &= \mathbb{E}\left[\psi\left(\frac{|\theta^\top z|}{2\|z\|}\right)\right] \\
&\le \mathbb{E}[\psi(1/2)] \\
&= \sum_{i \ge 1} a_i 2^{-2i} \\
&\le \sum_{i \ge 1} 2^{-2i} \max_{j \ge 1} a_j \\
&< 1/2 \cdot \psi(1) < 1.
\end{aligned}$$

Thus $\|Z\|_\psi \le 2\|z\|$, as desired. □

Now, we introduce some notation and basic comparison results for projected moment bounds. Given $Z \sim \mu \in \mathcal{P}(\mathbb{R}^d)$, $r \in [d]$, and $q \ge 1$, we write $\sigma_{q,r}(\mu) \coloneqq \mathsf{W}_{q,r}(\mu, \delta_{\mathbb{E}[Z]})$ and $\sigma_q(\mu) = \sigma_{q,d}(\mu)$. This quantity captures the largest centered $q$th moment of an $r$-dimensional projection of $\mu$.

**Lemma 3** (Projected moment comparison)**.** *Fix $\mu \in \mathcal{P}(\mathbb{R}^d)$, dimension $r \in [d]$, and power $q \ge 1$. We then have $\sigma_{q,r}(\mu) \le \mathbb{E}[|S_1|^q]^{-1/q}\sigma_{q,1}(\mu)$, where $S \sim \mathrm{Unif}(\mathbb{S}^{r-1})$.*

*Proof.* Assume without loss of generality that $Z \sim \mu$ has mean zero. Fix any $U \in \mathbb{R}^{r \times d}$ with $UU^\top = I_r$, and let $S \sim \mathrm{Unif}(\mathbb{S}^{r-1})$. We then bound

$$\begin{aligned}
\sigma_{q,1}(\mu)^q &\ge \sigma_{q,1}(U_\# \mu)^q \\
&= \sup_{\theta \in \mathbb{S}^{r-1}} \mathbb{E}[|\theta^\top U Z|^q] \\
&\ge \mathbb{E}[|S^\top U Z|^q] \\
&= \mathbb{E}[|S_1|^q]\,\mathbb{E}[\|UZ\|^q],
\end{aligned}$$

where the last equality holds by rotational symmetry. Taking a supremum over $U$ gives the lemma. □

**Lemma 4** (Moment centering)**.** *Fix $\mu \in \mathcal{P}(\mathbb{R}^d)$, dimension $r \in [d]$, and power $q \geq 1$. Then for any $z \in \mathbb{R}^d$, we have $\sigma_{q,r}(\mu) \leq 2\mathsf{W}_{q,r}(\mu, \delta_z)$.*

*Proof.* Taking $Z \sim \mu$, we compute

$$\begin{aligned}\sigma_{q,r}(\mu) &= \mathsf{W}_{q,r}(\mu, \delta_{\mathbb{E}[Z]})\\ &\leq \mathsf{W}_{q,r}(\mu, \delta_z) + \mathsf{W}_{q,r}(\delta_z, \delta_{\mathbb{E}[Z]})\\ &\leq 2\mathsf{W}_{q,r}(\mu, \delta_z),\end{aligned}$$

where the final inequality follows by Jensen's inequality. □

Next, we recall two useful results for mean resilience.

**Lemma 5** (Mean resilience under moment bounds)**.** *For any $\varepsilon \in [0, 1)$ and $\mu \in \mathcal{P}(\mathbb{R})$, we have $\tau(\mu, \varepsilon) \leq \inf_{q \geq 1} \sigma_{q,1}(\mu)\varepsilon^{1-1/q}(1-\varepsilon)^{-1}$.*

*Proof.* This follows from Lemma E.2 of [48], using the Orlicz function $\psi(t) = t^q$ for each $q \geq 1$. □

**Lemma 6** (Mean resilience for large $\varepsilon$, [39], Lemma 10)**.** *For any $\varepsilon \in (0, 1)$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, we have $\tau(\mu, 1-\varepsilon) = \frac{1-\varepsilon}{\varepsilon}\tau(\mu, \varepsilon)$.*

Finally, we turn to Wasserstein resilience.

**Lemma 7** ($\mathsf{W}_2$ resilience and even moment bounds)**.** *Fix $\varepsilon \in (0, 1)$ and family $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$ satisfying Assumption 1. We then have*

$$\frac{1}{8}(1-\varepsilon)\tau_2(\mathcal{G}, \varepsilon)^2 \leq \sup_{\mu \in \mathcal{G}} \inf_{i \in \mathbb{N}_{>0}} \sigma_{2i}(\mu)^2 \varepsilon^{1-1/i} \leq 2\tau_2(\mathcal{G}, \varepsilon)^2.$$

*Proof.* Fix $\mu \in \mathcal{G}$ with mean zero. By the proof of [29, Theorem 2], we have

$$\begin{aligned}\tau_2(\mu, \varepsilon)^2 &\leq 4(1-\varepsilon)^{-1} \inf_{i>1} \sigma_{2i}(\mu)^2\, \mathbb{E}[\|Z\|^{2i}]^{1/i}\varepsilon^{1-1/i} + 4\varepsilon\sigma_2(\mu)^2\\ &\leq 8(1-\varepsilon)^{-1}\sigma_{2i}(\mu)^2\varepsilon^{1-1/i}.\end{aligned}$$

Taking a supremum over $\mu \in \mathcal{G}$ gives the first inequality (noting that the centering assumption is without loss of generality since $\mathcal{G}$ is closed under translations). For the second inequality, we again take mean zero $Z \sim \mu \in \mathcal{G}$. Then, by Assumption 1, we have $\sup_{\theta \in \mathbb{S}^{d-1}} \mathbb{E}_\mu[\psi(|\theta^\top Z|)] \leq 1$, where $\psi(x) = \sum_{i \geq 1} a_i x^{2i}$. Taking $S \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, we bound

$$\begin{aligned}1 &\geq \sup_{\theta \in \mathbb{S}^{d-1}} \mathbb{E}[\psi(|\theta^\top Z|)]\\ &= \sup_{\theta \in \mathbb{S}^{d-1}} \sum_{i \geq 1} a_i\, \mathbb{E}[|\theta^\top Z|^{2i}]\\ &\geq \sup_{\theta \in \mathbb{S}^{d-1}, i \geq 1} a_i\, \mathbb{E}[|\theta^\top Z|^{2i}]\\ &= \sup_{i \geq 1} a_i \sup_{\theta \in \mathbb{S}^{d-1}} \mathbb{E}[|\theta^\top Z|^{2i}]\\ &= \sup_{i \geq 1} a_i\, \sigma_{2i,1}(\mu)^{2i}\\ &= \sup_{i \geq 1} a_i\, \mathbb{E}[S_1^{2i}]\sigma_{2i}(\mu)^{2i},\end{aligned}$$

where the last equality follows by Lemma 3.

Next, we define the modified Orlicz functions

$$\phi(x) := \mathbb{E}[\psi(|S_1|\sqrt{x})] = \sum_{i \geq 1} a_i\, \mathbb{E}[S_1^{2i}]x^i, \qquad \underline{\phi}(x) = \sup_{i \geq 1} a_i\, \mathbb{E}[S_1^{2i}]x^i.$$

By design, we have

$$\underline{\phi}(x) \leq \phi(x) = \sum_{i \geq 1} a_i\, \mathbb{E}[S_1^{2i}](2x)^i 2^{-i} \leq \underline{\phi}(2x).$$

Since $\phi$ and $\underline{\phi}$ are increasing on $\mathbb{R}_+$, we have $\frac{1}{2}\underline{\phi}^{-1}(y) \leq \phi^{-1}(y) \leq \underline{\phi}^{-1}(y)$ for $y \geq 0$. Moreover, the inverse of this lower bound has closed form

$$\underline{\phi}^{-1}(y) = \inf_{i \geq i} (a_i \, \mathbb{E}[S_1^{2i}]/y)^{-1/i}.$$

We now bound

$$\begin{aligned} \inf_{i \geq 1} \sigma_{2i}(\mu)^2 \varepsilon^{1-1/i} &\leq \varepsilon \inf_{i \geq 1} (\varepsilon a_i \, \mathbb{E}[S_1^{2i}])^{-1/i} \\ &= \varepsilon \underline{\phi}^{-1}(1/\varepsilon) \\ &\leq 2\varepsilon \phi^{-1}(1/\varepsilon) \\ &= 2 \sup\big\{\varepsilon x^2 : x \geq 0, \mathbb{E}[\psi(|S_1|x)] \leq 1/\varepsilon\big\}. \end{aligned}$$

Finally, for any feasible $x$ for the final supremum, consider the random variable $Z \sim \nu$ defined by

$$Z = 0 \text{ w.p. } 1 - \varepsilon, \qquad Z = xS \text{ w.p. } \varepsilon.$$

By construction, we have

$$\tau_2(\nu, \varepsilon)^2 \geq \mathbb{E}[\|Z\|^2] = \varepsilon x^2,$$

and, for any $\theta \in \mathbb{S}^{d-1}$, we have

$$\mathbb{E}[\psi(|\theta^\top(Z - \mathbb{E}[Z])|)] = \varepsilon \, \mathbb{E}[\psi(|S_1|x)] \leq 1.$$

Combining, we have $\tau_2(\mathcal{G}, \varepsilon)^2 \geq \tau_2(\nu, \varepsilon)^2 \geq \varepsilon x^2 \geq \frac{1}{2} \inf_{i \geq 1} \sigma_{2i}(\mu)^2 \varepsilon^{1-1/i}$, as desired. □

From this result, we obtain the following two lemmas.

**Lemma 8.** *Fix $\varepsilon \in (0,1)$ and $\mu \in \mathcal{G}$ for $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$ satisfying Assumption 1. Then, for any $\nu \leq \frac{1}{\varepsilon}\mu$, we have $\varepsilon \sigma_2(\nu)^2 \leq 4\tau_2(\mathcal{G}, \varepsilon)^2$.*

*Proof.* Assume without loss of generality that $\mu$ has mean 0. Taking $Z \sim \mu$ and $Y \sim \nu$, we bound

$$\begin{aligned} \varepsilon \sigma_2(\nu)^2 &\leq 2\varepsilon \, \mathbb{E}[\|Y\|^2] && \text{(Lemma 4)} \\ &\leq 2\varepsilon \, \mathbb{E}[\|Z\|^2] + \varepsilon \tau(\|Z\|^2, 1 - \varepsilon) \\ &\leq 2\varepsilon \, \mathbb{E}[\|Z\|^2] + \inf_{i > 1} \mathbb{E}[\|Z\|^{2i}]^{1/i} \varepsilon^{1-1/i} && \text{(Lemma 5)} \\ &\leq 2 \inf_{i \geq 1} \mathbb{E}[\|Z\|^{2i}]^{1/i} \varepsilon^{1-1/i}. \end{aligned}$$

Applying Lemma 7 gives the lemma. □

**Lemma 9.** *If $\varepsilon \in (0,1)$ and $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$ satisfies Assumption 1, then $\tau(\mathcal{G}, \varepsilon) \leq 4 \frac{\sqrt{\varepsilon}}{(1-\varepsilon)} \tau_{2,1}(\mathcal{G}, \varepsilon)$.*

*Proof.* For each $\mu \in \mathcal{G}$, we bound

$$\begin{aligned} \frac{(1-\varepsilon)^2}{\varepsilon} \tau(\mu, \varepsilon)^2 &\leq 8 \inf_{q \geq 1} \sigma_{q,1}(\mu)^2 \varepsilon^{1-2/q} && \text{(Lemma 5)} \\ &\leq 8 \inf_{i \geq 1} \sigma_{2i,1}(\mu)^2 \varepsilon^{1-1/i} \\ &\leq 16 \tau_{2,1}(\mathcal{G}, \varepsilon). && \text{(Lemma 7)} \end{aligned}$$

Taking a supremum over $\mu \in \mathcal{G}$ gives the lemma. □

# B Generic DRO Regularizer Bounds

This section considers a generic DRO problem and a corresponding notion of regularization. As special cases, we highlight results for WDRO and TV DRO that underlie our proof of Theorem 1.

Fix a distribution class $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$ and a loss family $\mathcal{L} \subseteq \cap_{\mu \in \mathcal{G}} L^1(\mu)$. Let $\mathsf{C} : \mathcal{G} \to \mathcal{P}(\mathbb{R}^d)$ be a corruption channel taking $\mu \in \mathcal{G}$ to a set of potential $\tilde{\mu} \in \mathsf{C}(\mu)$. Then, for any such $\tilde{\mu}$, one can consider the generic DRO problem

$$\inf_{\ell \in \mathcal{L}} \sup_{\nu \in \mathcal{G} \cap \mathsf{C}^{-1}(\tilde{\mu})} \mathbb{E}_\nu[\ell]. \tag{7}$$

For a fixed $\nu \in \mathsf{C}(\mathcal{G})$ and $\ell \in \mathcal{L} \cap L^1(\nu)$, we define the *DRO regularizer*

$$\Omega(\ell; \nu, \mathcal{G}, \mathsf{C}) \coloneqq \sup_{\nu' \in \mathcal{G} \cap \mathsf{C}^{-1}(\nu)} \mathbb{E}_{\nu'}[\ell] - \mathbb{E}_\nu[\ell].$$

Assuming that $\ell \in L^1(\tilde{\mu})$, one can rewrite (7) as the regularized minimization problem

$$\inf_{\ell \in \mathcal{L}} \tilde{\mu}(\ell) + \Omega(\ell; \tilde{\mu}, \mathcal{G}, \mathsf{C}).$$

In any case, this quantity controls the excess risk of DRO. Writing $\mathsf{C}^{-1} \circ \mathsf{C}$ for the composite corruption channel taking $\mu \in \mathcal{G}$ to $\nu \in \mathcal{G}$ with $\mathsf{C}(\mu) \cap \mathsf{C}(\nu) \neq \emptyset$, we have the following.

**Lemma 10** (Risk bound for generic DRO)**.** *Fix $\mu \in \mathcal{G}$ and $\tilde{\mu} \in \mathsf{C}(\mu)$. If $\hat{\ell}$ minimizes* (7)*, then* $\mathbb{E}_\mu[\hat{\ell}] \leq \inf_{\ell \in \mathcal{L}} \mathbb{E}_\mu[\ell] + \Omega(\ell; \mu, \mathcal{G}, \mathsf{C}^{-1} \circ \mathsf{C})$.

*Proof.* We simply bound

$$\begin{aligned} \mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell] &\leq \sup_{\nu \in \mathcal{G} \cap \mathsf{C}^{-1}(\tilde{\mu})} \mathbb{E}_\nu[\hat{\ell}] - \mathbb{E}_\mu[\ell] \\ &\leq \sup_{\nu \in \mathcal{G} \cap \mathsf{C}^{-1}(\tilde{\mu})} \mathbb{E}_\nu[\ell] - \mathbb{E}_\mu[\ell] \\ &\leq \sup_{\nu \in \mathcal{G} \cap \mathsf{C}^{-1}(\mathsf{C}(\mu))} \mathbb{E}_\nu[\ell] - \mathbb{E}_\mu[\ell] = \Omega_{\mathsf{D}}(\ell, r; \mu, \mathcal{G}, \mathsf{C}^{-1} \circ \mathsf{C}). \end{aligned}$$

Infimizing over $\ell \in \mathcal{L}$ gives the lemma. □

When $\mathsf{C}(\mu) = \{\tilde{\mu} \in \mathcal{P}(\mathbb{R}^d) : \mathsf{D}(\tilde{\mu}, \mu) \leq r\}$ for a statistical distance $\mathsf{D} : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$ and radius $r \geq 0$, we write $\Omega_{\mathsf{D}}(\ell, r; \nu, \mathcal{G}) = \Omega(\ell; \nu, \mathcal{G}, \mathsf{C})$. If distributional assumptions play a minor role, we may opt to consider $\Omega_{\mathsf{D}}(\ell, r; \nu) \coloneqq \Omega_{\mathsf{D}}(\ell, r; \nu, \mathcal{P}(\mathbb{R}^d))$.

### B.1 WDRO Regularization

The $\mathsf{W}_p$ regularizer, corresponding to $\mathsf{D} = \mathsf{W}_p$, appears explicitly and implicitly throughout the WDRO literature. We now recall standard bounds on this quantity.

**Lemma 11** ($\Omega_{\mathsf{W}_1}$ bound, [11], Lemma 1)**.** *Fix $\nu \in \mathcal{P}_1(\mathbb{R}^d)$, Lipschitz $\ell : \mathbb{R}^d \to \mathbb{R}$, and $\rho \geq 0$. We then have $\Omega_{\mathsf{W}_1}(\ell, \rho; \nu) \leq \rho \|\ell\|_{\mathrm{Lip}}$, with equality if $\ell$ is convex and $\mathcal{Z} = \mathbb{R}^d$.*

**Lemma 12** ($\Omega_{\mathsf{W}_2}$ bound, [11], Lemma 2)**.** *Fix $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\alpha$-smooth $\ell : \mathbb{R}^d \to \mathbb{R}$, and $\rho \geq 0$. We then have $|\Omega_{\mathsf{W}_2}(\ell, \rho; \nu) - \rho \|\ell\|_{\dot{H}^{1,2}(\nu)}| \leq \frac{1}{2}\alpha\rho^2$.*

### B.2 TV DRO Regularization

We introduce new bounds (to the best of our knowledge) for the DRO regularizer with $\mathsf{D} = \mathsf{TV}$.

**Lemma 13** ($\Omega_{\mathsf{TV}}$ bound under Lipschitzness)**.** *Fix $\mu \in \mathcal{G} \subseteq \mathcal{P}_1(\mathbb{R}^d)$ and l.s.c. $\ell : \mathbb{R}^d \to \mathbb{R}$ with $\sup_{z \in \mathbb{R}^d} \frac{|\ell(z)|}{1 + \|z\|} < \infty$. If $\ell$ is Lipschitz, then*

$$\Omega_{\mathsf{TV}}(\ell, \varepsilon; \mu, \mathcal{G}) \leq \Omega_{\mathsf{W}_1}(\ell, 2\tau_1(\mathcal{G}, \varepsilon); \mu).$$

*Proof.* Fix $\nu \in \mathcal{G}$ with $\|\nu - \mu\|_{\mathsf{TV}} \leq \varepsilon$, and write $\kappa = \frac{1}{(\nu \wedge \mu)(\mathbb{R}^d)} \nu \wedge \mu$ for their midpoint distribution. Note that $(\nu \wedge \mu)(\mathbb{R}^d) \geq 1 - \varepsilon$ by the TV bound. We then have $\mathsf{W}_1(\nu, \mu) \leq \mathsf{W}_1(\nu, \kappa) + \mathsf{W}_1(\kappa, \mu) \leq 2\tau_1(\mathcal{G}, \varepsilon)$, implying the lemma. □

**Lemma 14** ($\Omega_{\mathsf{TV}}$ bound under smoothness)**.** *Fix $\mu \in \mathcal{G}$ for $\mathcal{G} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ satisfying Assumption 1, and let $\ell : \mathbb{R}^d \to \mathbb{R}$ be l.s.c. with $\sup_{z\in\mathbb{R}^d} \frac{|\ell(z)|}{1+\|z\|^2} < \infty$. If $\ell$ is $\alpha$-smooth, then*

$$\Omega_{\mathsf{TV}}(\ell, \varepsilon; \mu, \mathcal{G}) \le 2\|\nabla\ell(\mathbb{E}_\mu[Z])\|\tau(\mathcal{G}, \varepsilon) + 44\alpha(1-\varepsilon)^{-1}\tau_2(\mathcal{G}, \varepsilon)^2.$$

*Proof.* Fix any $\nu \in \mathcal{G}$ with $\|\nu - \mu\|_{\mathsf{TV}} \le \varepsilon$, and decompose $\nu = \mu + \varepsilon(\kappa_+ - \kappa_-)$, where $\kappa_\pm \in \mathcal{P}(\mathcal{Z})$ with $\varepsilon\kappa_- \le \mu$ and $\varepsilon\kappa_+ \le \nu$. Let $Z \sim \mu$, $Y \sim \kappa_-$, $X \sim \nu$, and $W \sim \kappa_+$. We bound

$$\begin{aligned}\mathbb{E}[\ell(X) - \ell(Z)] &= \varepsilon\,\mathbb{E}[\ell(W) - \ell(Y)]\\ &= \varepsilon\,\mathbb{E}\big[\ell(W) - \ell(\mathbb{E}[W])\big] + \varepsilon\big[\ell(\mathbb{E}[W]) - \ell(\mathbb{E}[Y])\big] + \varepsilon\,\mathbb{E}\big[\ell(\mathbb{E}[Y]) - \ell(Y)\big].\end{aligned}$$

To bound the first and last terms, we observe that for $V \sim \kappa = \kappa_\pm$, we have

$$\begin{aligned}\varepsilon\,\mathbb{E}\big[\ell(V) - \ell(\mathbb{E}[V])\big] &\le \alpha\varepsilon\,\mathbb{E}[\|V - \mathbb{E}[V]\|^2]\\ &\le \alpha\varepsilon\sigma_2(\kappa)^2\\ &\le 4\alpha\tau_2(\mathcal{G}, \varepsilon)^2,\end{aligned}$$

by $\alpha$-smoothness of $\tilde{\ell}$ and Lemma 8. For the second term, write $I = \mathrm{conv}(\{\mathbb{E}[W], \mathbb{E}[Y]\})$ for the line segment connecting $\mathbb{E}[W]$ and $\mathbb{E}[Y]$. By the definition of mean resilience, we bound

$$\begin{aligned}\|\,\mathbb{E}[W] - \mathbb{E}[X]\| &\le \tau(\mathcal{G}, 1-\varepsilon),\\ \|\,\mathbb{E}[Y] - \mathbb{E}[Z]\| &\le \tau(\mathcal{G}, 1-\varepsilon),\\ \|\,\mathbb{E}[Z] - \mathbb{E}[X]\| &\le 2\tau(\mathcal{G}, \varepsilon),\end{aligned}$$

where the last inequality follows by the same midpoint argument applied in the proof of Lemma 13. Writing $L = \|\nabla\ell(\mathbb{E}[Z])\|$, we have for each $x \in I$ that

$$\begin{aligned}\|\nabla\ell(x)\| &\le L + \alpha\|x - \mathbb{E}[Z]\|\\ &\le L + \alpha\max\{\|\,\mathbb{E}[W] - \mathbb{E}[Z]\|, \|\,\mathbb{E}[Y] - \mathbb{E}[Z]\|\}\\ &\le L + \alpha\max\{\tau(\mathcal{G}, 1-\varepsilon) + 2\tau(\mathcal{G}, \varepsilon), \tau(\mathcal{G}, 1-\varepsilon)\}\\ &\le L + \alpha\left(\frac{1-\varepsilon}{\varepsilon} + 2\right)\tau(\mathcal{G}, \varepsilon),\end{aligned}$$

again using smoothness of $\ell$. We then bound

$$\begin{aligned}\varepsilon\big[\ell(\mathbb{E}[W]) - \ell(\mathbb{E}[Y])\big] &\le \varepsilon\max_{x\in I}\|\nabla\ell(x)\|\|\,\mathbb{E}[X] - \mathbb{E}[Z]\|\\ &= \max_{x\in I}\|\nabla\ell(x)\|\|\,\mathbb{E}[X] - \mathbb{E}[Z]\|\\ &\le \left[L + \alpha\left(\frac{1-\varepsilon}{\varepsilon} + 2\right)\tau(\mathcal{G}, \varepsilon)\right]2\tau(\mathcal{G}, \varepsilon)\\ &= 2L\tau(\mathcal{G}, \varepsilon) + 2\alpha\left(\frac{1-\varepsilon}{\varepsilon} + 2\right)\tau(\mathcal{G}, \varepsilon)^2\\ &= 2L\tau(\mathcal{G}, \varepsilon) + 4\alpha\tau_2(\mathcal{G}, \varepsilon)^2 + 2\alpha\frac{1-\varepsilon}{\varepsilon}\tau(\mathcal{G}, \varepsilon)^2\\ &\le 2L\tau(\mathcal{G}, \varepsilon) + 4\alpha\tau_2(\mathcal{G}, \varepsilon)^2 + 32\alpha(1-\varepsilon)^{-1}\tau_{2,1}(\mathcal{G}, \varepsilon)^2 \qquad \text{(Lemma 9)}\\ &\le 2L\tau(\mathcal{G}, \varepsilon) + 36\alpha(1-\varepsilon)^{-1}\tau_{2,1}(\mathcal{G}, \varepsilon)^2.\end{aligned}$$

Combining the above, we obtain

$$\begin{aligned}\mathbb{E}[\ell(X)] - \mathbb{E}[\ell(Z)] &\le 8\alpha\tau_2(\mathcal{G}, \varepsilon) + 2L\tau(\mathcal{G}, \varepsilon) + 36\alpha(1-\varepsilon)^{-1}\tau_{2,1}(\mathcal{G}, \varepsilon)^2\\ &\le 2L\tau(\mathcal{G}, \varepsilon) + 44\alpha(1-\varepsilon)^{-1}\tau_2(\mathcal{G}, \varepsilon)^2,\end{aligned}$$

as desired. ☐

# C Proofs for Section 3

## C.1 Proof of Theorem 1

Our proof follows by analyzing the $\mathsf{W}_p^\varepsilon$ regularizer

$$\Omega_{\mathsf{W}_p^\varepsilon}(\ell, \rho; \mu, \mathcal{G}) = \sup_{\substack{\nu \in \mathcal{G} \\ \mathsf{W}_p^\varepsilon(\nu,\mu) \le \rho}} \mathbb{E}_\nu[\ell] - \mathbb{E}_\mu[\ell].$$

We bound this quantity from above by a $\mathsf{W}_p$ regularizer and a $\mathsf{TV}$ regularizer maximized over a Wasserstein ball centered at $\mu$.

**Lemma 15.** *Fix $\varepsilon \in [0, 1)$ and $\rho \ge 0$. For any $\mu \in \mathcal{G} \subseteq \mathcal{P}_p(\mathbb{R}^d)$ and $\ell : \mathbb{R}^d \to \mathbb{R}$ l.s.c. with $\sup_{z \in \mathbb{R}^d} \frac{|\ell(z)|}{1+\|z\|^p} < \infty$, we have*

$$\Omega_{\mathsf{W}_p^\varepsilon}(\ell, \rho; \mu, \mathcal{G}) \le \Omega_{\mathsf{W}_p}(\ell, \rho; \mu) + \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}^d) \\ \mathsf{W}_p(\nu,\mu) \le \rho}} \Omega_{\mathsf{TV}}(\ell, \varepsilon; \nu, \mathcal{G}).$$

*Proof.* Fix any $\kappa \in \mathcal{G}$ with $\mathsf{W}_p^\varepsilon(\kappa, \mu) \le \rho$. By the definition of $\mathsf{W}_p^\varepsilon$, there exists $\mu' \in \mathcal{P}(\mathbb{R}^d)$ with $\mathsf{W}_p(\mu', \mu) \le \rho$ and $\|\mu' - \kappa\|_{\mathsf{TV}} \le \varepsilon$. We thus bound

$$\begin{aligned} \mathbb{E}_\kappa[\ell] - \mathbb{E}_\mu[\ell] &= (\mathbb{E}_\kappa[\ell] - \mathbb{E}_{\mu'}[\ell]) + (\mathbb{E}_{\mu'}[\ell] - \mathbb{E}_\mu[\ell]) \\ &\le \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}^d) \\ \mathsf{W}_p(\nu,\mu) \le \rho}} \Omega_{\mathsf{TV}}(\ell, \varepsilon; \nu, \mathcal{G}) + \Omega_{\mathsf{W}_p}(\ell, \rho; \mu). \end{aligned}$$

Supremizing over $\kappa$ gives the lemma. □

Next, we show that, under the affine structure of $\ell_\star$, one can instead consider DRO in $\mathbb{R}^k$. In particular, writing $\mathcal{G}_k = \mathcal{G} \cap \mathcal{P}(\mathbb{R}^k)$ for some $U \in \mathbb{R}^{k \times d}$ with $UU^\top = I_k$ (the choice is not important due to rotational symmetry), we have the following.

**Lemma 16.** *Under Assumption 2, we may decompose $\ell_\star = \tilde{\ell} \circ Q$ for $Q \in \mathbb{R}^{k \times d}$ with $QQ^\top = I_k$ and l.s.c. $\tilde{\ell}$ with $\sup_{z \in \mathbb{R}^d} \frac{|\tilde{\ell}(z)|}{1+\|z\|^p} < \infty$. For any such decomposition, we have*

$$\sup_{\substack{\nu \in \mathcal{G} \\ \mathsf{W}_p^\varepsilon(\nu,\tilde{\mu}) \le \rho}} \mathbb{E}_\nu[\ell_\star] = \sup_{\substack{\nu \in \mathcal{G}_k \\ \mathsf{W}_p^\varepsilon(\nu,Q_\#\tilde{\mu}) \le \rho}} \mathbb{E}_\nu[\tilde{\ell}].$$

*Proof.* By Assumption 2, we can write $\ell_\star = \underline{\ell} \circ A$ for $A : \mathbb{R}^d \to \mathbb{R}^k$ affine and $\underline{\ell}$ l.s.c. with $\sup_{z \in \mathbb{R}^d} \frac{|\tilde{\ell}(z)|}{1+|z|^p} < \infty$. We further decompose $A(z) = RQz + z_0$, where $Q \in \mathbb{R}^{k \times d}$ with $QQ^\top = I_k$, $R \in \mathbb{R}^{k \times k}$, and $z_0 \in \mathbb{R}^k$. Note that the orthogonality condition ensures that $Q^\top$ isometrically embeds $\mathbb{R}^k$ into $\mathbb{R}^d$. We can then choose $\tilde{\ell}(w) = \underline{\ell}(Rw + z_0)$.

Next, given any $\nu \in \mathcal{G}$, we have $Q_\#\nu \in \mathcal{G}_k$ with $\mathsf{W}_p^\varepsilon(Q_\#\nu, Q_\#\tilde{\mu}) \le \mathsf{W}_p^\varepsilon(\nu, \tilde{\mu})$, and $\mathbb{E}_\nu[\ell] = \mathbb{E}_{Q_\#\nu}[\tilde{\ell}]$. Thus, the RHS supremum is always at least as large as the LHS. It remains to show the reverse.

Fix $\nu \in \mathcal{G}_k$ with $\mathsf{W}_p^\varepsilon(\nu, Q_\#\tilde{\mu})$. Take any $\nu' \in \mathcal{P}(\mathbb{R}^k)$ with $\mathsf{W}_p(\nu, \nu') \le \rho$ and $\|\nu' - Q_\#\tilde{\mu}\|_{\mathsf{TV}} \le \varepsilon$. Write $\kappa = Q_\#^\top \nu \in \mathcal{G}$ and $\kappa' = Q_\#^\top \nu'$. Since $Q^\top$ is an isometric embedding, we have $\kappa \in \mathcal{G}$, $\mathsf{W}_p(\kappa, \kappa') = \mathsf{W}_p(\nu, \nu') \le \rho$, and $\|\kappa' - \tilde{\mu}\|_{\mathsf{TV}} = \|\nu' - Q_\#\tilde{\mu}\|_{\mathsf{TV}} \le \varepsilon$. Finally, we have $\mathbb{E}_\nu[\ell] = \mathbb{E}_\kappa[\tilde{\ell}]$. Thus, the RHS supremum is no greater than the LHS, and we have the desired equality. □

We are now equipped to prove the theorem. Applying Lemma 16, we decompose $\ell_\star = \tilde{\ell} \circ Q$. We bound risk by

$$\begin{aligned} \mathbb{E}_\mu[\hat{\ell}] &\le \sup_{\substack{\nu \in \mathcal{G} \\ \mathsf{W}_p^\varepsilon(\nu,\tilde{\mu}) \le \rho}} \mathbb{E}_\nu[\hat{\ell}] \\ &\le \sup_{\substack{\nu \in \mathcal{G} \\ \mathsf{W}_p^\varepsilon(\nu,\tilde{\mu}) \le \rho}} \mathbb{E}_\nu[\ell_\star] \end{aligned}$$

$$\leq \sup_{\substack{\nu \in \mathcal{G}_k \\ \mathsf{W}_p^\varepsilon(\nu, Q_\# \tilde{\mu}) \leq \rho}} \mathbb{E}_\nu[\tilde{\ell}] \qquad \text{(Lemma 16)}$$

$$\leq \sup_{\substack{\nu \in \mathcal{G}_k \\ \mathsf{W}_p^{2\varepsilon}(\nu, Q_\# \mu) \leq 2\rho}} \mathbb{E}_\nu[\tilde{\ell}].$$

Writing $\mu_k = Q_\# \mu$, we can then bound excess risk by

$$\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star] \leq \sup_{\substack{\nu \in \mathcal{G}_k \\ \mathsf{W}_p^{2\varepsilon}(\nu, \mu_k) \leq 2\rho}} \mathbb{E}_\nu[\tilde{\ell}] - \mathbb{E}_{\mu_k}[\tilde{\ell}].$$

Noting that the RHS is just the $\mathsf{W}_p^\varepsilon$ regularizer of $\ell$ in $\mathbb{R}^k$, we apply Lemma 15 to obtain

$$\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star] \leq \Omega_{\mathsf{W}_p}(\tilde{\ell}, 2\rho; \mu_k) + \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}^k) \\ \mathsf{W}_p(\nu, \mu_k) \leq \rho}} \Omega_{\mathsf{TV}}(\tilde{\ell}, 2\varepsilon; \nu, \mathcal{G}_k),$$

If $p = 1$ and $\ell_\star$ is Lipschitz, we apply Lemma 13 and Lemma 11 to obtain

$$\begin{aligned}
\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star] &\leq \|\tilde{\ell}\|_{\mathrm{Lip}}(2\rho + 2\tau_1(\mathcal{G}_k, 2\varepsilon)) \\
&\leq \|\tilde{\ell}\|_{\mathrm{Lip}}(2\rho + 2\tau_1(\mathcal{G}_k, 2\varepsilon)) \\
&\leq \|\ell_\star\|_{\mathrm{Lip}}(2\rho + 2\tau_{1,k}(\mathcal{G}, 2\varepsilon))
\end{aligned}$$

If $p = 2$ and $\ell_\star$ is $\alpha$-smooth, we apply Lemma 14 and Lemma 12 to bound $\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell_\star]$ by

$$\begin{aligned}
&2\rho\|\tilde{\ell}\|_{\dot{H}^{1,2}(\mu_k)} + 4\alpha\rho^2 + \sup_{\substack{\nu \in \mathcal{P}(\mathbb{R}^k) \\ \mathsf{W}_p(\nu, \mu_k) \leq \rho}} 2\|\nabla\tilde{\ell}(\mathbb{E}_\nu[Z])\|\tau(\mathcal{G}_k, 2\varepsilon) + 44\alpha(1-2\varepsilon)^{-1}\tau_2(\mathcal{G}_k, 2\varepsilon)^2 \\
&\leq 2\rho\|\tilde{\ell}\|_{\dot{H}^{1,2}(\mu_k)} + 4\alpha\rho^2 + 2(\|\nabla\tilde{\ell}(\mathbb{E}_{\mu_k}[Z])\| + \alpha\rho)\tau(\mathcal{G}_k, 2\varepsilon) + 44\alpha(1-2\varepsilon)^{-1}\tau_2(\mathcal{G}_k, 2\varepsilon)^2 \\
&\leq 2\rho\|\tilde{\ell}\|_{\dot{H}^{1,2}(\mu_k)} + 2\|\nabla\tilde{\ell}(\mathbb{E}_{\mu_k}[Z])\|\tau(\mathcal{G}_k, 2\varepsilon) + 44\alpha(1-2\varepsilon)^{-1}\big(\rho^2 + \rho\tau(\mathcal{G}_k, 2\varepsilon) + \tau_2(\mathcal{G}_k, 2\varepsilon)^2\big) \\
&\leq 2\rho\|\tilde{\ell}\|_{\dot{H}^{1,2}(\mu_k)} + 2\|\nabla\tilde{\ell}(\mathbb{E}_{\mu_k}[Z])\|\tau(\mathcal{G}_k, 2\varepsilon) + 44\alpha(1-\varepsilon)^{-1}(\rho + \tau_2(\mathcal{G}_k, 2\varepsilon))^2 \\
&= 2\rho\|\ell_\star\|_{\dot{H}^{1,2}(\mu)} + 2\|\nabla\ell_\star(\mathbb{E}_\mu[Z])\|\tau(\mathcal{G}_k, 2\varepsilon) + 44\alpha(1-2\varepsilon)^{-1}(\rho + \tau_2(\mathcal{G}_k, 2\varepsilon))^2 \\
&= 2\rho\|\ell_\star\|_{\dot{H}^{1,2}(\mu)} + 2\|\nabla\ell_\star(\mathbb{E}_\mu[Z])\|\tau(\mathcal{G}, 2\varepsilon) + 44\alpha(1-2\varepsilon)^{-1}(\rho + \tau_{2,k}(\mathcal{G}, 2\varepsilon))^2,
\end{aligned}$$

as desired. □

### C.2 Risk bounds in Table 1

The upper bounds for OR-WDRO follow by combining Theorem 1 with Proposition 2.

To see that these are minimax optimal, we start by proving that no $\hat{\ell}$ chosen as a function of $\tilde{\mu}$ can obtain risk less than $L\rho$ in the worst-case, for any of the considered settings. We fix $\tilde{\mu} = \delta_{0_d}$ and consider two candidates $\mu_\pm = \delta_{\pm\rho e_1}$ for $\mu$. We let $\mathcal{L}$ consist of the two $L$-Lipschitz loss functions

$$\ell_+(z) \coloneqq L e_1^\top(\rho - z), \quad \ell_-(z) \coloneqq L e_1^\top z.$$

By construction, $\mu_+$ and $\mu_-$ both belong to $\mathcal{G} \in \{\mathcal{G}_{\mathrm{cov}}, \mathcal{G}_{\mathrm{subG}}\}$ and, for $\mu = \mu_\pm$, we have that $\|\ell_\pm\|_{\mathrm{Lip}} = \|\ell_\pm\|_{\dot{H}^{1,2}(\mu)} = L$. Moreover, we have

$$\mathbb{E}_{\mu_+}[\ell_+] = 0, \mathbb{E}_{\mu_+}[\ell_-] = L\rho, \mathbb{E}_{\mu_-}[\ell_+] = 0, \mathbb{E}_{\mu_-}[\ell_-] = -L\rho.$$

Thus, for any $\hat{\ell}$ selected as a function of $\tilde{\mu}$ (with $\mathsf{W}_p(\tilde{\mu}, \mu) \leq \rho$), there exists $\mu \in \{\mu_+, \mu_-\}$ such that

$$\mu(\hat{\ell}) - \inf_{\ell \in \mathcal{L}} \mu(\ell) \geq L\rho.$$

Next, we fix $p = 1$. For ease of presentation, suppose $d = 2m$ is even. Consider $\mathbb{R}^d$ as $\mathbb{R}^m \times \mathbb{R}^m$, and let $\mathcal{L}$ consist of the two $L$-Lipschitz loss functions

$$\ell_+(x, y) \coloneqq L\|x + y\|, \quad \ell_-(x, y) \coloneqq L\|x - y\|$$

Fixing corrupted measure $\tilde{\mu} = \delta_0$, we consider the following candidates for the clean measure $\mu$:

$$\mu_+ \coloneqq (1-\varepsilon)\delta_0 + \varepsilon(\mathrm{Id}, -\mathrm{Id})_{\#}\kappa$$
$$\mu_- \coloneqq (1-\varepsilon)\delta_0 + \varepsilon(\mathrm{Id}, +\mathrm{Id})_{\#}\kappa$$

where $\mathrm{Id} : x \mapsto x$ is the identity map and $\kappa \in \mathcal{P}(\mathbb{R}^m)$ will be selected later as a function of $\mathcal{G}$. By design, we have $\|\tilde{\mu} - \mu_+\|, \|\tilde{\mu} - \mu_-\|_{\mathsf{TV}} \le \varepsilon$ and

$$\mathbb{E}_{\mu_+}[\ell_+] = \mathbb{E}_{\mu_-}(\ell_-) = 0$$
$$\mathbb{E}_{\mu_+}[\ell_-] = \mathbb{E}_{\mu_-}[\ell_+] = 2L\varepsilon\,\mathbb{E}_\kappa[\|Z\|]$$

Thus, for any $\hat{\ell}$ selected as a function of $\tilde{\mu}$, there exists $\mu \in \{\mu_+, \mu_-\}$ such that

$$\mu(\hat{\ell}) - \inf_{\ell \in \mathcal{L}} \mu(\ell) = \mu(\hat{\ell}) \ge 2L\varepsilon\,\mathbb{E}_\kappa[\|Z\|].$$

When $\mathcal{G} = \mathcal{G}_{\mathrm{cov}}$, taking $\kappa = \mathcal{N}(0_m, \frac{1}{\varepsilon} I_m)$ ensures that $\mu_\pm \in \mathcal{G}_{\mathrm{cov}}$, and $L\varepsilon\,\mathbb{E}_\kappa[\|Z\|] \gtrsim L\sqrt{d\varepsilon}$, as desired. When $\mathcal{G} = \mathcal{G}_{\mathrm{subG}}$, taking $\kappa = \mathcal{N}(0_m, I_m)$ ensures that $\mu_\pm \in \mathcal{G}_{\mathrm{subG}}$, and $L\varepsilon\,\mathbb{E}_\kappa[\|Z\|] \gtrsim L\varepsilon\sqrt{d}$. The alternative choice of $\kappa = \delta_{\sqrt{\log(1/\varepsilon)} e_1}$ also ensures $\mu_\pm \in \mathcal{G}_{\mathrm{subG}}$ and $L\varepsilon\,\mathbb{E}_\kappa[\|Z\|] \gtrsim L\varepsilon\sqrt{\log(1/\varepsilon)}$. Combining gives a minimax lower lower bound of $L\varepsilon\sqrt{d + \log(1/\varepsilon)}$ for $\mathcal{G}_{\mathrm{subG}}$. These match the claimed lower bounds for $p = 1$ when $k = d$; for smaller $k$, we simply apply the same construction with $m = k/2$, ignoring the extra $d - k$ coordinates.

For $p = 2$, take $\mathcal{L}$ consisting of the $\alpha$-smooth loss functions $\ell_\pm(x, y) = \alpha\|x \mp y\|^2$. For $\mu_\pm$ as above with $\kappa = \mathcal{N}(0_m, \frac{1}{\varepsilon} I_m)$, we have $\|\ell_\pm\|_{\dot{H}^{1,2}(\mu_\pm)} = 0$. The same argument as above gives a lower bound of $\alpha d$ for $\mathcal{G}_{\mathrm{cov}}$. Repeating with the corresponding measures for $\mathcal{G}_{\mathrm{subG}}$ gives a lower bound of $\alpha d\varepsilon \log(1/\varepsilon)$. Going through this process with $\ell_\pm(x, y) = Le_1^\top(x - y)$ adds a mean resilience term of $L\sqrt{\varepsilon}$ for $\mathcal{G}_{\mathrm{cov}}$ and $L\varepsilon\sqrt{\log(1/\varepsilon)}$ for $\mathcal{G}_{\mathrm{subG}}$. Taking $\ell_+(z) = \alpha(\rho^2 - \|z\|^2)$ and $\ell_-(z) = \alpha\|z\|^2$ with $\mu_\pm = \delta_{\pm\rho e_1}$ adds a final $\alpha\rho^2$ to both lower bounds. We may substitute $d$ by $k$ as above.

In all of the table's cases, we find that the minimax lower bound matches the upper bound for OR-WDRO given by Theorem 1.

### C.3 Proof of Proposition 3

This is an immediate consequence of Markov's inequality and the empirical convergence bound $\mathbb{E}[\mathsf{W}_1(\hat{\mu}_n, \mu)] \lesssim \sqrt{d} n^{-1/d}$, which follows by [23, Theorem 3.1] since $\mu \in \mathcal{G}_{\mathrm{cov}}$. □

## D Proofs for Section 4

### D.1 Proof of Proposition 4

For $\mu \in \mathcal{G}_{\mathrm{cov}}$, we bound

$$\begin{aligned}
\mathbb{E}_\mu[\|Z - z_0\|^2] &\le 2\,\mathbb{E}_\mu[\|Z - \mathbb{E}_\mu[Z]\|^2] + 2\|\,\mathbb{E}_\mu[Z] - z_0\|^2 \\
&= 2\,\mathrm{tr}(\Sigma_\mu) + 2\|\,\mathbb{E}_\mu[Z] - z_0\|^2 \\
&\le 2d + 2\|\,\mathbb{E}_\mu[Z] - z_0\|^2 \\
&\le 2(\sqrt{d} + \|\,\mathbb{E}_\mu[Z] - z_0\|)^2.
\end{aligned}$$

Consequently, we have $\mu \in \mathcal{G}_2(\sigma, z_0)$ for $\sigma = \sqrt{2d} + \sqrt{2}\|\,\mathbb{E}_\mu[Z] - z_0\|$.

Next, we note that $\mathsf{W}_p^\varepsilon(\tilde{\mu}_n, \mu) \le \rho_0 + \mathsf{W}_p(\hat{\mu}_n, \mu)$. Thus, applying Theorem 1 with $\mathcal{G} = \mathcal{G}_2(\sigma, z_0)$ and using the resilience bound from Proposition 2 gives that for $\rho = \rho_0 + \mathsf{W}_p(\hat{\mu}_n, \mu) + 8\sigma\varepsilon^{1/p - 1/2}(1 - \varepsilon)^{-1/p}$, the desired excess risk bounds hold so long as $\|\,\mathbb{E}_\mu[Z] - z_0\| = \rho_0 + O(\sqrt{d})$. Indeed, under these conditions with $p = 1$, we have for each $\ell \in \mathcal{L}$ that

$$\begin{aligned}
\mathbb{E}_\mu[\hat{\ell}] - \mathbb{E}_\mu[\ell] &\le c\|\ell\|_{\mathrm{Lip}}\big(\rho + 2\tau_1(\mathcal{G}_2(\sigma, z_0)\big) \\
&\lesssim \|\ell\|_{\mathrm{Lip}}\big(\rho + \sigma\sqrt{\varepsilon}\big) && \text{(Proposition 2)} \\
&\lesssim \|\ell\|_{\mathrm{Lip}}\big(\rho_0 + \mathsf{W}_1(\hat{\mu}_n, \mu) + \sigma\sqrt{\varepsilon}\big) \\
&\lesssim \|\ell\|_{\mathrm{Lip}}\big(\rho_0 + \mathsf{W}_1(\hat{\mu}_n, \mu) + \sqrt{d\varepsilon}\big),
\end{aligned}$$

as desired. □

## D.2 Proof of Proposition 5

Since iterative filtering works by identifying a subset of samples with bounded covariance and $\mathsf{W}_1$ perturbations can arbitrarily increase second moments, it is not immediately clear how to apply this method. Fortunately, trimming out a small fraction of samples ensures that second moments do not increase too much.

**Lemma 17.** *For any $\tau \in (0,1]$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we have $\mathsf{W}_2^\tau(\mu,\nu) \leq \mathsf{W}_1(\mu,\nu)\sqrt{2/\tau}$.*

*Proof.* Let $(X,Y)$ be a coupling of $\mu$ and $\nu$ such that $\mathbb{E}[\|X-Y\|] = \mathsf{W}_1(\mu,\nu)$. Write $\Delta = \|X-Y\|$, let $F$ denote its CDF, and note that $F^{-1}(1-\tau) \leq \mathsf{W}_1(\mu,\nu)/\varepsilon$ by Markov's inequality. Thus,

$$\begin{aligned}
\mathsf{W}_2^\tau(\mu,\nu)^2 &\leq \mathbb{E}[\Delta^2 \mid \Delta \leq F^{-1}(1-\tau)]\\
&\leq \mathbb{E}[\Delta^2 \mid \Delta \leq \mathsf{W}_1(\mu,\nu)/\tau]\\
&= \int_0^{\mathsf{W}_1(\mu,\nu)^2\tau^{-2}} \Pr\Big[\Delta > \sqrt{t} \mid \Delta \leq \mathsf{W}_1(\mu,\nu)/\tau\Big]\,\mathrm{d}t\\
&\leq \int_0^{\mathsf{W}_1(\mu,\nu)^2\tau^{-2}} \Big(\mathbb{E}[\Delta \mid \Delta \leq \mathsf{W}_1(\mu,\nu)/\tau]\, t^{-1/2} \wedge 1\Big)\,\mathrm{d}t\\
&\leq \int_0^{\mathsf{W}_1(\mu,\nu)^2\tau^{-2}} \Big(\mathsf{W}_1(\mu,\nu)\, t^{-1/2} \wedge 1\Big)\,\mathrm{d}t\\
&= \mathsf{W}_1(\mu,\nu)^2 + \mathsf{W}_1(\mu,\nu)\int_{\mathsf{W}_1(\mu,\nu)^2}^{\mathsf{W}_1(\mu,\nu)^2\tau^{-2}} t^{-1/2}\,\mathrm{d}t\\
&= \mathsf{W}_1(\mu,\nu)^2 + \mathsf{W}_1(\mu,\nu)\cdot 2\sqrt{t}\,\Big|_{\mathsf{W}_1(\mu,\nu)^2}^{\mathsf{W}_1(\mu,\nu)^2\tau^{-2}}\\
&= \mathsf{W}_1(\mu,\nu)^2 + 2\mathsf{W}_1(\mu,\nu)^2/\tau - 2\mathsf{W}_1(\mu,\nu)^2\\
&\leq 2\mathsf{W}_1(\mu,\nu)^2/\tau.
\end{aligned}$$

Taking a square root gives the claim. □

Write $\mu_n'$ for any uniform discrete measure over $n$ points such that $\mathsf{W}_1(\mu_n', \hat{\mu}_n) \leq \rho_0$ and $\|\mu_n' - \tilde{\mu}_n\|_{\mathsf{TV}} \leq \varepsilon$. It is well known that the empirical measure $\hat{\mu}_n$ will inherit the bounded covariance of $\mu$ for $n$ sufficiently large, so long as a small fraction of samples are trimmed out. In particular, by Lemma 4.2 of [18] and our sample complexity requirement, there exists a uniform discrete measure $\alpha_m$ over a subset of $m = (1-\varepsilon/120)n$ points, such that $\|\mathbb{E}_{\alpha_m}[Z] - \mathbb{E}_\mu[Z]\| \lesssim \sqrt{\varepsilon}$ and $\Sigma_{\alpha_m} \preceq O(1)I_d$ with probability at least 0.99. Moreover, by Lemma 17 with $\tau = \varepsilon/120$, there exists $\beta \in \mathcal{P}(\mathbb{R}^d)$ with $\|\beta - \mu_n'\|_{\mathsf{TV}} \leq \varepsilon/120$ and $\mathsf{W}_2^{\varepsilon/120}(\beta, \hat{\mu}_n) \leq \sqrt{240/\varepsilon}\rho_0$. Combining, we have that $\mathsf{W}_2^{\varepsilon/120+\varepsilon/120+\varepsilon}(\alpha_m, \tilde{\mu}_n) = \mathsf{W}_2^{61\varepsilon/60}(\alpha_m, \tilde{\mu}_n) \leq \sqrt{240/\varepsilon}\rho_0$.

Thus, there exists a uniform discrete measure $\gamma_m$ with support size $m$ such that $\|\gamma_m - \tilde{\mu}_n\|_{\mathsf{TV}} \leq 61/60\varepsilon$, $\mathsf{W}_2(\gamma_m, \alpha_m) \leq \sqrt{240/\varepsilon}\rho_0$, and $\mathsf{W}_1(\gamma_m, \alpha_m) \leq \rho_0$. The $\mathsf{W}_2$ bound implies that $\Sigma_{\gamma_m} \preceq O(1+\rho_0^2\varepsilon^{-1})I_d$. Thus, by the proof of Theorem 4.1 in [18] and our sample complexity requirement, the iterative filtering algorithm (Algorithm 1 therein) applied with an outlier fraction of $61/60\varepsilon \leq 1/10$ returns a reweighting of $\tilde{\mu}_m$ whose mean $z_0 \in \mathbb{R}^d$ is within $O(\sqrt{\varepsilon} + \rho_0)$ of that of $\gamma_m$. By a triangle inequality, the same error bound holds with respect to the mean of $\mu$. □

## D.3 Proof of Proposition 6

We have

$$\sup_{\substack{\nu \in \mathcal{G}_2(\sigma, z_0)\\ \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \leq \rho}} \mathbb{E}_\nu[\ell] = \sup_{\substack{\mu', \nu \in \mathcal{P}(\mathbb{R}^d)\\ \pi \in \Pi(\mu', \nu)}} \left\{ \mathbb{E}_\nu[\ell] : \begin{array}{c} \mathbb{E}_\nu[\|Z - z_0\|^2] \leq \sigma^2, \\ \mathbb{E}_\pi[\|Z' - Z\|^p] \leq \rho^p, \\ \mu' \leq \frac{1}{1-\varepsilon}\tilde{\mu}_n \end{array} \right\}$$

$$= \sup_{\substack{m \in \mathbb{R}^n \\ \nu_1, \dots, \nu_n \in \mathcal{P}(\mathbb{R}^d)}} \left\{ \sum_{i \in [n]} m_i \, \mathbb{E}_{\nu_i}[\ell] : \begin{array}{l} \sum_{i \in [n]} m_i \, \mathbb{E}_{\nu_i}[\|Z_i - z_0\|^2] \le \sigma^2, \\ \sum_{i \in [n]} m_i \, \mathbb{E}_{\nu_i}[\|\tilde{z}_i - Z_i\|^p] \le \rho^p, \\ 0 \le m_i \le \frac{1}{n(1-\varepsilon)}, \ \forall i \in [n] \\ \sum_{i \in [n]} m_i = 1 \end{array} \right\},$$

where the first equality follows from the definitions of $\mathcal{G}_2(\sigma, z_0)$ and $\mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu)$. The second equality holds because $\tilde{\mu}_n = \frac{1}{n} \sum_{i \in [n]} \delta_{\tilde{z}_i}$, which implies that the distributions $\mu', \nu$ and $\pi$ take the form $\mu' = \sum_{i \in [n]} m_i \delta_{\tilde{z}_i}$, $\nu = \sum_{i \in [n]} m_i \nu_i$, and $\pi = \sum_{i \in [n]} m_i \delta_{\tilde{z}_i} \otimes \nu_i$, respectively. Note that the distribution $\nu_i$ models the probability distribution of the random variable $Z$ condition on the event that $Z' = \tilde{z}_i$. Using the definition of the expectation operator and introducing the positive measure $\nu_i' = m_i \nu_i$ for every $i \in [n]$, we arrive at

$$\sup_{\substack{\nu \in \mathcal{G}_2(\sigma, z_0) \\ \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \le \rho}} \mathbb{E}_\nu[\ell] = \sup_{\substack{m \in \mathbb{R}^n \\ \nu_1', \dots, \nu_n' \ge 0}} \left\{ \sum_{i \in [n]} \mathbb{E}_{\nu_i'}[\ell] : \begin{array}{l} \sum_{i \in [n]} \int_{\mathbb{R}^d} \|z_i - z_0\|^2 d\nu_i'(z_i) \le \sigma^2, \\ \sum_{i \in [n]} \int_{\mathcal{Z}} \|z_i - \tilde{z}_i\|^p d\nu_i'(z_i) \le \rho^p, \\ 0 \le m_i \le \frac{1}{n(1-\varepsilon)}, \ \forall i \in [n], \\ \sum_{i \in [n]} m_i = 1 \\ \int_{\mathcal{Z}} d\nu_i'(z_i) = m_i, \ \forall i \in [n] \end{array} \right\}$$

$$= \inf_{\substack{\lambda_1, \lambda_2 \in \mathbb{R}_+ \\ r, s \in \mathbb{R}^n, \alpha \in \mathbb{R}}} \left\{ \lambda_1 \sigma^q + \lambda_2 \rho^p + \frac{\sum_{i \in [n]} s_i}{n(1-\varepsilon)} + \alpha : \begin{array}{r} s_i \ge \max\{0, r_i - \alpha\}, \ \forall i \in [n], \\ r_i \ge \ell(\xi) - \lambda_1 \|\xi - z_0\|^2 - \lambda_2 \|\xi - \tilde{z}_i\|^p, \\ \forall \xi \in \mathbb{R}^d, \forall i \in [n] \end{array} \right\},$$

where the second equality follows from strong duality, which holds because the Slater condition outlined in [37, Proposition 3.4] is satisfied thanks to Assumption 3. The proof concludes by removing the decision variables $r$ and $s$ and using the definition of $\tilde{\mu}_n$. □

### D.4 Proof of Theorem 2

The proof requires the following preparatory lemma. We say that the function $f$ is proper if $f(x) > -\infty$ and $\mathrm{dom}(f) \neq \emptyset$.

**Lemma 18.** *The followings hold.*

*(i) Let $f(x) = \lambda g(x - x_0)$, where $\lambda \ge 0$ and $g : \mathbb{R}^d \to \mathbb{R}$ is l.s.c. and convex. Then, $f^*(y) = x_0^\top y + \lambda g^*(y/\lambda)$.*

*(ii) Let $f(x) = \|x\|^p$ for some $p \ge 1$. Then, $f^*(y) = h(y)$, where the function $h$ is defined as in* (5)*.*

*(iii) Let $f(x) = x^\top \Sigma x$ for some $\Sigma \succ 0$. Then, $f^*(y) = \frac{1}{4} y^\top \Sigma^{-1} y$.*

*Proof.* The claims follows from [17, §E, Proposition 1.3.1 ], [47, Lemma B.8 (ii)] and [17, §E, Example 1.1.3], respectively. □

Now, by Proposition 6 and exploiting the definition of $\tilde{\mu}_n$, we have

$$\sup_{\substack{\nu \in \mathcal{G}_2(\sigma, z_0) \\ \mathsf{W}_p^\varepsilon(\tilde{\mu}_n \| \nu) \le \rho}} \mathbb{E}_\nu[\ell] \tag{8}$$

$$= \left\{ \begin{array}{ll} \inf & \lambda_1 \sigma^2 + \lambda_2 \rho^p + \alpha + \dfrac{1}{n(1-\varepsilon)} \displaystyle\sum_{i \in [n]} s_i \\ \text{s.t.} & \lambda_1, \lambda_2 \in \mathbb{R}_+, \ s \in \mathbb{R}_+^n \\ & s_i \ge \displaystyle\sup_{\xi \in \mathcal{Z}} \ \ell(\xi) - \lambda_1 \|\xi - z_0\|^2 - \lambda_2 \|\xi - \tilde{z}_i\|^p - \alpha \quad \forall i \in [n] \end{array} \right.$$

$$
= \begin{cases} \inf & \lambda_1\sigma^2 + \lambda_2\rho^p + \alpha + \frac{1}{n(1-\varepsilon)}\sum_{i\in[n]} s_i \\ \text{s.t.} & \lambda_1,\lambda_2 \in \mathbb{R}_+,\ s\in\mathbb{R}^n_+ \\ & s_i \geq \sup_{\xi\in\mathcal{Z}} \ell_j(\xi) - \lambda_1\|\xi - z_0\|^2 - \lambda_2\|\xi - \tilde{z}_i\|^p - \alpha \quad \forall i\in[n], \forall j \in [J] \end{cases} \tag{9}
$$

where the second equality follows from Assumption 4. For any fixed $i \in [n]$ and $j \in [J]$, we have

$$
\begin{aligned}
&\sup_{\xi\in\mathcal{Z}} \ell_j(\xi) - \lambda_1\|\xi - z_0\|^2 - \lambda_2\|\xi - \tilde{z}_i\|^p - \alpha \\
=& \begin{cases} \inf & (-\ell_j)^*(\zeta^\ell_{ij}) + z_0^\top \zeta^{\mathcal{G}}_{ij} + \tau_{ij} + \tilde{z}_i^\top \zeta^{\mathsf{W}}_{ij} + P_h(\zeta^{\mathsf{W}}_{ij}, \lambda_2) - \alpha \\ \text{s.t.} & \tau_{ij} \in \mathbb{R}^n_+,\ \zeta^\ell_{ij}, \zeta^{\mathcal{G}}_{ij}, \zeta^{\mathsf{W}}_{ij},\ \zeta^\ell_{ij} + \zeta^{\mathcal{G}}_{ij} + \zeta^{\mathsf{W}}_{ij} + = 0,\ \|\zeta^{\mathcal{G}}_{ij}\|^2 \leq \lambda_1\tau_{ij} \end{cases}
\end{aligned}
$$

where the equality is a result of strong duality due to [47, Theorem 2] and Lemma 18. The claim follows by substituting all resulting dual minimization problems into (9) and eliminating the corresponding minimization operators. □

# E Additional Experiments

In addition to the experiments in the main body, we also present applications to classification and multivariate regression. Code for all experiments is provided at `https://anonymous.4open.science/r/outlier-robust-WDRO-14EB/`. We first consider linear classification with the hinge loss, i.e. $\mathcal{L} = \{\ell_\theta(x,y) = \max\{0, 1 - y(\theta^\top x)\} : \theta \in \mathbb{R}^{d-1}\}$. This time (to ensure that the resulting optimization problem is convex), our approach supports Euclidean Wasserstein perturbations in the feature space, but no Wasserstein perturbations in the label space (this corresponds to using $\mathcal{Z} = \mathbb{R}^{d-1} \times \mathbb{R}$ equipped with the (extended) norm $\|(x,y)\| = \|x\|_2 + \infty \cdot \mathbb{1}\{y \neq 0\}$. We consider clean data $(X, \theta_0^\top X) \sim \mu$ as defined in Section 5. The corrupted data $(\tilde{X}, \tilde{Y}) \sim \tilde{\mu}$ satisfies $(\tilde{X}, \tilde{Y}) = (X + \rho e_1, Y)$ with probability $1 - \varepsilon$ and $(\tilde{X}, \tilde{Y}) = (20X, -20\theta_0^\top X)$ with probability $\varepsilon$, so that $\mathsf{W}_p^\varepsilon(\tilde{\mu}\|\mu) \leq \rho$. In Figure 2 (left), we fix $d = 10$ and compare the excess risk $\mathbb{E}_\mu[\ell_{\hat{\theta}}] - \mathbb{E}_\mu[\ell_{\theta_0}]$ of standard WDRO and outlier-robust WDRO with $\mathcal{A} = \mathcal{G}_2$, as described by Proposition 4 and implemented via Theorem 2. The results are averaged over $T = 20$ runs for sample size $n \in \{10, 20, 50, 75, 100\}$. We note that this contamination example cannot drive the excess risk of standard WDRO to infinity, so the separation between standard and outlier-robust WDRO is less striking than regression, though still present.

Finally, we present results for multivariate regression. This time, we consider $\mathcal{Z} = \mathbb{R}^{d\times k}$ equipped with the $\ell_2$ norm, and use the loss family $\mathcal{L} = \{\ell_M(x,y) = \|Mx - y\|_1 : M \in \mathbb{R}^{k\times d}\}$. We consider clean data $(X, M_0^\top X) \sim \mu$, where $M_0 \in \mathbb{R}^{k\times d}$ and $X$ have standard normal entries. The corrupted data $(\tilde{X}, \tilde{Y}) \sim \tilde{\mu}$ satisfies $(\tilde{X}, \tilde{Y}) = (X + \rho e_1, Y)$ with probability $1 - \varepsilon$ and $(\tilde{X}, \tilde{Y}) = (20X, -20M_0X)$ with probability $\varepsilon$, so that $\mathsf{W}_p^\varepsilon(\tilde{\mu}\|\mu) \leq \rho$. In Figure 2 (right), we fix $d = 10$ and $k = 3$, and compare the excess risk $\mathbb{E}_\mu[\ell_{\hat{\theta}}] - \mathbb{E}_\mu[\ell_{\theta_0}]$ of standard WDRO and outlier-robust WDRO with $\mathcal{A} = \mathcal{G}_2$, as described by Proposition 4 and implemented via Theorem 2. The results are averaged over $T = 10$ runs for sample size $n \in \{10, 20, 50, 75, 100\}$. We are restricted to low $k$ since the $\ell_1$ norm in the losses is expressed as the maximum of $2^k$ concave functions (specifically, we use that $\ell_M(x,y) = \max_{\alpha\in\{-1,1\}^k} \alpha^\top(Mx - y)$).

In both cases, confidence bands are plotted representing the top and bottom 10% quantiles among 100 bootstrapped means from the $T$ runs. The additional experiments were performed on an M1 Macbook Air with 16GB RAM in roughly 30 minutes each.

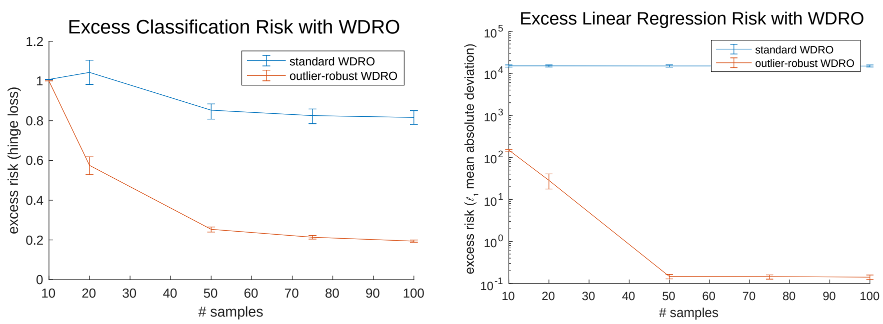


**Figure 2:** Excess risk of standard WDRO and outlier-robust WDRO for classification and multivariate linear regression under $\mathrm{W}_p$ and TV corruptions, with varied sample size.