Proceedings of Machine Learning Research vol 196:1–29, 2024 37th Annual Conference on Learning Theory

Robust Distribution Learning with Local and Global Adversarial Corruptions

author names withheld

Editor: Under Review for COLT 2024

Abstract

We consider learning in an adversarial environment, where an ε-fraction of samples from a distribution P are arbitrarily modified (*global* corruptions) and the remaining perturbations have average magnitude bounded by ρ (*local* corruptions). Given access to these n corrupted samples, we seek a computationally efficient estimator $\hat{P}_n$ that minimizes the Wasserstein distance $\mathsf{W}_1(\hat{P}_n, P)$. In fact, we attack the fine-grained task of minimizing $\mathsf{W}_1(\Pi_\sharp \hat{P}_n, \Pi_\sharp P)$ for all orthogonal projections $\Pi \in \mathbb{R}^{d\times d}$, with performance scaling with $\mathrm{rank}(\Pi) = k$. This allows us to account simultaneously for mean estimation ($k = 1$), distribution estimation ($k = d$), as well as the settings interpolating between these two extremes. We characterize the population-limit optimal risk for this task and then proceed to develop an efficient finite-sample algorithm with error bounded by $\sqrt{\varepsilon k} + \rho + d^{O(1)}\tilde{O}(n^{-1/k})$ when P is sub-Gaussian. For data distributions with bounded covariance, our finite-sample bounds match the minimax population-level optimum for large sample sizes. Our efficient procedure relies on a novel trace norm approximation of an ideal yet intractable 2-Wasserstein projection estimator. We apply this algorithm to robust stochastic optimization, and, in the process, uncover a new method for overcoming the curse of dimensionality in Wasserstein distributionally robust optimization.

Keywords: robust statistics, optimal transport, distributionally robust optimization

1. Introduction

In robust statistics and adversarial machine learning, estimation and decision-making are treated as a two-player game between the learner and a budget-constrained adversary. Through this lens, researchers have developed learning algorithms with strong guarantees despite adversarial corruptions. For example, Huber's ε-contamination model in classical robust statistics (Huber, 1964) and the total variation (TV) ε-contamination model (Donoho and Liu, 1988) give the adversary an ε fraction of data to arbitrarily and globally corrupt. Popularized recently in the setting of adversarial training (Sinha et al., 2018), Wasserstein corruption models permit all of the data to be locally perturbed, bounding the average perturbation size by some radius $\rho \geq 0$. Recall that the p-Wasserstein distance is defined between distributions P, Q by

$$\mathsf{W}_p(P, Q) := \inf_{\pi\in\Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi} \left[\|X - Y\|_2^p\right]^{\frac{1}{p}},$$

where $\Pi(P, Q)$ is the set of their couplings. This metric naturally lifts the geometry of $\mathbb{R}^d$ to the space of distributions $\mathcal{P}(\mathbb{R}^d)$ with finite p-th absolute moments.

Ideally, a corruption model should be flexible enough to capture multiple types of data contamination. Towards this goal, we investigate learning under combined TV and Wasserstein adversarial corruptions, recently introduced in the setting of distributionally robust optimization (DRO) (Nietert et al., 2023b). Formally, we consider learning where clean samples $X_1, \ldots, X_n \sim P$ are arbitrarily perturbed to obtain $\{\tilde{X}_i\}_{i=1}^n$ such that $\sum_{i\in S} \|\tilde{X}_i - X_i\|_2 \leq \rho$, where $S \subseteq [n]$ with $|S| \geq (1 - \varepsilon)n$. Denoting the clean and corrupted empirical measures by P_n and $\tilde{P}_n$, respectively, this corruption model is characterized by an outlier-robust variant of the Wasserstein distance defined in (2) ahead, whereby $\mathsf{W}_1^\varepsilon(P_n, \tilde{P}_n) \leq \rho$. We ask:

How can we learn effectively and efficiently with both local and global adversarial corruptions?

Under this combined model, we seek an estimate $\hat{P}_n$ that can approximate P in a variety of downstream applications. In particular, we explore the distribution learning task of recovering P under W_1 itself. Since the sample complexity and risk bounds associated with standard W_1 suffer a curse of dimensionality, we focus on the fine grained-goal of estimating k-dimensional projections of P, with performance scaling with k. Quantitatively, we seek $\hat{P}_n$ such that the *k-dimensional max-sliced Wasserstein distance*

$$\mathsf{W}_{1,k}(\hat{P}_n, P) := \sup_{\substack{U \in \mathbb{R}^{k\times d} \\ UU^\top = I_k}} \mathsf{W}_1(U_\# \hat{P}, U_\# P) = \sup_{\substack{U \in \mathbb{R}^{k\times d}, f \in \mathrm{Lip}_1(\mathbb{R}^k) \\ UU^\top = I_k}} \mathbb{E}_{\hat{P}_n}[f(UX)] - \mathbb{E}_P[f(UX)]$$

is appropriately small for all $k \in [d]$. Since our approach cleanly addresses all slicing dimensions simultaneously, we focus on providing bounds which are uniform in k. By doing so, we account not only for the said distribution estimation task ($k = d$), but also mean estimation $k = 1$

While this task is relatively straightforward under TV corruption alone (we show in Section 2.3 that a standard iterative filtering algorithm (Diakonikolas et al., 2016) suffices), and immediate under Wasserstein corruption alone where the corrupted distribution $\tilde{P}_n$ is itself minimax optimal, the combined model requires a new algorithmic approach and analysis to obtain suitable risk bounds. Eventually, we revisit the Wasserstein DRO setting that introduced in (Nietert et al., 2023b). Here, the steps we take to employ our estimate lead to a new perspective on generalization and radius selection when employing Wasserstein ambiguity sets for distributionally robust stochastic optimization.

1.1. Our Results

Our results are most cleanly stated when P is sub-Gaussian (although they hold for a breadth of additional settings). In this case, supposing that that $\mathsf{W}_1^\varepsilon(\tilde{P}_n, P_n) \le \rho$, our algorithm W2PROJECT (Algorithm 1) efficiently computes an estimate $\hat{P}_n$ such that

$$\mathsf{W}_{1,k}(\hat{P}_n, P) \lesssim \sqrt{k\varepsilon} + \rho + d^{O(1)} \tilde{O}\big(n^{-\frac{1}{k \vee 2}}\big)$$

for all $k \in [d]$. In Section 2.2, we show that this algorithm serves as a tractable proxy for the minimum distance estimate

$$\hat{P}_{\mathrm{MDE}} = \mathrm{argmin}_{Q \le \frac{1}{1-\varepsilon}\tilde{P}_n} \mathsf{W}_2(Q, \mathcal{G}_{\mathrm{cov}}),$$

where Q ranges over all distributions obtained by deleting an ε-fraction of mass from $\tilde{P}_n$ and renormalizing, and $\mathsf{W}_2(Q, \mathcal{G}_{\mathrm{cov}}) = \inf_{R : \Sigma_R \preceq I_d} \mathsf{W}_2(Q, R)$. In the infinite-sample population-limit (Section 2.1), we prove sharper risk bounds using a family of related minimum distance estimators. In particular, incorporating the sub-Gaussian assumption more directly leads to a tight risk bound of $\sqrt{k}\tilde{O}(\varepsilon) + \rho$. While the $\sqrt{k\varepsilon}$ term above seems sub-optimal in light of this fact, we emphasize that beating a $\sqrt{\varepsilon}$ dependence in algorithmic robust statistics typically requires strong assumptions like full Gaussianity or identity covariance. For this reason, we view our performance guarantee for W2PROJECT as a strong starting point for algorithm design in the W_1^ε corruption model.

Given such an algorithm, we then explore applications to robust stochastic optimization. Suppose that we have an estimate $\hat{P}$ of known quality $\mathsf{W}_{1,k}(\hat{P}, P) \le \tau$, perhaps from the procedure above. Given a family of Lipschitz loss functions $\mathcal{L}$ which operate on k-dimensional linear features (e.g. k-variate linear regression) we prove that the Wasserstein DRO estimate

$$\hat{\ell} = \mathrm{argmin}_{\ell \in \mathcal{L}} \sup_{Q : \mathsf{W}_1(Q, \hat{P}_n) \lesssim \tau} \mathbb{E}_Q[\ell] \tag{1}$$

achieves risk bounds typically associated with the sliced-Wasserstein DRO problem

$$\min_{\ell \in \mathcal{L}} \sup_{Q:\mathsf{W}_{1,k}(Q,\hat{P}_n)\lesssim\tau} \mathbb{E}_Q[\ell],$$

even though P need not belong to the Wasserstein ambiguity set in (1). In particular, we prove that $\hat{\ell}$ satisfies the excess risk bound $\mathbb{E}_P[\hat{\ell}] - \mathbb{E}_P[\ell_\star] \lesssim \|\ell_\star\|_{\mathrm{Lip}}\tau$, where $\ell_\star = \mathrm{argmin}_{\ell\in\mathcal{L}} \mathbb{E}_P[\ell]$. Plugging in the algorithmic results above improves upon existing results for outlier-robust WDRO, exhibiting tight dependence on k and with sampling error scaling as $n^{-1/k}$ rather than $n^{-1/d}$. This risk bound would be immediate if the W_1 ball above was replaced with a $\mathsf{W}_{1,k}$ ball. The fact that this is not necessary is essential for computational tractability, and provides a new framework for avoiding the curse of dimensionality (CoD) in Wasserstein DRO. We note that this result is new even when $\varepsilon = 0$ and $\rho > 0$ is taken to model only stochastic sampling error. Previous results on avoiding the CoD required $k = 1$ or involved significantly more complicated analysis. In particular, for the rank-one linear structure with $k = 1$, including univariate linear regression/classification, the bounds of $O(\sqrt{n})$ have been established in (Shafieezadeh-Abadeh et al., 2019; Chen and Paschalidis, 2018; Olea et al., 2022; Wu et al., 2022). On the other hand, Gao (2022) relaxes the rank-one structural assumption and achieves $O(\sqrt{n})$ bounds as long as the data generating distribution satisfies certain transport inequalities. Nonetheless, the required assumptions are not easily verifiable.

1.2. Related Work

Robust statistics. Learning from data under TV ε-corruptions, a staple of classical robust statistics, dates back to Huber (1964). Various robust and sample-efficient estimators, particularly for mean and scale parameters, have been developed in the robust statistics community; see Ronchetti and Huber (2009) for a comprehensive survey. Recently, Zhu et al. (2022a), significantly expanding the celebrated results of Donoho and Liu (1988), developed a unified statistical framework for robust statistics based on minimum distance estimation and a generalized resilience quantity, providing sharp population-limit and strong finite-sample guarantees for tasks including mean and covariance estimation.

Over the past decade, the focus in the computer science community has shifted to the high-dimensional setting, where they have developed computationally efficient estimators achieving optimal estimation rates for many problems (Diakonikolas et al., 2016; Cheng et al., 2019; Diakonikolas and Kane, 2023). In computational learning theory, older work has explored probably approximately correct (PAC) learning framework with adversarially corrupted labels (Angluin and Laird, 1988; Bshouty et al., 2002).

Robust optimal transport. The robust optimal transport (OT) literature has a close connection with unbalanced OT theory, which deals with transportation problems between measures of different mass. Unbalanced OT problems involve f-divergences that account for differences in mass, which can appear either in the constraints (Balaji et al., 2020) or in the objective function as regularizers (Piccoli and Rossi, 2014; Chizat et al., 2018a; Liero et al., 2018; Schmitzer and Wirth, 2019; Hanin, 1992). The constraint version is usually more difficult to solve, whereas primal-dual type algorithms have been developed to solve the regularized version (Mukherjee et al., 2021; Chizat et al., 2018b; Fatras et al., 2021; Fukunaga and Kasai, 2022; Le et al., 2021; Nath, 2020). An alternative approach to model robustness in OT is through partial OT problems, where only a fraction of mass needs to be transported (Caffarelli and McCann, 2010; Figalli, 2010; Nietert et al., 2023a; Chapel et al., 2020). Partial OT has been previously used in the context of DRO problems; however, it was introduced to address stochastic programs with side information (Esteban-Pérez and Morales, 2022).

Sliced optimal transport. Max-sliced optimal transport, as used to define $\mathsf{W}_{1,k}$, is also known as k-dimensional optimal transport (Niles-Weed and Rigollet, 2022) and projection-robust optimal transport (Lin et al., 2020, 2021). In general, $\mathsf{W}_{1,k}$ defines a metric on the space of d-dimensional distributions by measuring discrepancy between k-dimensional projections thereof. The structural, statistical, and computational properties of $\mathsf{W}_{1,k}$ are well-studied in (Lin et al., 2021; Niles-Weed and Rigollet, 2022; Nietert et al., 2022b; Bartl and Mendelson, 2022; Nadjahi et al., 2020), with the tightest results established for $k = 1$. Max-sliced OT has been used in the context of DRO problems; however, it was introduced for rank-one linear structures (Olea et al., 2022).

Distributionally robust optimization. Wasserstein distributionally robust optimization has emerged as a powerful modeling approach for addressing uncertainty in the data generating distribution. In this approach, the ambiguity set around the empirical distribution is constructed by the Wasserstein distance. Modern convex approach, leveraging duality theory (Mohajerin Esfahani and Kuhn, 2018; Blanchet and Murthy, 2019; Gao and Kleywegt, 2023), has led to significant computational advantages. Despite its computational success, some studies have raised concerns about the sensitivity of the standard DRO formulations to outliers (Hashimoto et al., 2018; Hu et al., 2018; Zhu et al., 2022a). To address potential overfitting to outliers, Zhai et al. (2021) propose a refined risk function based on a family of f-divergences. Nevertheless, this approach is not robust to local perturbations, and the risk bounds require a moment condition to hold uniformly over Θ. Another related work in (Bennouna and Van Parys, 2022; Bennouna et al., 2023) constructs the ambiguity set using an f-divergence for statistical errors and the Prokhorov distance for outliers. This provides computational efficiency and statistical reliability but lacks analysis of minimax optimality and robustness to Huber contamination. Furthermore, Nietert et al. (2023b) constructs the ambiguity set using the robust Wasserstein distance introduced in (Nietert et al., 2021). We revisit this setting in Section 3.

1.3. Notation and Preliminaries

Let $\|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^d$. We write $\mathcal{P}(\mathbb{R}^d)$ for the family of Borel probability measures on $\mathbb{R}^d$, equipped with the TV norm between $P, Q \in \mathcal{P}(\mathbb{R}^d)$ defined by $\|P - Q\|_{\mathsf{TV}} := \frac{1}{2}|P - Q|(\mathcal{Z})$. We say that Q is an ε-deletion of P if $Q \le \frac{1}{1-\varepsilon} P$ (where such inequalities are set-wise). We write $\mathbb{E}_P[f(X)]$ for expectation of $f(X)$ with $X \sim P$; when clear from the context, the random variable is dropped and we write $\mathbb{E}_P[f]$. Let μ_P denote the mean and Σ_P the covariance matrix of $P \in \mathcal{P}(\mathbb{R}^d)$, and let $\mathcal{P}_p(\mathbb{R}^d) := \{P \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_P[\|X - \mu_P\|_2^p] < \infty\}$. The push-forward of f through $P \in \mathcal{P}(\mathbb{R}^d)$ is $f_\sharp P(\cdot) := P(f^{-1}(\cdot))$. 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, and let $a \vee b := \max\{a, b\}$. For a matrix $A \in \mathbb{R}^{d\times d}$, we write $\|A\|_{\mathrm{op}} := \sup_{\|x\|_2=1} \|Ax\|_2$ for its operator norm. If A is further is diagonalizable, we write $\lambda_{\max}(A) = \lambda_1(A) \ge \cdots \ge \lambda_d(A)$ for its eigenvalues.

Classical and outlier-robust Wasserstein distances. For $p \in [1, \infty)$, the p-*Wasserstein distance* between $P, Q \in \mathcal{P}_p(\mathbb{R}^d)$ is $\mathsf{W}_p(P, Q) := \inf_{\pi\in\Pi(P,Q)} \big(\mathbb{E}_\pi\big[\|X - Y\|^p\big]\big)^{1/p}$, where $\Pi(P, Q) := \{\pi \in \mathcal{P}(\mathcal{Z}^2) : \pi(\cdot \times \mathcal{Z}) = P,\ \pi(\mathcal{Z} \times \cdot) = Q\}$ is the set of all their couplings. Some basic properties of W_p are (see, e.g., Villani (2003); Santambrogio (2015)): (i) W_p is a metric on $\mathcal{P}_p(\mathcal{Z})$; (ii) the distance is monotone in the order, i.e., $\mathsf{W}_p \le \mathsf{W}_q$ for $p \le q$; and (iii) W_p metrizes weak convergence plus convergence of pth moments: $\mathsf{W}_p(P_n, P) \to 0$ if and only if $P_n \stackrel{w}{\to} P$ and $\int \|x\|^p dP_n(x) \to \int \|x\|^p dP(x)$. For a family of measures $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$, we write $\mathsf{W}_p(P, \mathcal{G}) := \inf_{R\in\mathcal{G}} \mathsf{W}_p(P, R)$.

To handle corrupted data, we employ the ε-*outlier-robust* p-*Wasserstein distance*[1], defined by

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

The second equality is a useful consequence of Lemma 4 in Nietert et al. (2023a) (see Appendix A of Nietert et al. (2023b) for further details).

1. While not a metric, W_p^ε is symmetric and satisfies an approximate triangle inequality (Nietert et al. (2023a), Proposition 3).

2. Robust Distribution Learning

We now turn to robust distribution estimation under combined TV-W_1 contamination. Given corrupted samples from an unknown distribution P, we aim to produce an estimate $\hat{P}$ such that $\mathsf{W}_{1,k}(\hat{P}, P)$ is appropriately small for all $k \in [d]$. When $k = d$, we recover standard W_1. When $k = 1$, we shall see that the resulting estimation task is of essentially the same complexity as mean estimation. Omitted proofs appear in Appendices A and B.

2.1. The Population Limit

We first examine the information-theoretic limits of this problem without sampling error, namely, allowing computationally-intractable estimators and access to population distributions (rather than samples). We consider learning under the following environment.

Setting A: Fix corruption levels $0 \le \varepsilon \le 0.49$[2] and $\rho \ge 0$, along with a clean distribution family $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$. Nature selects $P \in \mathcal{G}$, and the learner observes $\tilde{P}$ such that $\mathsf{W}_1^{\varepsilon}(\tilde{P}, P) \le \rho$.

Given $\tilde{P}$, we seek an estimator $\hat{P}$ such that $\mathsf{W}_{p,k}(\hat{P}, P)$ is small for all k. To ensure that effective learning is possible, we impose a stability condition on the clean measure P.

Definition 1 (Stability) *Let $0 < \varepsilon < 1$ and $\delta \ge \varepsilon$. We say that a distribution $P \in \mathcal{P}(\mathbb{R}^d)$ is (ε, δ)-stable if, for all $Q \le \frac{1}{1-\varepsilon} P$, we have $\|\mu_Q - \mu_P\|_2 \le \delta$ and $\|\Sigma_Q - \Sigma_P\|_{\mathrm{op}} \le \delta^2/\varepsilon$. Write $\mathcal{S}(\varepsilon, \delta)$ for the family of (ε, δ)-stable P such that $\Sigma_P \preceq I_d$, and $\mathcal{S}_{\mathrm{iso}}(\varepsilon, \delta)$ for the subfamily of those for which $\Sigma_P \succeq (1 - \delta^2/\varepsilon) I_d$.*

A distribution is stable if its first two moments vary minimally under ε-deletions. The near-isotropic subfamily $\mathcal{S}_{\mathrm{iso}}$ coincides with a popular definition in algorithmic robust statistics (see, e.g., Chapter 2 of Diakonikolas and Kane, 2023). Stability is a flexible notion that connects to many standard tail bounds.

Example 1 (Concrete stability bounds) *Fix $0 < \varepsilon \le 0.99$[3] and $P \in \mathcal{P}(\mathbb{R}^d)$ with $\Sigma_P \preceq I_d$. Then:* [proof]

- ***Bounded covariance**: with no further assumptions, $P \in \mathcal{S}_{\mathrm{iso}}(\varepsilon, O(\sqrt{\varepsilon}))$;*
- ***Sub-Gaussian**: if P is 1-sub-Gaussian, then $P \in \mathcal{S}(\varepsilon, O(\varepsilon\sqrt{\log(1/\varepsilon)}))$;*
- ***Log-concave**: if P is log-concave, then $P \in \mathcal{S}(\varepsilon, O(\varepsilon \log(1/\varepsilon)))$;*
- ***Bounded moments of order $q \ge 2$**: if $\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}_P[|v^\top(Z - \mu_P)|^q] \le 1$, then $P \in \mathcal{S}(\varepsilon, O(\varepsilon^{1-1/q}))$.*

We refer to the families of distributions satisfying these properties by $\mathcal{G}_{\mathrm{cov}}$, $\mathcal{G}_{\mathrm{subG}}$, $\mathcal{G}_{\mathrm{lc}}$, and $\mathcal{G}_q$, respectively. Note that $\mathcal{G}_{\mathrm{cov}} = \mathcal{G}_2$. Similar bounds are derived in Chapter 2 of Diakonikolas and Kane (2023).

We now present our primary risk bound for the population-limit.

Theorem 2 (Population-limit risk bound) *Under Setting A, take $\eta = \min\{2\varepsilon, 1/4 + \varepsilon/2\}$ and assume that $\mathcal{G} \subseteq \mathcal{S}(2\eta, \delta)$. Then the minimum distance estimate*[4] *$\hat{P}_{\mathrm{MDE}} = \mathrm{argmin}_{Q \le \frac{1}{1-\eta}} \mathsf{W}_2(Q, \mathcal{G})$ satisfies*

$$\mathsf{W}_{1,k}(\hat{P}_{\mathrm{MDE}}, P) \lesssim \sqrt{k}\delta + \rho, \quad \forall k \in [d].$$

2. As ε approaches the optimal breakdown point of 1/2, it becomes information-theoretically impossible to distinguish inliers from outliers. The quantity 0.49 can be replaced with any constant bounded away from 1/2
3. Similarly to the $\varepsilon \le 0.49$ bound above, here 0.99 can be replaced with any constant bounded away from 1.
4. Here and throughout, the existence of such minimizers is not consequential but simply assumed for cleaner statements. Approximate minimizers up to some additive error provide the same risk bounds up to said error.

The minimum distance estimate $\hat{P}_{\mathrm{MDE}}$ involves an infinite dimensional optimization problem, which is computationally intractable. In the subsequent subsection we propose an iterative filtering algorithm that approximately solves a surrogate optimization problem on a finite sample set.

Proof The constant η is selected so that $\eta - \varepsilon \gtrsim \varepsilon$ while keeping $2\eta \leq 0.99$ bounded away from 1. For concreteness, the reader may want to focus on the case where $\varepsilon \leq 1/6$ and $\eta = 2\varepsilon$.

To treat combined Wasserstein and TV contamination, we first show that any W_1 perturbation can be decomposed into a W_2 perturbation followed by a TV perturbation.

[proof] **Lemma 3 (W_1 decomposition)** *Fix $0 < \tau < 1$ and $P, Q \in \mathcal{P}(\mathbb{R}^d)$ with $\mathsf{W}_1(P,Q) \leq \rho$. Then there exists $R \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_1(P,R) \leq \rho$, $\mathsf{W}_2(P,R) \leq \sqrt{2}\rho/\sqrt{\tau}$, and $\|R - Q\|_{\mathsf{TV}} \leq \tau$.*

The proof in Appendix A.2 takes $Z \sim P$ and $Z + \Delta \sim Q$, where $\mathbb{E}[\|\Delta\|_2] \leq \rho$. Letting E be the event such that $\|\Delta\|_2$ is less than its $1-\tau$ quantile, we conclude by setting R to the law of $Z + \Delta \mathbb{1}_E$.

Next, we prove a version of the theorem when $\rho = 0$, but the clean measure is close to $\mathcal{S}(\varepsilon, \delta)$ under W_2.

[proof] **Lemma 4 (Risk bound when $\rho = 0$)** *Fix $0 \leq \eta < 1/2$, $\lambda \geq 0$, and $\mathcal{G} \subseteq \mathcal{S}(2\eta, \delta)$. Take $R, \tilde{R} \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_2(R, \mathcal{G}) \leq \lambda$ and $\|R - \tilde{R}\|_{\mathsf{TV}} \leq \eta$. Then the estimate $\hat{R} = \operatorname{argmin}_{Q \leq \frac{1}{1-\eta}\tilde{R}} \mathsf{W}_2(Q, \mathcal{G})$ satisfies $\mathsf{W}_{1,k}(\hat{R}, R) \lesssim \frac{1}{1-2\eta}(\sqrt{k}\delta + \lambda\sqrt{\eta})$, for all $k \in [d]$.*

The proof in Appendix A.3 observes that R satisfies a generalized stability bound: for all $R' \leq \frac{1}{1-\varepsilon}R$ and $M \succeq 0$, we have $\|\mu_{R'} - \mu_R\|_2 \lesssim \delta + \lambda\sqrt{\varepsilon}$ and

$$\left|\operatorname{tr}\left(M(\Sigma_{R'} - \Sigma_R)\right)\right| \lesssim \frac{\delta^2}{\varepsilon}\operatorname{tr}(M) + \lambda^2\|M\|_{\mathrm{op}}. \tag{3}$$

In words, the latter bounds shows that the gap $\Sigma_{R'} - \Sigma_R$ lies in the Minkowski sum of an operator norm ball of radius $O(\delta^2/\varepsilon)$ and a trace norm ball of radius $O(\lambda^2)$. In contrast, the more direct guarantee $\|\Sigma_{R'} - \Sigma_R\|_{\mathrm{op}} \lesssim \delta^2/\varepsilon + \lambda^2$ would only give a suboptimal risk bound of $\sqrt{k}\delta + \sqrt{k}\lambda$.

Given the above lemmas, we are ready to prove the theorem. Fix $P \in \mathcal{G}$ and $\tilde{P}$ such that $\mathsf{W}_1^{\varepsilon}(\tilde{P}, P) \leq \rho$. This requires the existence of Q such that $\mathsf{W}_1(P,Q) \leq \rho$ and $|Q - \tilde{P}|_{\mathsf{TV}} \leq \epsilon$. Applying Lemma 3 to P and Q with $\tau = \eta - \varepsilon$ implies that there exists R such that $\mathsf{W}_1(P,R) \leq \rho$, $\mathsf{W}_2(P,R) \leq \sqrt{2}\rho/\sqrt{\eta - \varepsilon} =: \lambda$, and $\|R - \tilde{P}\|_{\mathsf{TV}} \leq \|R - Q\|_{\mathsf{TV}} + \|Q - \tilde{P}\|_{\mathsf{TV}} \leq \eta - \varepsilon + \varepsilon = \eta$. Applying Lemma 4 with TV corruption level η and W_2 bound $\lambda \lesssim \rho/\sqrt{\varepsilon}$, we find that $\hat{P}_{\mathrm{MDE}}$ from the theorem statement satisfies

$$\mathsf{W}_{1,k}(\hat{P}_{\mathrm{MDE}}, P) \leq \mathsf{W}_{1,k}(\hat{P}_{\mathrm{MDE}}, R) + \mathsf{W}_{1,k}(R, P) \lesssim \frac{1}{1-2\eta}\left(\sqrt{k}\delta + \lambda\sqrt{\eta}\right) \lesssim \frac{1}{1-2\eta}\left(\sqrt{k}\delta + \rho\right),$$

as desired. ■

This bound is tight for many distribution families, including those in Example 1.

[proof] **Corollary 5** *The minimum distance estimate $\hat{P}_{\mathrm{MDE}}$ from Theorem 2 achieves error*

$$\mathsf{W}_{1,k}(\hat{P}_{\mathrm{MDE}}, P) \lesssim \begin{cases} \sqrt{k\varepsilon} + \rho, & \mathcal{G} = \mathcal{G}_{\mathrm{cov}} \\ \sqrt{k}\varepsilon\sqrt{\log(1/\varepsilon)} + \rho, & \mathcal{G} = \mathcal{G}_{\mathrm{subG}} \\ \sqrt{k}\varepsilon\log(1/\varepsilon) + \rho, & \mathcal{G} = \mathcal{G}_{\mathrm{lc}} \\ \sqrt{k}\varepsilon^{1-1/q} + \rho, & \mathcal{G} = \mathcal{G}_q \end{cases},$$

and each of these guarantees is minimax optimal up to logarithmic factors in ε^{-1}.

For the minimax lower bounds, we employ existing constructions for the setting where $\rho = 0$. To strengthen these bounds when $\rho > 0$, we show that the learner cannot distinguish between translations of magnitude ρ.

Remark 6 (Comparison to other minimum distance estimators) *Estimators related to that in Theorem 2 are standard in robust statistics (see, e.g., Donoho and Liu (1988); Zhu et al. (2022a) for methods based on (smoothed) TV projection) and robust optimal transport (see, e.g., Nietert et al. (2023a), which employs projection under W_p^ε). The risk bounds from Lemma 4 match those in the literature for robust mean and distribution estimation when $\rho = 0$ (recalling that our results extend to mean estimation since $\|\mu_P - \mu_Q\|_2 \leq \mathsf{W}_{1,1}(P, Q)$). We diverge from these existing estimators by returning $\hat{P} = \mathsf{T}(\tilde{P})$ which lies not in $\mathcal{G}$ but nearby $\mathcal{G}$ under W_2. The fact that $\hat{P}$ is an ε-deletion of $\tilde{P}$ is essential in turning this approach into a practical algorithm in Section 2.2.*

2.2. Finite-Sample Algorithms

We now turn to the finite-sample setting. Here, our rates are only tight when $\delta \gtrsim \sqrt{\varepsilon}$, so we restrict to the family $\mathcal{G}_{\mathrm{cov}}$ of distributions $P \in \mathcal{P}(\mathbb{R}^d)$ with $\Sigma_P \preceq I_d$. Indeed, $\mathcal{G}_{\mathrm{cov}} \subseteq \mathcal{S}(\varepsilon, O(\sqrt{\varepsilon}))$ by Example 1.

Setting B: Let $0 < \varepsilon < \varepsilon_0$, where ε_0 is a sufficiently small absolute constant[5]. Fix $\rho \geq 0$ and sample size $n = \Omega(d \log d/\varepsilon)$. Nature samples $X_1, \ldots, X_n$ i.i.d. from $P \in \mathcal{G}_{\mathrm{cov}}$, with empirical measure P_n. The learner observes $\tilde{X}_1, \ldots \tilde{X}_n$ with empirical measure $\tilde{P}_n$ such that $\mathsf{W}_p^\varepsilon(\tilde{P}_n, P_n) \leq \rho$.

We aim to match the bound of Theorem 2, computing an estimate $\hat{P}_n$ such that $\mathsf{W}_{1,k}(\hat{P}_n, P) \lesssim \sqrt{k\varepsilon} + \rho$ for sufficiently large n. In order to turn the W_2 projection procedure into an efficient algorithm, we replace the intractable objective $\mathsf{W}_2(Q, \mathcal{G}_{\mathrm{cov}})$ with the tractable trace norm objective $\mathrm{tr}(\Sigma_Q - I_d)_+ = \sum_i [\lambda_i(\Sigma_Q) - 1]_+$, which can be computed via eigen-decomposition.

Lemma 7 (Trace norm comparison) *For $Q \in \mathcal{P}(\mathbb{R}^d)$, we have* [proof]

$$\tfrac{1}{2}\,\mathrm{tr}(\Sigma_Q - 2I_d)_+ \leq \mathsf{W}_2\big(Q, \mathcal{G}_{\mathrm{cov}}\big)^2 \leq \mathrm{tr}(\Sigma_Q - I_d)_+.$$

This result underlies W2PROJECT (Algorithm 1), which approximately solves the optimization problem $\min_{Q \leq \frac{1}{1-O(\varepsilon)}\tilde{P}_n} \mathrm{tr}(\Sigma_Q - \sigma^2 I_d)_+$ using a variant of iterative filtering (Diakonikolas et al., 2016). In the algorithm description, we identify a multiset $T \subseteq \mathbb{R}^d$ with the uniform distribution $\mathrm{Unif}(T)$.

Theorem 8 *Under Setting B, W2PROJECT$(\tilde{P}_n, \varepsilon, \rho)$ returns $\hat{P}_n$ in time $\mathrm{poly}(n, d)$ such that* [proof]

$$\mathsf{W}_{1,k}(\hat{P}_n, P) \lesssim \sqrt{k\varepsilon} + \rho + \mathsf{W}_{1,k}(P, P_n), \qquad \forall k \in [d].$$

with probability at least $2/3$.

Over $P \in \mathcal{G}_{\mathrm{cov}} \subseteq \mathcal{S}(\varepsilon, O(\sqrt{\varepsilon}))$, this guarantee attains the minimax optimal error from Corollary 5 as the sample size n increases (whence the empirical estimation error vanishes). Our proof shows that the estimate $\hat{P}_n$ satisfies $\mathrm{tr}(\Sigma_{\hat{P}_n} - O(1)I_d)_+ \lesssim \rho^2/\varepsilon$, mirroring the martingale-based analysis of iterative filtering with the simpler objective $\lambda_{\max}(\Sigma_Q)$; see, e.g., Section 2.4 of Diakonikolas and Kane (2023). Via Lemma 7, we then convert this trace norm bound into a W_2 bound, and proceed with the analysis of the W_2 projection

5. We make no effort to optimize the breakdown point ε_0. Similar results for robust mean estimation first required $\varepsilon_0 \ll 1/2$, but this was later alleviated (Hopkins et al., 2020; Zhu et al., 2022b; Dalalyan and Minasyan, 2022), We expect that similar improvements are possible under our setting but defer such optimization future work—see Section 4 for additional discussion.

Algorithm 1: W2PROJECT

Input: Contamination levels ε and ρ, uniform discrete measure $\tilde{P}_n$ supported on $T \subseteq \mathbb{R}^d$
Output: Uniform discrete measure $\hat{P}_n$

$\sigma \leftarrow 50, C \leftarrow 10^{10}$
Compute eigen-decomposition $\lambda_1, \dots, \lambda_d \in \mathbb{R}$, $v_1, \dots, v_d \in \mathbb{R}^d$ of $\Sigma_T - \sigma^2 I_d$
$\Pi \leftarrow \sum_{i:\lambda_i \geq 0} v_i v_i^\top$
if $\mathrm{tr}(\Pi(\Sigma_T - \sigma^2 I_d)) < C\varepsilon + C\rho^2/\varepsilon$ **then return** $\hat{P}_n = \mathrm{Unif}(T)$ `// LHS equals` $\mathrm{tr}(\Sigma_T - \sigma^2 I_d)_+$
else
  $g(x) \leftarrow \|\Pi(x - \mu_T)\|_2^2$ for $x \in T$
  Let $L \subseteq T$ be set of $6\varepsilon|T|$ points for which $g(x)$ is largest
  $f(x) \leftarrow g(x)$ for $x \in L$ and $f(x) \leftarrow 0$ otherwise
  Remove each point $x \in T$ from T with probability $f(x)/\max_{x\in T} f(x)$
  Return to Step 1 with new set T
end

from Theorem 2 to arrive at the risk bound above. As with Theorem 2, the generalized stability bound (3) is essential for avoiding a $\sqrt{k}\rho$ dependence.

The remaining empirical convergence term, $\mathsf{W}_{1,k}(P, P_n)$, can always be bounded by $\mathsf{W}_1(P, P_n)$, and the covariance bound implies that $\mathbb{E}[\mathsf{W}_1(P, P_n)] = \tilde{O}(\sqrt{d}n^{-1/d})$ for $d \geq 2$ (see, e.g., Theorem 3.1 of LEI, 2020). Generally, we would hope for a faster $n^{-1/k}$ rate, and this is indeed the case under appropriate additional assumptions on the clean distribution P. To name a few instances, Lin et al., 2021 derive such rates for general k under a Bernstein tail condition or a Poincaré inequality assumption, while Niles-Weed and Rigollet, 2022 provide rates when P satisfies a transport inequality (Niles-Weed and Rigollet, 2022) (which, in particular, holds for sub-Gaussian distributions). Empirical convergence rates in additional settings have been derived in the $k = 1$ case, e.g., for log-concave distributions (Nietert et al., 2022a) and under certain isotropic and moment boundedness assumptions (Bartl and Mendelson, 2022). For concreteness, below we instantiate the bound from Theorem 8 to the case of sub-Gaussian P, relying on the rate bounds form (Niles-Weed and Rigollet, 2022).

[proof] **Corollary 9 (Statistical performance)** *Under Setting B with $P \in \mathcal{G}_{\mathrm{subG}}$,* W2PROJECT$(\tilde{P}_n, \varepsilon, \rho)$ *returns $\hat{P}_n$ in time* $\mathrm{poly}(n, d)$ *such that*

$$\mathsf{W}_{1,k}(\hat{P}_n, P) \lesssim \sqrt{k\varepsilon} + \rho + d^{2.5} n^{-\frac{1}{k \vee 2}} \log(n)^{\mathbb{1}\{k=2\}} + \sqrt{\frac{d^5 k \log n}{n}}, \qquad \forall k \in [d].$$

with probability at least $2/3$.

Here, one multiple of d in the latter two terms arrives from taking a union bound over k. If one has a fixed slicing dimension in mind, a priori, this factor can be eliminated. Another factor of d arises from converting sub-Gaussianity into a $p = 1$ transport inequality.

Remark 10 (Recovering standard filtering via sliced W_2 projection) *We note that Lemma 7 can be adapted to the sliced W_2 setting. In particular, one can approximate $\mathsf{W}_{2,1}(Q, \mathcal{G}_{\mathrm{cov}})$ by the operator norm $\|(\Sigma_Q - I_d)_+\|_{\mathrm{op}} = [\|\Sigma_Q\|_{\mathrm{op}} - 1]_+$. This is equivalent to the standard objective $\|\Sigma_Q\|_{\mathrm{op}}$ for iterative filtering (Diakonikolas et al., 2016) (when $\|\Sigma_Q\|_{\mathrm{op}} > 1$, and otherwise the algorithm will have terminated), providing a new perspective on this standard algorithm.*

2.3. Other Corruption Models and Robust Mean Estimation

We now discuss some complementary results. First, in the case that $\rho = 0$ and we only suffer TV corruption, standard iterative filtering resolves the question of efficient distribution learning for near-isotropic P.

Proposition 11 *Under Setting B with $\rho = 0$ and $P \in \mathcal{S}_{\mathrm{iso}}(4\varepsilon, \delta)$, any estimate $\hat{P} \leq \frac{1}{1-4\varepsilon}\tilde{P}$ such that $\|\Sigma_{\hat{P}}\|_{\mathrm{op}} \leq 1 + O(\delta^2/\varepsilon)$ satisfies $\mathsf{W}_{1,k}(\hat{P}, P) \lesssim \sqrt{k}\delta + \mathsf{W}_{1,k}(P, P_n)$, for all $k \in [d]$.* [proof]

Indeed, this $\lambda_{\max}$ bound is achieved by all stability-based algorithms for robust mean estimation (see, e.g., Theorem 2.11 of Diakonikolas and Kane, 2023). Our proof employs a refined version of the certificate lemma for stable distributions (see Lemma 2.7 of Diakonikolas and Kane, 2023). The isotropic restriction is standard in algorithmic robust statistics; hardness results suggest it cannot be eliminated without imposing further assumptions like Gaussianity or losing computational tractability (Hopkins and Li, 2019).

Next, we comment on the simpler task of robust mean estimation. Under Setting B, we can simply return the mean of $\hat{P}_n = \mathrm{W2PROJECT}(\tilde{P}_n, \varepsilon, \rho)$ to obtain error $O(\sqrt{\varepsilon} + \rho)$. It is not hard to show that standard iterative filtering also suffices, employing the W_1 decomposition in Lemma 3 to bound the extent to which the Wasserstein corruption can perturb second moments after filtering out its ε-tails. However, neither approach generalizes to $\mathcal{S}_{\mathrm{iso}}(\varepsilon, \delta)$, leaving us with a simple open question: *under Setting B with $P = \mathcal{N}(\mu, I_d)$ and $n = \mathrm{poly}(d, 1/\varepsilon)$, can one efficiently compute $\hat{\mu}$ from $\tilde{P}_n$ such that $\|\hat{\mu} - \mu\|_2 = \tilde{O}(\varepsilon + \rho)$?*

3. Robust Stochastic Optimization

Finally, we present an application to robust stochastic optimization. We consider a setting where the learner seeks to make a decision $\hat{\theta} \in \Theta$ that performs well on a data distribution P, given only a corrupted observation $\tilde{P}_n$. More precisely, given a loss function $L : \Theta \times \mathbb{R}^d \to \mathbb{R}$, we seek to minimize the risk $\mathbb{E}_P[L(\hat{\theta}, X)]$. In the following we suppress dependence of L on the model parameters $\theta \in \Theta$ and write $\ell(\cdot) = L(\theta, \cdot)$ for a specific function. We also introduce the set $\mathcal{L} = \{L(\theta, \cdot)\}_{\theta \in \Theta}$ for the whole class, and impose the following.

Assumption 1 *Fix $p \geq 1$. Take $\mathcal{L}$ to be a family of real-valued loss functions on $\mathbb{R}^d$, such that each $\ell \in \mathcal{L}$ is of the form $\ell = \underline{\ell} \circ A$, where $A : \mathbb{R}^d \to \mathbb{R}^k$ is affine and $\underline{\ell} : \mathbb{R}^k \to \mathbb{R}^d$ is l.s.c. with $\sup_{z \in \mathcal{Z}} \frac{\ell(z)}{1 + \|z\|^p} < \infty$.*

In addition to mild regularity conditions, we assume that the loss functions operate on k-dimensional linear features of the data. For example, this captures k-variate least-squares regression if $\Theta \subseteq \mathbb{R}^{d \times k}$ and L maps $\theta \in \Theta$ and $(x, y) \in \mathbb{R}^{d-k} \times \mathbb{R}^k$ to $L(\theta, (x, y)) = \|\theta x - y\|_2^2$. This structural assumption is not restrictive as we may always set $k = d$.

If it is known that $\tilde{P}_n$ and P are close under W_p, a popular decision-making procedure is *Wasserstein distributionally robust optimization (WDRO)*, which selects

$$\hat{\ell}_{\mathrm{WDRO}} \coloneqq \mathrm{argmin}_{\ell \in \mathcal{L}} \sup_{Q : \mathsf{W}_p(Q, \tilde{P}_n) \leq r} \mathbb{E}_Q[\ell].$$

Indeed, if $\mathsf{W}_p(P, \tilde{P}_n) \leq r$, then it is easy to prove the excess risk bound

$$\mathbb{E}_P[\hat{\ell}_{\mathrm{WDRO}}] - \mathbb{E}_P[\ell_\star] \lesssim \sup_{\mathsf{W}_p(Q, P) \leq 2r} \mathbb{E}_Q[\ell_\star] - \mathbb{E}_P[\ell_\star] =: \mathcal{R}_p(\ell_\star; P, 2r), \tag{4}$$

where $\ell_\star = \mathrm{argmin}_{\ell \in \mathcal{L}} \mathbb{E}_P[\ell]$. The right-hand side, denoted $\mathcal{R}_p$, is termed the *p-Wasserstein regularizer* and characterizes a certain variational complexity of the optimal loss function (see, e.g., Gao and Kleywegt, 2023). In particular, we have $\mathcal{R}_1(\ell; P, r) \leq r \|\ell\|_{\mathrm{Lip}}$.

Alas, the assumption that $\mathsf{W}_p(P, \tilde{P}_n)$ is small is quite conservative, especially in the high-dimensional setting, where Wasserstein empirical convergence rates suffer from the curse of dimensionality. In fact, given the low-dimensional structure imposed in Assumption 1, it is natural to expect that a much smaller Wasserstein radius would suffice, e.g., as captured by the k-dimensional sliced distance. The next theorem indeed shows that the inner WDRO maximization problem automatically adapts to the dimensionality of a given loss function, which provide a new perspective on beating the curse of dimensionality in WDRO, as discussed in detail at the end of this section.

Theorem 12 *Fix $P, \hat{P} \in \mathcal{P}_p(\mathbb{R}^d)$ with $\mathsf{W}_{p,k}(P, \hat{P}) \leq \tau$, for some $\tau \geq 0$. Under Assumption 1, we have*

$$\mathbb{E}_P[\ell] \leq \sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_p(\hat{P}, Q) \leq \tau} \mathbb{E}_Q[\ell]$$

for each $\ell \in \mathcal{L}$.

Proof Take $\ell = \underline{\ell} \circ A$ to be the decomposition guaranteed by Assumption 1. Assume without loss of generality that A is linear. By the QR decomposition, we can rewrite ℓ as $\underline{\ell} \circ BU$ for $U \in \mathbb{R}^{k \times d}$ such that $UU^\top = I_k$. Take $\tilde{\ell} = \underline{\ell} \circ B$. We now show that

$$\sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_p(\hat{P}, Q) \leq \tau} \mathbb{E}_Q[\ell] = \sup_{R \in \mathcal{P}(\mathbb{R}^k): \mathsf{W}_p(U_\sharp \hat{P}, R) \leq \tau} \mathbb{E}_R[\tilde{\ell}]. \tag{5}$$

For the "$\leq$" direction, note that for any feasible Q for the left supremum, $R = U_\sharp Q$ is feasible for the right hand side with equal objective value. For the "$\geq$" direction, take any R feasible for the right supremum. Let (UX, Y) be an optimal coupling for the $\mathsf{W}_p(U_\sharp \hat{P}, R)$ problem, where $X \sim \hat{P}$. Taking Q to be the law of $X + U^\top(Y - UX)$, we have that $\mathsf{W}_p(\hat{P}, Q)^p \leq \mathbb{E}[\|U^\top(Y - UX)\|_2^p] = \mathbb{E}[\|Y - UX\|_2^p] = \mathsf{W}_p(U_\sharp \hat{P}, R)^p \leq \tau$, and $\mathbb{E}_Q[\ell] = \mathbb{E}[\tilde{\ell}(UX + UU^\top(Y - UX))] = \mathbb{E}[\tilde{\ell}(Y)] = \mathbb{E}_R[\tilde{\ell}]$, as desired.

At this point, we note that $\mathsf{W}_p(U_\sharp P, U_\sharp \hat{P}) \leq \mathsf{W}_{p,k}(P, \hat{P}) \leq \tau$ and bound

$$\begin{aligned}
\mathbb{E}_P[\hat{\ell}] &= \mathbb{E}_{U_\sharp P}[\tilde{\ell}] \\
&\leq \sup_{R \in \mathcal{P}(\mathbb{R}^k): \mathsf{W}_1(R, U_\sharp \hat{P}) \leq \tau} \mathbb{E}_R[\tilde{\ell}] \\
&= \sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_1(Q, \hat{P}) \leq \tau} \mathbb{E}_Q[\hat{\ell}] - \mathbb{E}_P[\ell_\star], && (5)
\end{aligned}$$

as desired. ■

Theorem 12 implies that we may center the WDRO procedure around any distribution $\hat{P}$, for which we have control over its $\mathsf{W}_{1,k}$ distance from the true population P. Remarkably, the W2PROJECT algorithm provides a computationally efficient way to find such a distribution, and Theorem 8 further yields the required bound on the $\mathsf{W}_{1,k}$ error. We have the following.

Corollary 13 *Under Setting B and Assumption 1 with $p = 1$, take $\hat{P}_n = \text{W2PROJECT}(\tilde{P}_n, \varepsilon, \rho)$ and let τ be any upper bound on the error $\mathsf{W}_{1,k}(\hat{P}_n, P) \lesssim \sqrt{k\varepsilon} + \rho + \mathsf{W}_{1,k}(P, P_n)$. Then the WDRO estimate*

$$\hat{\ell} = \operatorname{argmin}_{\ell \in \mathcal{L}} \sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_p(\hat{P}_n, Q) \leq \tau} \mathbb{E}_Q[\ell]$$

satisfies the excess risk bound $\mathbb{E}_P[\ell] - \mathbb{E}[\ell_\star] \leq 2\|\ell_\star\|_{\mathrm{Lip}}\tau$, where $\ell_\star = \operatorname{argmin}_{\ell \in \mathcal{L}} \mathbb{E}_P[\ell]$.

Proof Let $\hat{\ell} = \tilde{\ell} \circ U$ denote the decomposition given by Theorem 12. Observe that $\mathsf{W}_1(U_\sharp P, U_\sharp \hat{P}_n) \leq \mathsf{W}_{1,k}(P, \hat{P}_n) \leq \tau$. We thus bound

$$\begin{aligned}
\mathbb{E}_P[\hat{\ell}] - \mathbb{E}_P[\ell_\star] &\leq \sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_1(Q, \hat{P}_n) \leq \tau} \mathbb{E}_Q[\hat{\ell}] - \mathbb{E}_P[\ell_\star] && \text{(Theorem 12)} \\
&\leq \sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_1(Q, \hat{P}_n) \leq \tau} \mathbb{E}_Q[\ell_\star] - \mathbb{E}_P[\ell_\star] && (\hat{\ell} \text{ minimizing}) \\
&\leq \sup_{Q \in \mathcal{P}(\mathbb{R}^d): \mathsf{W}_1(Q, P) \leq 2\tau} \mathbb{E}_Q[\ell_\star] - \mathbb{E}_P[\ell_\star] \\
&\leq \|\ell_\star\|_{\mathrm{Lip}} 2\tau,
\end{aligned}$$

as desired. ■

3.1. Beating the Curse of Dimensionality in WDRO

Despite promising applications, the classic Wasserstein DRO approach suffers from the curse of dimensionality. The rate of empirical convergence under the Wasserstein distance scales as $n^{-1/d}$, which cannot be generally improved when $d \geq 3$ (Fournier and Guillin, 2015; LEI, 2020). In light of this rate, Mohajerin Esfahani and Kuhn (2018) showed that if the WDRO radius is chosen as $\rho = O(n^{-1/d})$, then the worst-case expected loss over all distributions in the Wasserstein ambiguity set of that radius would be an upper bound for the expected loss with respect to the true data-generating distribution. This provided the first non-asymptotic guarantee for the Wasserstein robust solution, but the bound deteriorates exponentially fast as d grows.

To address the curse of dimensionality, an empirical likelihood approach was proposed in (Blanchet and Kang, 2021; Blanchet et al., 2022, 2019) to find the smallest radius ρ such that, with high probability, there exists $Q \in \mathcal{P}(\mathbb{R}^d)$ with $\mathsf{W}_p(Q, \hat{P}_n) \leq \rho$ and $\ell^\star \in \mathcal{L}$ satisfying

$$\ell^\star \in \operatorname{argmin}_{\ell \in \mathcal{L}} \mathbb{E}_Q[\ell] \cap \operatorname{argmin}_{\ell \in \mathcal{L}} \mathbb{E}_P[\ell].$$

This choice leads to a confidence region around the optimal solution, which enables working with a radius $\rho = O(n^{-1/2})$. However, this result is only asymptotic in nature, yet finite-sample bounds are crucial for applications. For certain WDROs with linear structure, such as linear regression/classification and kernelized versions thereof, non-asymptotic bounds of $O(n^{-1/2})$, have been established in (Shafieezadeh-Abadeh et al., 2019; Chen and Paschalidis, 2018; Olea et al., 2022; Wu et al., 2022). To relax these structural assumptions, Gao (2022) demonstrated that if the data-generating distribution satisfies a transport-entropy inequality, a radius of $O(n^{-1/2})$ is again sufficient. However, the transport-entropy inequality assumption on the unknown data distribution is restrictive and may be hard to verify in practice. Furthermore, the loss function requires to be α-smooth over the family $\mathcal{L}$ and admits a sub-root function in order to establish local Rademacher complexity bounds.

Corollary 13 present a clean route to overcome the curse of dimensionality in the classical Wasserstein DRO setting, when $\varepsilon = 0$, and obtain finite-sample results without relying on transport inequalities. Moreover, it provides a simple procedure for achieving the excess risk bounds for outlier-robust WDRO presented in Nietert et al. (2023b) when $p = 1$. In fact, the algorithm therein matches our $\sqrt{k\varepsilon} + \rho$ risk bound only when $k = \Theta(1)$ or $k = \Theta(d)$, but not in between. Their approach further requires solving new optimization problems that are more complicated than standard WDRO. Finally, the analysis in that work led to finite-sample excess risk bounds including a term scaling like $n^{-1/d}$ even when $k = O(1)$. The result of Theorem 13 accounts for all these limitations, yielding optimal rates uniformly in k via simple and computationally efficient procedures.

4. Concluding Remarks and Future Work

In this work, we have provided the first polynomial time algorithm for robust distribution estimation under combined Wasserstein and TV corruptions. In order to apply its guarantees to Wasserstein DRO, we uncovered a practical and conceptually simple technique for alleviating the curse of dimensionality that often manifests itself in this setting. There are numerous directions for future work; some of particular interest include:

- For distributions that are $(2\varepsilon, \delta)$-stable and isotropic, we have efficient algorithms for robust mean estimation up to error δ under TV ε-corruptions. Can we extend these results to obtain $\mathsf{W}_{p,k}$ estimation error $\delta\sqrt{k} + \rho$ under our combined model? For an even simpler challenge, as posed in Section 2.3 — can one estimate the mean of a spherical Gaussian up to ℓ_2 error $\tilde{O}(\varepsilon) + \rho$ with both W_1 and TV corruption? We suspect that there may be similar obstacles as those known for robust mean estimation with stable but non-isotropic distributions (Hopkins and Li, 2019).

- Relatedly, algorithms for robust mean estimation have been refined and optimized in many ways, improving their breakdown points (Hopkins et al., 2020; Zhu et al., 2022b; Dalalyan and Minasyan, 2022), running time (Cheng et al., 2019), and memory usage (Diakonikolas et al., 2017, 2022). We expect many of these improvements to translate to our model.

- Finally, for WDRO, can we tractably achieve dependence on the dimensionality $k_\star$ of the optimal loss function if $k_\star \ll k$ (recalling that k is a uniform bound over the loss function family)? We suspect this can be achieved by integrating the objective of W2PROJECT into the Wasserstein DRO ambiguity set.

References

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In *Advances in Neural Information Processing Systems*, 2020.

Daniel Bartl and Shahar Mendelson. Structure preservation via the wasserstein distance. *arXiv preprint arXiv:2209.07058*, 2022.

Amine Bennouna and Bart Van Parys. Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*, 2022.

Amine Bennouna, Ryan Lucas, and Bart Van Parys. Certified robust neural networks: Generalization and corruption resistance. In *International Conference on Machine Learning*, 2023.

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

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Jose Blanchet, Karthyek Murthy, and Nian Si. Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, 109(2):295–315, 2022.

Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Mathematics. Second Series*, 171(2):673–730, 2010.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, 2020.

Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(1):517–564, 2018.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *SIAM Symposium on Discrete Algorithms*, 2019.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018a.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018b.

Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *The Annals of Statistics*, 50(2):1193–1219, 2022.

Ilias Diakonikolas and Daniel M Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE Symposium on Foundations of Computer Science*, 2016.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, 2017.

Ilias Diakonikolas, Daniel M Kane, Ankit Pensia, and Thanasis Pittas. Streaming algorithms for high-dimensional robust statistics. In *International Conference on Machine Learning*, 2022.

David L. Donoho and Richard C. Liu. The "Automatic" Robustness of Minimum Distance Functionals. *The Annals of Statistics*, 16(2):552 – 586, 1988.

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

Kilian Fatras, Thibault Sejourne, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, 2021.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195: 533–560, 2010.

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

Takumi Fukunaga and Hiroyuki Kasai. Block-coordinate Frank-Wolfe algorithm and convergence analysis for semi-relaxed optimal transport problem. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

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

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

Leonid G. Hanin. Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352, 1992.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Samuel B. Hopkins and Jerry Li. How hard is robust mean estimation? In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory*, 2019.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, 2018.

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

Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. In *Advances in Neural Information Processing Systems*, 2021.

JING LEI. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust Wasserstein distance and Riemannian optimization. In *Advances in Neural Information Processing Systems*, 2020.

Tianyi Lin, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael I Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, 2021.

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

Debarghya Mukherjee, Aritra Guha, Justin Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, 2021.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, 2020.

J. Saketha Nath. Unbalanced optimal transport using integral probability metric regularization. *arXiv preprint arXiv:2011.05001*, 2020.

Sloan Nietert, Rachel Cummings, and Ziv Goldfeld. Outlier-robust optimal transport with applications to generative modeling and data privacy. In *Theory and Practice of Differential Privacy Workshop at ICML*, 2021.

Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced wasserstein distances. In *Advances in Neural Information Processing Systems*, 2022a.

Sloan Nietert, Ritwik Sadhu, Ziv Goldfeld, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, 2022b.

Sloan Nietert, Rachel Cummings, and Ziv Goldfeld. Robust estimation under the Wasserstein distance. *arXiv preprint arXiv:2302.01237*, 2023a.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust Wasserstein DRO. In *Advances in Neural Information Processing Systems*, 2023b.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

José Luis Montiel Olea, Cynthia Rush, Amilcar Velez, and Johannes Wiesel. On the generalization error of norm penalty linear regression models. *arXiv preprint arXiv:2211.07608*, 2022.

Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

Elvezio M Ronchetti and Peter J Huber. *Robust Statistics*. John Wiley & Sons Hoboken, 2009.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.

Bernhard Schmitzer and Benedikt Wirth. A framework for Wasserstein-1-type metrics. *Journal of Convex Analysis*, 26(2):353–396, 2019.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Cédric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

Qinyu Wu, Jonathan Yu-Meng Li, and Tiantian Mao. On generalization and regularization via Wasserstein distributionally robust optimization. *arXiv preprint arXiv:2212.05716*, 2022.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. DORO: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, 2021.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256 – 2283, 2022a.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Robust estimation via generalized quasi-gradients. *Information and Inference: A Journal of the IMA*, 11(2):581–636, 2022b.

Appendix A. Proofs for Section 2.1

Throughout this section, we prove results under a more general learning environment.

Setting A2: Fix TV corruption level $0 \le \varepsilon \le 0.49$ and W_p corruption level $\rho \ge 0$, where $p \in \{1, 2\}$. Let $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$. Nature selects $P \in \mathcal{G}$, and the learner observes $\tilde{P}$ such that $\mathsf{W}_p^\varepsilon(\tilde{P}, P) \le \rho$.

We begin with some auxiliary definitions and lemmas.

Definition 14 (Resilience) *For $P \in \mathcal{P}(\mathbb{R}^d)$ and $0 \le \varepsilon < 1$, the mean ε-resilience of P is given by*

$$\tau(P, \varepsilon) \coloneqq \sup_{Q \in \mathcal{P}(\mathbb{R}^d): Q \le \frac{1}{1-\varepsilon} P} \|\mu_Q - \mu_P\|_2.$$

For $p \ge 1$, the pth-order ε-resilience of $P \in \mathcal{P}_p(\mathbb{R}^d)$ is defined by $\tau_p(P, \varepsilon) \coloneqq \tau(f_\sharp P, \varepsilon)$, where $f(z) = \|z - \mu_P\|_2^p$. For a family $\mathcal{G} \subseteq \mathcal{P}(\mathbb{R}^d)$, we define $\tau(\mathcal{G}, \varepsilon) \coloneqq \sup_{P \in \mathcal{G}} \tau(P, \varepsilon)$ and $\tau_p(\mathcal{G}, \varepsilon) \coloneqq \sup_{P \in \mathcal{G}} \tau_p(P, \varepsilon)$.

It generally suffices to analyze resilience for ε bounded away from 1, due to the following result.

Lemma 15 *For each $P \in \mathcal{P}(\mathbb{R}^d)$ and $0 < \varepsilon < 1$, we have $\tau(P, 1 - \varepsilon) = \frac{1-\varepsilon}{\varepsilon} \tau(P, \varepsilon)$.*

Proof If $P = (1 - \varepsilon)Q + \varepsilon R$ for $Q, R \in \mathcal{P}(\mathbb{R}^d)$, we have

$$\|\mu_P - \mu_R\|_2 = \frac{1 - \varepsilon}{\varepsilon} \|\mu_P - \mu_Q\|_2 \le \frac{1 - \varepsilon}{\varepsilon} \tau(P, \varepsilon).$$

Supremizing over R gives one direction, and substituting $\varepsilon \leftarrow 1 - \varepsilon$ gives the other. ∎

We also observe a certain monotonicity of pth moment resilience terms in p.

Lemma 16 *Fix $1 \le p < q$, $0 \le \varepsilon < 1$, and $P \in \mathcal{P}(\mathbb{R}^d)$. Then $\tau(P, \varepsilon) \le \tau_p(P, \varepsilon)^{1/p} \le \tau_q(P, \varepsilon)^{1/q}$.*

Proof Take $X \sim P$ and $Y \sim Q$, for any $Q \in \mathcal{P}(\mathbb{R}^d)$ such that $Q \le \frac{1}{1-\varepsilon} P$. We then bound

$$\|\mu_Q - \mu_P\|_2 \le \mathbb{E}[\|Y - X\|_2] \le \big|\mathbb{E}[\|Y - \mu_P\|_2 - \|X - \mu_P\|_2]\big| \le \tau_1(P, \varepsilon).$$

Moreover, writing $a = \mathbb{E}[\|X - \mu_P\|_2^p]$, $b = \mathbb{E}[\|Y - \mu_P\|_2^p]$, and $r = q/p \ge 1$, we have

$$\big|\mathbb{E}[\|Y - \mu_P\|_2^p - \|X - \mu_P\|_2^p]\big| \le |a - b| \le |a^r - b^r|^{1/r} = \big|\mathbb{E}[\|Y - \mu_P\|_2^q - \|X - \mu_P\|_2^q]\big|^{p/q}.$$

Raising both sides to the $(1/p)$th power and supremizing over Q completes the proof. ∎

Stability essentially captures resilience in first and second moments.

Lemma 17 *Let $0 < \varepsilon < 1$ and $\delta \ge \varepsilon$. For $P \in \mathcal{S}(\varepsilon, \delta)$, we have $\tau(P, \varepsilon) \le \delta$ and $\tau_2(P, \varepsilon) \le 2d\delta^2/\varepsilon$.*

Proof Mean resilience follows directly from the definition of (ε, δ)-stability. For second moment resilience, we fix any $Q \le \frac{1}{1-\varepsilon} P$ and bound

$$\big|\operatorname{tr}(\Sigma_Q(\mu_P)) - \operatorname{tr}(\Sigma_P)\big| = |\operatorname{tr}(\Sigma_Q - \Sigma_P)| + \|\mu_Q - \mu_P\|_2^2 \le \frac{\delta^2}{\varepsilon} + \delta^2 \le \frac{2\delta^2}{\varepsilon}.$$

Supremizing over Q gives the lemma. ∎

Next, we compute useful resilience bounds for distributions that lie near $\mathcal{S}(\varepsilon, \delta)$ under W_2. For brevity, we define $\mathcal{S}(\varepsilon, \delta, \lambda) \coloneqq \{P \in \mathcal{P}(\mathbb{R}^d) : \mathsf{W}_2(p, \mathcal{S}(\varepsilon, \delta)) \le \lambda\}$.

Lemma 18 *Let $0 < \varepsilon < 1$, $\delta \geq \varepsilon$, and $\lambda \geq 0$. For $P \in \mathcal{S}(\varepsilon, \delta, \lambda)$, we have*

$$\tau(P, \varepsilon) \leq \delta + \frac{2\lambda\sqrt{\varepsilon}}{1-\varepsilon}$$
$$\tau_2(P, \varepsilon) \leq \frac{4d\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\lambda^2}{1-\varepsilon}$$

Finally, we have $\mathrm{tr}(\Sigma_P) \leq 2d + 4\lambda^2$ and, for any $Q \leq \frac{1}{1-\varepsilon}P$,

$$\mathrm{tr}(\Sigma_Q) \leq \frac{6d\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{20\lambda^2}{1-\varepsilon}.$$

Proof Fix $P_0 \in \mathcal{S}(\varepsilon, \delta)$ such that $\mathsf{W}_2(P_0, P) \leq \lambda$, and let X, Y be optimal coupling for $\mathsf{W}_2(P_0, P)$. Fix any $Q \leq \frac{1}{1-\varepsilon}P$. Augmenting the probability space if necessary, we can realize Q as the law of Y conditioned on an event E with probability $1-\varepsilon$. Write P_0' for the law of X conditioned on E. We then bound

$$\begin{aligned}
\|\mu_Q - \mu_P\|_2 &= \frac{\varepsilon}{1-\varepsilon}\| \mathbb{E}[Y] - \mathbb{E}[Y|E^c]\|_2 \\
&\leq \frac{\varepsilon}{1-\varepsilon}(\| \mathbb{E}[X] - \mathbb{E}[X|E^c]\|_2 + \| \mathbb{E}[Y] - \mathbb{E}[X]\|_2 + \| \mathbb{E}[Y|E^c] - \mathbb{E}[X|E^c]\|_2) \\
&\leq \frac{\varepsilon}{1-\varepsilon}\left(\| \mathbb{E}[X] - \mathbb{E}[X|E^c]\|_2 + \mathsf{W}_2(P_0, P) + \frac{1}{\sqrt{\varepsilon}}\mathsf{W}_2(P_0, P)\right) \\
&\leq \|\mu_{P_0} - \mu_{P_0'}\|_2 + \frac{2\sqrt{\varepsilon}}{1-\varepsilon}\lambda \\
&\leq \delta + \frac{2\sqrt{\varepsilon}}{1-\varepsilon}\lambda.
\end{aligned}$$

Supremizing over Q gives the mean resilience bound. Next, we use Minkowski's inequality to bound

$$\begin{aligned}
\left|\mathrm{tr}\big(\Sigma_P - \Sigma_{P_0}\big)\right| &= \left|\mathbb{E}[\|Y - \mathbb{E}[Y]\|_2^2] - \mathrm{tr}(\Sigma_{P_0})\right| \\
&\leq \left|\Big(\mathbb{E}[\|X - \mathbb{E}[Y]\|_2^2]^{\frac{1}{2}} + \mathbb{E}[\|Y - X\|_2^2]^{\frac{1}{2}} + \| \mathbb{E}[X] - \mathbb{E}[Y]\|_2\Big)^2 - \mathrm{tr}(\Sigma_{P_0})\right| \\
&\leq \left|\Big(\sqrt{\mathrm{tr}(\Sigma_{P_0})} + 2\lambda\Big)^2 - \mathrm{tr}(\Sigma_{P_0})\right| \\
&\leq 4\lambda\sqrt{\mathrm{tr}(\Sigma_{P_0})} + 4\lambda^2 \\
&\leq 4\lambda\sqrt{d} + 4\lambda^2, \qquad\qquad (\Sigma_{P_0} \preceq I_d)
\end{aligned}$$

which implies the desired bound on $\mathrm{tr}(\Sigma_P)$. The same argument via Minkowski's inequality gives

$$\left|\mathbb{E}[\|X - \mathbb{E}[X]\|_2^2|E^c] - \mathbb{E}[\|Y - \mathbb{E}[Y]\|_2^2|E^c]\right| \leq \frac{4\lambda\sqrt{d}}{\sqrt{\varepsilon}} + \frac{4\lambda^2}{\varepsilon}.$$

We then compute

$$\begin{aligned}
\left|\mathrm{tr}(\Sigma_Q(\mu_P) - \Sigma_P)\right| &= \left|\mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2 \mid E\right] - \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right]\right| \\
&= \frac{\varepsilon}{1-\varepsilon}\left|\mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right] - \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2 \mid E^c\right]\right| \\
&\le \frac{\varepsilon}{1-\varepsilon}\left(\left|\mathbb{E}\left[\|X - \mathbb{E}[X]\|_2\right] - \mathbb{E}\left[\|X - \mathbb{E}[X]\|_2 \mid E^c\right]\right| + \frac{8\lambda\sqrt{d}}{\sqrt{\varepsilon}} + \frac{8\lambda^2}{\varepsilon}\right) \\
&= \left|\mathbb{E}\left[\|X - \mathbb{E}[X]\|_2 \mid E\right] - \mathbb{E}\left[\|X - \mathbb{E}[X]\|_2\right]\right| + \frac{8\lambda\sqrt{\varepsilon d}}{1-\varepsilon} + \frac{8\lambda^2}{1-\varepsilon} \\
&\le \tau_2(P_0, \varepsilon) + \frac{8\lambda\sqrt{\varepsilon d}}{1-\varepsilon} + \frac{8\lambda^2}{1-\varepsilon} \\
&\le \frac{2d\delta^2}{\varepsilon} + \frac{8\lambda\sqrt{\varepsilon d}}{1-\varepsilon} + \frac{8\lambda^2}{1-\varepsilon} && \text{(Theorem 17)} \\
&\le \frac{1}{1-\varepsilon}\left(\frac{\sqrt{2d}\delta}{\sqrt{\varepsilon}} + \sqrt{8}\lambda\right)^2 && (\delta \ge \varepsilon) \\
&\le \frac{4d\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\lambda^2}{1-\varepsilon}
\end{aligned}$$

Supremizing over Q gives the second moment resilience bound. Finally, we bound

$$\begin{aligned}
\mathrm{tr}(\Sigma_Q) &= \mathrm{tr}(\Sigma_P) + \mathrm{tr}(\Sigma_Q(\mu_P) - \Sigma_P) - \|\mu_P - \mu_Q\|_2^2 \\
&\le \mathrm{tr}(\Sigma_P) + \tau_2(P, \varepsilon) \\
&\le 2d + 4\lambda^2 + \frac{4d\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\lambda^2}{1-\varepsilon} \\
&\le \frac{6d\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{20\lambda^2}{1-\varepsilon},
\end{aligned}$$

as desired.

■

Finally, we prove a technical lemma used throughout.

Lemma 19 *Fix $0 < \varepsilon < 1$ and $P \in \mathcal{S}(\varepsilon, \delta, \lambda)$. Suppose that $Q = (1-\varepsilon)P' + \varepsilon R$ for some $P', R \in \mathcal{P}(\mathbb{R}^d)$ such that $P' \le \frac{1}{1-\varepsilon}P$. Then, for $1 \le q \le 2$, we have*

$$\mathsf{W}_q(P, Q) \le \frac{7\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\mathrm{tr}(\Sigma_R(\mu_P))}.$$

Proof Write $P = (1-\varepsilon)P' + \varepsilon S$, for some $S \in \mathcal{P}(\mathbb{R}^d)$. For any $q \in [1, 2]$, we have

$$\begin{aligned}
\mathsf{W}_q(P, Q)^q &\le \varepsilon \mathsf{W}_q(S, R)^q \\
&\le 2^q \varepsilon\big(\mathsf{W}_q(S, \delta_{\mu_P})^q + \mathsf{W}_q(\delta_{\mu_P}, R)^q\big) \\
&\le 2^q \varepsilon\big(\mathsf{W}_q(P, \delta_{\mu_P})^q + \tau_q(P, 1-\varepsilon) + \mathsf{W}_q(\delta_{\mu_P}, R)^q\big).
\end{aligned}$$

Taking pth roots and applying Lemma 18, we obtain

$$
\begin{aligned}
\mathsf{W}_q(P,Q) &\leq 2\varepsilon^{\frac{1}{q}}\mathsf{W}_q(P,\delta_{\mu_P}) + 2\varepsilon^{\frac{1}{q}}\tau_q(P,1-\varepsilon)^{\frac{1}{q}} + 2\varepsilon^{\frac{1}{q}}\mathsf{W}_q(\delta_{\mu_P},R)\\
&\leq 2\varepsilon^{\frac{1}{q}}\mathsf{W}_2(P,\delta_{\mu_P}) + 2\varepsilon^{\frac{1}{q}}\sqrt{\tau_2(P,1-\varepsilon)} + 2\varepsilon^{\frac{1}{q}}\mathsf{W}_2(\delta_{\mu_P},R)\\
&\leq 2\varepsilon^{\frac{1}{q}}\sqrt{\mathrm{tr}(\Sigma_P)} + 2\varepsilon^{\frac{1}{q}-\frac{1}{2}}\sqrt{\tau_2(P,\varepsilon)} + 2\varepsilon^{\frac{1}{q}}\sqrt{\mathrm{tr}(\Sigma_R(\mu_P))}\\
&\leq 2\varepsilon^{\frac{1}{q}}\sqrt{2d+4\lambda^2} + 2\varepsilon^{\frac{1}{q}-\frac{1}{2}}\sqrt{\frac{4d\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\lambda^2}{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\mathrm{tr}(\Sigma_R(\mu_P))}\\
&\leq \frac{7\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\mathrm{tr}(\Sigma_R(\mu_P))},
\end{aligned}
$$

as desired. ■

As a consequence, we can bound the Wasserstein distance between a stable distribution and any of its ε-deletions. In Nietert et al. (2023a), this is called a "Wasserstein resilience" bound.

Lemma 20 (Wasserstein resilience from stability) *Fix $0<\varepsilon<1$, $P\in\mathcal{S}(\varepsilon,\delta,\lambda)$, and $Q\leq\frac{1}{1-\varepsilon}P$. Then, for $1\leq q\leq 2$, we have*

$$
\mathsf{W}_q(P,Q) \leq \frac{12\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{21\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}}.
$$

Proof Writing $P=(1-\varepsilon)Q+\varepsilon R$ for some $R\in\mathcal{P}(\mathbb{R}^d)$, we use Lemma 18 to bound

$$
\begin{aligned}
\mathrm{tr}(\Sigma_Q(\mu_P)) &\leq \mathrm{tr}(\Sigma_P) + \tau_2(P,\varepsilon)\\
&\leq 2d+4\lambda^2 + \frac{4d\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\lambda^2}{1-\varepsilon}\\
&\leq \frac{6d\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{20\lambda^2}{1-\varepsilon}.
\end{aligned}
$$

Noting that $Q=(1-\varepsilon)Q+\varepsilon Q$ and $Q\leq\frac{1}{1-\varepsilon}P$, we use Lemma 19 to bound

$$
\begin{aligned}
\mathsf{W}_q(P,Q) &\leq \frac{7\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\mathrm{tr}(\Sigma_Q(\mu_P))}\\
&\leq \frac{7\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\frac{6d\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{20\lambda^2}{1-\varepsilon}}\\
&\leq \frac{12\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{21\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}},
\end{aligned}
$$

as desired. ■

A.1. Proof of bounds in Example 1

We first observe that a distribution is (ε,δ)-stable if and only if all of its 1-dimensional orthogonal projections are (ε,δ)-stable. We thus assume that $d=1$ without loss of generality.

Next we prove a useful stability bound under the more general condition of an Orlicz norm bound. Recall that an Orlicz function is any convex, non-decreasing function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$. For a real random variable X, we define its Orlicz norm with respect to ψ by $\|X\|_\psi \coloneqq \sup\{\sigma \ge 0 : \mathbb{E}[\psi(|X|/\sigma)] \le 1\}$.

Lemma 21 *Suppose that $\|X - \mathbb{E}[X]\|_\psi \le \sigma$, where ψ is an Orlicz function satisfying $\psi(x) = \phi(x^2)$ for another Orlicz function ϕ. Then X is $\big(\varepsilon, O(\sigma\varepsilon\psi^{-1}(1/\varepsilon))\big)$-stable for $0 \le \varepsilon \le 0.99$.*

Proof We assume without loss of generality that $\mathbb{E}[X] = 0$. For mean resilience, take E to be any event with probability $1 - \varepsilon' \ge 1 - \varepsilon$, and bound

$$
\begin{aligned}
|\,\mathbb{E}[X|E]| &= \frac{\varepsilon'}{1-\varepsilon'}|\,\mathbb{E}[X|E^c]| \\
&\le \frac{\varepsilon'}{1-\varepsilon'}\,\mathbb{E}\big[|X| \bigm| E^c\big] \\
&\le \frac{\sigma\varepsilon'}{1-\varepsilon'}\psi^{-1}\left(\mathbb{E}\left[\psi\left(\frac{|X|}{\sigma}\right) \Bigm| E^c\right]\right) \\
&\le \frac{\sigma\varepsilon'}{1-\varepsilon'}\psi^{-1}\left(\frac{1}{\varepsilon'}\right) \\
&\le \frac{\sigma\varepsilon}{1-\varepsilon}\psi^{-1}\left(\frac{1}{\varepsilon}\right) =: \delta.
\end{aligned}
$$

We note that this approach is well-known; see, e.g., Lemma E.2 of Zhu et al. (2022a). For the second moment condition, we apply the bound above to X^2 (with $\|X^2\|_\phi \le \sigma^2$), deducing

$$
\begin{aligned}
|\,\mathbb{E}[X^2|E] - \mathbb{E}[X^2]| &\le \frac{\sigma^2\varepsilon}{1-\varepsilon}\phi^{-1}\left(\frac{1}{\varepsilon}\right) \\
&\le \frac{\sigma^2\varepsilon}{1-\varepsilon}\psi^{-1}\left(\frac{1}{\varepsilon}\right)^2 \\
&\le \frac{\delta^2}{\varepsilon}.
\end{aligned}
$$

Combining, we bound $|\mathrm{Var}[X|E] - \mathrm{Var}[X]| \le \delta^2/\varepsilon + \mathbb{E}[X|E]^2 \le 2\delta^2/\varepsilon$. Thus, X is $(\varepsilon, \sqrt{2}\delta)$-stable. ■

We now turn to the specific examples.

Bounded covariance: We apply the lemma with $\psi(x) = x^2$ and $\phi(x) = x$, giving $(\varepsilon, O(\sqrt{\varepsilon}))$-stability. The near-isotropic restriction is vacuous since $\delta \gtrsim \sqrt{\varepsilon}$.

Sub-Gaussian: We apply the lemma with $\psi(x) = \exp(x^2) - 1$ and $\phi(x) = \exp(x) - 1$, giving $(\varepsilon, O(\varepsilon\sqrt{\log(1/\varepsilon)}))$-stability.

Log-concave: If X is log-concave with bounded variance, then $\|X\|_{\psi_0} = O(1)$ for $\psi_0(x) = \exp(x) - 1$, by Borell's lemma. There is a very slight non-convexity to $\exp(\sqrt{x}) - 1$, which we remedy by taking

$$
\phi(x) \coloneqq \begin{cases} \exp(\sqrt{x}) - 1, & x \ge 1 \\ e/2 + e|x|/2 - 1, & 0 \le x < 1 \end{cases}
$$

and $\psi(x) = \phi(x^2)$. We still have $\|X\|_\psi = O(1)$, giving $(\varepsilon, O(\varepsilon\log(1/\varepsilon))$-stability.

Bounded qth moments, $q \geq 2$: We apply the lemma with $\psi(x) = x^q$, giving $(\varepsilon, O(\varepsilon^{1-1/q}))$-stability. ■

A.2. Proof of Lemma 3

We prove a stronger result.

Lemma 22 *Fix $0 < \tau < 1$, $1 \leq p < q$, and $P, Q \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_p(P,Q) \leq \rho$. Then there exists $R \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_p(P,R) \leq \rho$, $\mathsf{W}_q(P,R) \leq \left(\frac{q}{q-p}\right)^{1/q} \rho\tau^{1/q-1/p}$, and $\|R - Q\|_{\mathsf{TV}} \leq \tau$.*

Proof Let (X,Y) be an optimal coupling for $\mathsf{W}_p(P,Q)$, and let $\Delta = Y - X$. Writing τ for the $1-\tau$ quantile of $\|\Delta\|_2$, let E denote the event that $\|\Delta\|_2 \leq \tau$. By Markov's inequality, we have

$$\tau = \Pr(\|\Delta\|_2 > \tau) = \Pr(\|\Delta\|_2^p > \tau^p) \leq \frac{\rho^p}{\tau^p},$$

and so $\tau \leq \rho\tau^{-1/p}$. Consider the random variable Z which equals Y if $\|\Delta\|_2 \leq \tau$ and X otherwise. Taking R to be the law of Z, we have $\|R - Q\|_{\mathsf{TV}} \leq \tau$,

$$\mathsf{W}_p(P,R)^p \leq \mathbb{E}[\|X - Z\|_2^p] \leq \mathbb{E}[\|X - Y\|_2^p] = \rho^p,$$

and

$$\begin{aligned}
\mathsf{W}_q(P,R)^q &\leq \mathbb{E}[\|X - Z\|_2^q] \\
&\leq \mathbb{E}\big[\|\Delta\|_2^q \mid E\big] \\
&\leq \int_0^{\tau^q} \Pr\big[\|\Delta\|_2^q > t \mid E\big]\,\mathrm{d}t \\
&= \int_0^{\tau^q} \Pr\Big[\|\Delta\|_2^p > t^{p/q} \mid E\Big]\,\mathrm{d}t \\
&\leq \mathbb{E}[\|\Delta\|_2^p | E] \int_0^{\tau^q} t^{-p/q}\,\mathrm{d}t \\
&\leq \rho^p \frac{t^{1-p/q}}{1-p/q}\Big|_0^{\tau^q} \\
&= \frac{q}{q-p}\rho^p\tau^{q-p} \\
&\leq \frac{q}{q-p}\rho^p(\rho\tau^{-1/p})^{q-p} \\
&= \frac{q}{q-p}\rho^q\tau^{1-q/p}.
\end{aligned}$$

Taking qth roots gives the lemma. ■

As a consequence, we also have the following.

Lemma 23 *Fix $0 < \varepsilon < 1/2$, integers $p, q \geq 1$, and $P, Q \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_p^\varepsilon(P,Q) \leq \rho$. Then there exists $R \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_p(P,R) \leq \rho$, $\mathsf{W}_q(P,R) \leq \sqrt{2}\rho\varepsilon^{-[1/p-1/q]_+}$, and $\|R - Q\|_{\mathsf{TV}} \leq 2\varepsilon$.*

Proof By the W_p^ε bound, there exists $P' \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathsf{W}_p(P,P') \leq \rho$ and $\|P' - \tilde{P}\|_{\mathsf{TV}} \leq \varepsilon$. If $q \leq p$, then we can simply take $R = P'$. Otherwise, $q \geq p+1$, and we can apply Lemma 22 between P and P' to obtain R such that $\mathsf{W}_q(P,R) \leq \left(\frac{q}{q-p}\right)^{1/q}\rho\varepsilon^{1/q-1/p} \leq \sqrt{2}\rho\varepsilon^{-[1/p-1/q]_+}$ and $\|R - Q\|_{\mathsf{TV}} \leq \|R - P'\|_{\mathsf{TV}} + \|P' + \tilde{P}\|_{\mathsf{TV}} \leq 2\varepsilon$. ■

A.3. Proof of Lemma 4

Defining $\mathcal{S}(\varepsilon,\delta,\lambda) \coloneqq \{P \in \mathcal{P}(\mathbb{R}^d) : \mathsf{W}_2(P, \mathcal{S}(\varepsilon,\delta)) \le \lambda\}$, we prove a stronger statement.

Lemma 24 *Fix $0 \le \varepsilon < 1/2$ and $\mathcal{G} \subseteq \mathcal{S}(2\varepsilon,\delta,\lambda)$. Then, for all $1 \le q < 2$ and $k \in [d]$, we have $\mathcal{R}_{\mathsf{W}_{q,k}}(\mathcal{G}, \|\cdot\|_{\mathsf{TV}} \le \varepsilon) \lesssim \frac{1}{1-2\varepsilon}(\sqrt{k}\varepsilon^{1/q-1}\delta + \lambda\varepsilon^{1/q-1/2})$, for all $k \in [d]$. This risk is achieved by the minimum distance estimator T satisfying $\mathsf{T}(\tilde{P}) \in \operatorname{argmin}_{Q \le \frac{1}{1-\varepsilon}\tilde{P}} \mathsf{W}_2(Q,\mathcal{G})$, where $\mathsf{W}_2(Q,\mathcal{G}) \coloneqq \inf_{R\in\mathcal{G}} \mathsf{W}_2(Q,R)$.*

Proof Given $P \in \mathcal{G}$ and $\tilde{P}$ such that $\|\tilde{P} - P\|_{\mathsf{TV}} \le \varepsilon$, we decompose $\tilde{P} = (1-\varepsilon)P' + \varepsilon R$ for distribution P' and R such that $P' \le \frac{1}{1-\varepsilon}P$. Observe that the estimate $\hat{P} = \mathsf{T}(\tilde{P})$ satisfies

$$\begin{aligned}
\mathsf{W}_2(\hat{P},\mathcal{G}) &= \min_{Q \le \frac{1}{1-\varepsilon}\tilde{P}} \mathsf{W}_2(Q,\mathcal{G}) \\
&\le \mathsf{W}_2(P',\mathcal{G}) \\
&\le \mathsf{W}_2(P',P) \\
&\le \frac{12\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{21\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}}. && \text{(Lemma 20)}
\end{aligned}$$

Thus, $\hat{P} \in \mathcal{S}(2\varepsilon,\delta,\lambda')$, where

$$\lambda' \coloneqq \frac{12\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{21\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{\sqrt{1-\varepsilon}} + \lambda \le \frac{12\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{22\lambda}{\sqrt{1-\varepsilon}}.$$

Of course $P \in \mathcal{S}(2\varepsilon,\delta,\lambda')$ as well, and $\|\hat{P} - P\|_{\mathsf{TV}} \le \|\hat{P} - \tilde{P}\|_{\mathsf{TV}} + \|\tilde{P} - P\|_{\mathsf{TV}} \le 2\varepsilon$. Defining the midpoint distribution $Q = (1 - \|\hat{P} - P\|_{\mathsf{TV}})^{-1}\hat{P} \wedge P$, we again apply Lemma 20 to bound

$$\begin{aligned}
\mathsf{W}_q(\hat{P},P) &\le \mathsf{W}_q(\hat{P},Q) + \mathsf{W}_q(Q,P) \\
&\le \frac{24(2\varepsilon)^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-2\varepsilon}} + \frac{42(2\varepsilon)^{\frac{1}{q}-\frac{1}{2}}\lambda'}{\sqrt{1-2\varepsilon}} \\
&\le \frac{528\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{1-2\varepsilon} + \frac{924\varepsilon^{\frac{1}{q}-\frac{1}{2}}\lambda}{1-2\varepsilon},
\end{aligned}$$

as desired. For $k < d$, we observe that, for each orthogonal projection $U \in \mathbb{R}^{k\times d}$, $UU^\top = I_k$, we still have $\|U_\sharp P - U_\sharp \hat{P}\|_{\mathsf{TV}} \le 2\varepsilon$ and $U_\sharp P, U_\sharp \hat{P} \in \mathcal{S}(2\varepsilon,\delta,\lambda')$, so the analysis above can be applied in $\mathbb{R}^k$ to give the desired bound under $\mathsf{W}_{q,k}$. ■

A.4. Proof of Corollary 5

Upper bounds follow by Theorem 2. Matching lower bounds (up to logarithmic factors) when $\rho = 0$ are shown in Theorem 2 of Nietert et al. (2023a). It is straightforward to raise these lower bounds by the needed term of ρ. Indeed, suppose the learner observes $\tilde{P} = \delta_0$ but the adversary flips a fair coin to select $P = \delta_0$ or $P = \delta_x$ with $\|x\|_2 = \rho$. Then, no estimate $\hat{P}$ can incur expected risk less than that of $\hat{P} = \delta_{\rho/2}$, for which $\mathsf{W}_{1,k}(\hat{P},P) = \rho/2$. (See proof of Theorem 2 in Nietert et al. (2023b) for a more formal treatment of this two point method).

■

Appendix B. Proofs for Section 2.2

Throughout this section, we prove results under a more general learning environment.

Setting B2: Fix $\varepsilon > 0$ sufficiently small, $\rho \geq 0$, and $p \in \{1, 2\}$. Nature selects a distribution $P \in \mathcal{G}_{\text{cov}}$ and produces P' such that $\mathsf{W}_p(P', P) \leq \rho$ and $\mathsf{W}_2(P', P) \leq \rho\varepsilon^{1/2-1/p}$. The learner observes $\tilde{P}$ such that $\|\tilde{P} - P'\|_{\mathsf{TV}} \leq \varepsilon$. All of P, P', and $\tilde{P}$ are uniform discrete measures over n points in $\mathbb{R}^d$.

The W_2 bound is without loss of generality by Lemma 3 (no such decomposition is necessary when $p = 2$, where we may simply take $P' = P$).

B.1. Proof of Lemma 7

We first prove that $\mathsf{W}_2(Q, \mathcal{G}_{\text{cov}})^2 \leq \operatorname{tr}(\Sigma_Q - I_d)_+$. Assume without loss of generality that Q has mean 0. Write $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ for the eigenvalues of Σ_Q, with accompanying eigenvectors $v_1, \dots, v_d \in \mathbb{R}^d$. Define $A = \sum_{i=1}^d \big(1 \wedge (1/\sqrt{\lambda_i})\big) v_i v_i^\top$, so that $R = A_\sharp Q$ satisfies $\Sigma_R \preceq I_d$. Taking $X \sim Q$, we bound

$$
\begin{aligned}
\mathsf{W}_2(Q, \mathcal{G}_{\text{cov}})^2 &\leq \mathsf{W}_2(Q, R)^2 \\
&\leq \mathbb{E}[\|X - AX\|_2^2] \\
&= \mathbb{E}[\|(I - A)X\|_2^2] \\
&= \operatorname{tr}((I - A)^2 \Sigma_Q) \\
&= \sum_{i:\lambda_i > 1} \big(\sqrt{\lambda_i} - 1\big)^2 \\
&\leq \sum_{i:\lambda_i > 1} \lambda_i - 1 \\
&= \operatorname{tr}(\Sigma_\mu - I_d)_+,
\end{aligned}
$$

as desired.

Next, we prove that $\operatorname{tr}(\Sigma_Q - 2I_d)_+ \leq 2\mathsf{W}_2\big(Q, \mathcal{G}_{\text{cov}}\big)^2$. Suppose that $\mathsf{W}_2(Q, R) \leq \lambda$ for some $R \in \mathcal{G}_{\text{cov}}$. Assume without loss of generality that R has mean 0, and fix any $\mathbb{R}^{d \times d}$ with $\|A\|_{\text{op}} \leq 1$. Taking (AX, AY) to be an optimal coupling for the $\mathsf{W}_2(A_\sharp Q, A_\sharp R)$ problem, so that (X, Y) is a coupling of Q and R, we have

$$
\begin{aligned}
\operatorname{tr}(A^\top A \Sigma_Q) &= \mathbb{E}[\|A(X - \mathbb{E}[X])\|_2^2] \\
&\leq \mathbb{E}[\|AX\|_2^2] \\
&= \mathbb{E}[\|AY + A(X - Y)\|_2^2] \\
&\leq \left(\sqrt{\mathbb{E}[\|AY\|_2^2]} + \sqrt{\mathbb{E}[\|A(X - Y)\|_2^2]}\right)^2 && \text{(Minkowski's inequality)} \\
&= \left(\sqrt{\operatorname{tr}(A^\top A \Sigma_R)} + \mathsf{W}_2(A_\sharp Q, A_\sharp R)\right)^2 \\
&\leq \left(\sqrt{\operatorname{tr}(A^\top A \Sigma_R)} + \lambda\right)^2 && (\|A\|_{\text{op}} \leq 1) \\
&\leq \operatorname{tr}(A^\top A(2\Sigma_R)) + 2\lambda^2 \\
&\leq \operatorname{tr}(A^\top A 2 I_d) + 2\lambda^2. && (R \in \mathcal{G}_{\text{cov}}(\sigma))
\end{aligned}
$$

Rearranging and supremizing over A, we find that $\operatorname{tr}(\Sigma_Q - 2I_d)_+ \leq 2\lambda^2$, as desired. ■

B.2. Proof of Theorem 8

We shall prove a stronger statement under Setting B2. We require a slight change to the termination condition when $p = 2$. Specifically, we define the modified algorithm W2PROJECT2 by changing the termination condition at Step 4 from

$$\operatorname{tr}(\Pi(\Sigma_T - \sigma^2 I_d)) < C\varepsilon + C\rho^2/\varepsilon$$

to

$$\operatorname{tr}(\Pi(\Sigma_T - \sigma^2 I_d)) < C\varepsilon + C\rho^2\varepsilon^{1-2/p}.$$

For this algorithm (which matches W2PROJECT when $p = 1$), we prove the following.

Lemma 25 *Under Setting B2 with $\varepsilon_0 \leq 2^{-20}$,* W2PROJECT2$(\tilde{P}, \varepsilon, \rho)$ *returns $\hat{P}$ in time* $\operatorname{poly}(n, d)$ *such that, for all $q \in \{1, 2\}$ and $k \in [d]$,*

$$\mathsf{W}_{q,k}(\hat{P}, P) \lesssim \varepsilon^{\frac{1}{q}-\frac{1}{2}}\sqrt{k} + \varepsilon^{-\left[\frac{1}{p}-\frac{1}{q}\right]_+}\rho$$

with probability at least $2/3$.

Proof First, it is easy to show that the algorithm runs in polynomial time. At least one point is removed at each iteration, so there can be at most n iterations, and each iteration can be performed in $\operatorname{poly}(n, d)$ time.

Next, we show that W2PROJECT approximately minimizes its trace norm objective. We begin with a technical claim about the function f computed in Steps 6-8. This result mirrors Proposition 2.19 of Diakonikolas and Kane (2023), which pertains to the simpler objective $\lambda_{\max}(\Sigma_Q)$.

Lemma 26 *Under Setting B2, let $S \subseteq \mathbb{R}^d$ denote the support of P'. Suppose that, in some iteration of* W2PROJECT*, the multiset T satisfies $|T \cap S| \geq (1 - 4\varepsilon)|S|$ and $\operatorname{tr}(\Sigma_T - \sigma^2 I_d)_+ \geq 2^{33}\varepsilon + 2^{19}\rho^2\varepsilon^{1-2/p}$. Then the function f computed in Steps 6-8 satisfies $\sum_{x\in T} f(x) \geq 2\sum_{x\in T\cap S} f(x)$.*

Proof Under Setting B2, we have $\mathsf{W}_2(S, \mathcal{G}_{\mathrm{cov}}) \leq \mathsf{W}_2(S, P) \leq \rho\varepsilon^{1/2-1/p} =: \lambda$. Since $\mathcal{G}_{\mathrm{cov}} \subseteq \mathcal{S}_{\mathrm{iso}}(6\varepsilon, 5\sqrt{\varepsilon})$, as described in Example 1 (the constant can be obtained from Lemma 21), we have that $S \in \mathcal{S}(6\varepsilon, 5\sqrt{\varepsilon}, \lambda)$, using the notation from the proof of Lemma 4.

Now, at Step 6, W2PROJECT computes the function $g(x) = \|\Pi(x - \mu_T)\|_2^2$, where Π is the orthogonal projection onto the non-negative eigenspace of $\Sigma_T - \sigma^2 I_d$. At Step 7, we take L to be the set of $6\varepsilon|T|$ elements of T for which $g(x)$ is largest. Then, Step 8 takes $f(x) = g(x)$ for $x \in L$ and $f(x) = 0$ otherwise. Define $\eta \coloneqq \operatorname{tr}(\Sigma_T - \sigma^2 I_d)_+ = \operatorname{tr}(\Pi(\Sigma_T - \sigma^2 I_d))$, which, by the lemma assumption, satisfies $\eta \geq 2^{33}\varepsilon + 2^{19}\lambda^2$. This lower bound will be used later. We first compute

$$\sum_{x\in T} g(x) = |T| \operatorname{tr}(\Pi\Sigma_T) = |T|(\eta + \sigma^2 \operatorname{tr}(\Pi)).$$

Next, we apply Lemma 18 to $\Pi_\sharp S$, noting that S, and hence $\Pi_\sharp S$, belongs to $\mathcal{S}(6\varepsilon, 5\sqrt{\varepsilon}, \lambda)$. For any $S' \subseteq S$ with $|S'| \geq (1 - 6\varepsilon)|S|$, this gives

$$\|\mu_S - \mu_{S'}\|_2 \leq 5\sqrt{\varepsilon} + \frac{2\lambda\sqrt{6\varepsilon}}{1 - 6\varepsilon} \leq 5\sqrt{\varepsilon} + 10\lambda\sqrt{\varepsilon} \qquad (\varepsilon \leq 1/12)$$

and

$$\begin{aligned}
|\operatorname{tr}(\Pi\Sigma_{S'}) - \operatorname{tr}(\Pi)| &\le \left|\operatorname{tr}\big(\Pi(\Sigma_{S'} - \Sigma_{S'}(\mu_S))\big)\right| + \left|\operatorname{tr}\big(\Pi(\Sigma_{S'}(\mu_S) - \Sigma_S)\big)\right| + \left|\operatorname{tr}\big(\Pi(\Sigma_S - I_d)\big)\right| \\
&\le \|\mu_S - \mu_{S'}\|_2^2 + \left(\frac{100}{1-6\varepsilon}\operatorname{tr}(\Pi) + \frac{16\lambda^2}{1-6\varepsilon}\right) + \max\{\operatorname{tr}(\Pi\Sigma_S), \operatorname{tr}(\Pi)\} \\
&\le \left(5\sqrt{\varepsilon} + \frac{2\lambda\sqrt{6\varepsilon}}{1-6\varepsilon}\right)^2 + \left(\frac{100}{1-6\varepsilon}\operatorname{tr}(\Pi) + \frac{16\lambda^2}{1-6\varepsilon}\right) + 2\operatorname{tr}(\Pi) + 4\lambda^2 \\
&\le \left(5\sqrt{\varepsilon} + \frac{2\lambda\sqrt{6\varepsilon}}{1-6\varepsilon}\right)^2 + \left(\frac{100}{1-6\varepsilon}\operatorname{tr}(\Pi) + \frac{16\lambda^2}{1-6\varepsilon}\right) + 2\operatorname{tr}(\Pi) + 4\lambda^2 \\
&\le \frac{152}{1-6\varepsilon}\operatorname{tr}(\Pi) + \frac{24\lambda^2}{1-6\varepsilon} && (\varepsilon \le 1/12) \\
&\le 304\operatorname{tr}(\Pi) + 48\lambda^2.
\end{aligned}$$

Moreover, since $|T| \le |S|$ and $|T \cap S| \ge (1-6\varepsilon)|S|$, we have that $\|T - S\|_{\mathsf{TV}} \le 6\varepsilon$. Using this, that $\operatorname{tr}(\Sigma_T - \sigma^2 I_d)_+ = \eta$, and that $S \in \mathcal{S}(6\varepsilon, 5\sqrt{\varepsilon}, \lambda)$, we apply Lemma 27 to T and S with $q = k = 1$ and $\varepsilon \leftarrow 6\varepsilon \le E_1\varepsilon$ to obtain

$$\begin{aligned}
\|\mu_T - \mu_S\|_2 &\le \mathsf{W}_{1,1}(T, S) \\
&\le \frac{21\big(5\sqrt{\varepsilon} + \sqrt{6\varepsilon}\sigma\big)}{1-6\varepsilon} + \frac{36\sqrt{\varepsilon}(\lambda + \sqrt{\eta})}{(1-6\varepsilon)^{3/2}} \\
&\le 2^{12}\sqrt{\varepsilon} + 102\sqrt{\varepsilon}(\lambda + \sqrt{\eta}) && (\varepsilon \le 1/12, \sigma \le 50)
\end{aligned}$$

Thus, the triangle inequality gives

$$\begin{aligned}
\|\mu_T - \mu_{S'}\|_2 &\le \|\mu_T - \mu_S\|_2 + \|\mu_S - \mu_{S'}\|_2 \\
&\le \big(2^{12}\sqrt{\varepsilon} + 102\sqrt{\varepsilon}(\lambda + \sqrt{\eta})\big) + \big(5\sqrt{\varepsilon} + 10\lambda\sqrt{\varepsilon}\big) \\
&\le 2^{13}\sqrt{\varepsilon} + 112\lambda\sqrt{\varepsilon} + 102\sqrt{\varepsilon\eta}
\end{aligned}$$

Combining the above, we have for such S' that

$$\begin{aligned}
\sum_{x \in S'} g(x) &= |S'|\big(\operatorname{tr}(\Pi\Sigma_{S'}) + \|\Pi(\mu_T - \mu_{S'})\|_2^2\big) \\
&\le |S|\Big(\operatorname{tr}(\Pi) + 304\operatorname{tr}(\Pi) + 48\lambda^2 + \big[2^{13}\sqrt{\varepsilon} + 112\lambda\sqrt{\varepsilon} + 102\sqrt{\eta\varepsilon}\big]^2\Big) \\
&\le |S|\big(305\operatorname{tr}(\Pi) + 2^{28}\varepsilon + 2^{14}\lambda^2 + 2^{15}\eta\varepsilon\big)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{x \in S'} g(x) &= |S'|\big(\operatorname{tr}(\Pi\Sigma_{S'}) + \|\Pi(\mu_T - \mu_{S'})\|_2^2\big) \\
&\ge |S'|\big(\operatorname{tr}(\Pi) - 304\operatorname{tr}(\Pi) - 2^{28}\varepsilon - 2^{14}\lambda^2 - 2^{15}\eta\varepsilon\big) \\
&\ge (1-\varepsilon)|S|\big(-303\operatorname{tr}(\Pi) - 2^{28}\varepsilon - 2^{14}\lambda^2 - 2^{15}\eta\varepsilon\big) \\
&\ge |S|\big(-303\operatorname{tr}(\Pi) - 2^{28}\varepsilon - 2^{14}\lambda^2 - 2^{15}\eta\varepsilon\big)
\end{aligned}$$

Since $|T| \geq (1-4\varepsilon)|S| \geq 2|S|/3$, combining the above gives

$$\begin{aligned}
\sum_{x \in T \setminus S} g(x) &\geq \sum_{x \in T} g(x) - \sum_{x \in S} g(x) \\
&= |T|\big(\eta + \sigma^2 \operatorname{tr}(\Pi)\big) - |S|\big(305 \operatorname{tr}(\Pi) + 2^{28}\varepsilon + 2^{14}\lambda^2 + 2^{15}\eta\varepsilon\big) \\
&\geq |S|\big(2\eta/3 + 1300 \operatorname{tr}(\Pi) - 2^{28}\varepsilon - 2^{14}\lambda^2 - 2^{15}\eta\varepsilon\big) && (\sigma \geq 50) \\
&\geq |S|(\eta/4 + 1300 \operatorname{tr}(\Pi)) && (\varepsilon \leq 2^{-18},\, \eta \geq 2^{30}\varepsilon + 2^{14}\lambda^2)
\end{aligned}$$

Moreover, since $|L| = 6\varepsilon|T| \geq 6\varepsilon|S|(1-4\varepsilon) \geq 4\varepsilon|S| \geq |T \setminus S|$, and g takes its largest values on points of L, we have

$$\sum_{x \in T} f(x) = \sum_{x \in L} g(x) \geq \sum_{x \in T \setminus S} g(x) \geq |S|(\eta/4 + 1300 \operatorname{tr}(\Pi)).$$

Finally, plugging in $S' = S$ and $S' = S \setminus L$ into the bounds above on $\sum_{x \in S'} g(x)$, we obtain

$$\begin{aligned}
\sum_{x \in S \cap T} f(x) &= \sum_{x \in S \cap L} g(x) \\
&= \sum_{x \in S} g(x) - \sum_{x \in S \setminus L} g(x) \\
&\leq |S|\big(609 \operatorname{tr}(\Pi) + 2^{29}\varepsilon + 2^{15}\lambda^2 + 2^{16}\eta\varepsilon\big) \\
&\leq |S|(609 \operatorname{tr}(\Pi) + \eta/8) && (\varepsilon \leq 2^{-20},\, \eta \geq 2^{33}\varepsilon + 2^{19}\lambda^2) \\
&\leq \frac{1}{2} \sum_{x \in T} f(x),
\end{aligned}$$

as desired. ∎

Now, by the exact martingale argument used to prove Theorem 2.17 in Diakonikolas and Kane (2023), Theorem 26 implies that W2PROJECT maintains the invariant $|S \cap T| \geq (1-4\varepsilon)|S|$ over all iterations with probability at least 2/3. Since at least one point is removed from T at each iteration, the algorithm must terminate while satisfying this invariant as well as the (updated) termination condition at Step 4: $\operatorname{tr}(\Sigma_T - \sigma^2 I_d)_+ < C\varepsilon + C\rho^2\varepsilon^{1-2/p}$. Consequently, the returned measure $\hat{P} = \mathrm{Unif}(T)$ satisfies

$$\|\hat{P} - P'\|_{\mathsf{TV}} \leq 4\varepsilon$$

and

$$\operatorname{tr}(\Sigma_{\hat{P}} - \Sigma_{P'} - \sigma^2 I_d)_+ \leq \operatorname{tr}(\Sigma_{\hat{P}} - \sigma^2 I_d)_+ \leq C\varepsilon + C\rho^2\varepsilon^{1-2/p}.$$

Thus, by Lemma 28 and the fact that $P \in \mathcal{S}(\varepsilon, O(\sqrt{\varepsilon}))$, we have

$$\begin{aligned}
\mathsf{W}_{q,k}(Q, P) &\lesssim \varepsilon^{\frac{1}{q}-\frac{1}{2}}\sqrt{k} + \varepsilon^{-[\frac{1}{p}-\frac{1}{q}]_+}\left(\rho + \varepsilon^{\frac{1}{p}-\frac{1}{2}}\sqrt{C\varepsilon + 2C\rho^2\varepsilon^{1-\frac{2}{p}}}\right) \\
&\lesssim \varepsilon^{\frac{1}{q}-\frac{1}{2}}\sqrt{k} + \varepsilon^{-[\frac{1}{p}-\frac{1}{q}]_+}\left(\rho + \varepsilon^{\frac{1}{p}}\right) \\
&\lesssim \varepsilon^{\frac{1}{q}-\frac{1}{2}}\sqrt{k} + \varepsilon^{-[\frac{1}{p}-\frac{1}{q}]_+}\rho,
\end{aligned}$$

as desired. ∎

B.3. Error Certificate Lemmas

We state a useful technical lemma extending the certificate lemma (Lemma 2.6) of Diakonikolas and Kane (2023). Note that here the name is less precise; since P' is not observed, we cannot certify this condition from our observation unless we approximate $\Sigma_{P'}$ by I_d.

Lemma 27 *Let $\lambda_1, \lambda_2, \eta \geq 0$, $\varepsilon \in (0,1)$, and $\delta \geq \varepsilon$. Fix $P' \in \mathcal{S}(\varepsilon, \delta, \eta)$ and $Q \in \mathcal{P}(\mathbb{R}^d)$ such that $\operatorname{tr}(\Sigma_Q - \Sigma_{P'} - \lambda_1 I_d)_+ \leq \lambda_2$ and $\|Q - P'\|_{\mathsf{TV}} \leq \varepsilon$. Then*

$$\mathsf{W}_{q,k}(Q, P') \leq \frac{21\varepsilon^{\frac{1}{q}-1}\tilde{\delta}\sqrt{k}}{1-\varepsilon} + \frac{36\varepsilon^{\frac{1}{q}-\frac{1}{2}}\tilde{\eta}}{(1-\varepsilon)^{3/2}}$$

for all $k \in [d]$, where $\tilde{\delta} = \delta + \sqrt{\lambda_1 \varepsilon}$ and $\tilde{\eta} = \eta + \sqrt{\lambda_2}$.

We now apply this result under Setting B2.

Lemma 28 *Let $\lambda_1, \lambda_2 \geq 0$ and $C \geq 1$. Under Setting B2 with $P \in \mathcal{S}(C\varepsilon, \delta)$, fix any $Q \in \mathcal{P}(\mathbb{R}^d)$ such that $\operatorname{tr}(\Sigma_Q - \Sigma_{P'} - \lambda_1 I_d)_+ \leq \lambda_2$ and $\|Q - P'\|_{\mathsf{TV}} \leq \tau$, where $\varepsilon \leq \tau \leq C\varepsilon$. Then*

$$\mathsf{W}_{q,k}(Q, P) \leq \frac{21\tau^{\frac{1}{q}-1}\tilde{\delta}\sqrt{k}}{1-\tau} + \frac{37\tau^{-[1/p-1/q]_+}\tilde{\rho}}{(1-\tau)^{\frac{3}{2}}}$$

for all $k \in [d]$, where $\tilde{\delta} = \delta + \sqrt{\lambda_1 \tau}$ and $\tilde{\rho} = \rho + \tau^{1/p-1/2}\sqrt{\lambda_2}$.

Proof Under Setting B2, we have $P' \in \mathcal{S}(C\varepsilon, \delta, \eta) \subseteq \mathcal{S}(\tau, \delta, \eta)$, where $\eta = \rho\varepsilon^{1/2-1/p} \geq \rho\tau^{1/2-1/p}$. Applying Lemma 27 to P' and Q with TV corruption level τ and plugging in our value of η gives

$$\mathsf{W}_{q,k}(Q, P') \leq \frac{21\tau^{\frac{1}{q}-1}\tilde{\delta}\sqrt{d}}{1-\tau} + \frac{36\tau^{\frac{1}{q}-\frac{1}{2}}\eta}{(1-\tau)^{3/2}} \leq \frac{21\tau^{\frac{1}{q}-1}\tilde{\delta}\sqrt{d}}{1-\tau} + \frac{36\tau^{\frac{1}{q}-\frac{1}{p}}\tilde{\rho}}{(1-\tau)^{3/2}}.$$

Moreover, we have $\mathsf{W}_{q,k}(P, P') \leq \rho\varepsilon^{-[1/p-1/q]_+} \leq \tilde{\rho}\tau^{-[1/p-1/q]_+}$. Noting that $\tau^{-[1/p-1/q]_+} \geq \tau^{1/q-1/p}$, the lemma follows by the triangle inequality. ∎

We now return to the initial technical lemma.

Proof of Lemma 27 By the TV bound, we can decompose $Q = (1-\varepsilon)\bar{P} + \varepsilon R$ for some $\bar{P} \leq \frac{1}{1-\varepsilon}P'$. Using this decomposition, we bound

$$\begin{aligned}
\lambda_2 &\geq \operatorname{tr}(\Sigma_Q - \Sigma_{P'} - \lambda_1 I_d) \\
&= \operatorname{tr}(\Sigma_{\bar{P}} - \Sigma_{P'}) + \varepsilon \operatorname{tr}(\Sigma_R) + \varepsilon(1-\varepsilon)\|\mu_{\bar{P}} - \mu_R\|_2^2 - \varepsilon \operatorname{tr}(\Sigma_{\bar{P}}) - d\lambda_1 \\
&\geq \operatorname{tr}(\Sigma_{\bar{P}} - \Sigma_{P'}) + \varepsilon \operatorname{tr}(\Sigma_R) - \varepsilon \operatorname{tr}(\Sigma_{\bar{P}}) - d\lambda_1.
\end{aligned}$$

Since $P' \in \mathcal{S}(\varepsilon, \delta, \eta)$, we can rearrange the above and apply Lemma 18 to obtain

$$
\begin{aligned}
\operatorname{tr}(\Sigma_R) &\leq \frac{\lambda_2}{\varepsilon} + \frac{1}{\varepsilon}\operatorname{tr}(\Sigma_{P'} - \Sigma_{\bar{P}}) + \operatorname{tr}(\Sigma_{\bar{P}}) + \frac{d\lambda_1}{\varepsilon} \\
&\leq \frac{\lambda_2}{\varepsilon} + \frac{1}{\varepsilon}\operatorname{tr}(\Sigma_{P'} - \Sigma_{\bar{P}}(\mu_{P'})) + \frac{1}{\varepsilon}\|\mu_{P'} - \mu_{\bar{P}}\|_2^2 + \operatorname{tr}(\Sigma_{\bar{P}}) + \frac{d\lambda_1}{\varepsilon} \\
&\leq \frac{\lambda_2}{\varepsilon} + \frac{1}{\varepsilon}\tau_2(P', \varepsilon) + \frac{1}{\varepsilon}\tau(P, \varepsilon)^2 + \operatorname{tr}(\Sigma_{\bar{P}}) + \frac{d\lambda_1}{\varepsilon} \\
&\leq \frac{\lambda_2}{\varepsilon} + \frac{1}{\varepsilon}\left[\frac{4d\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\eta^2}{1-\varepsilon}\right] + \frac{1}{\varepsilon}\left[\delta + \frac{2\sqrt{\varepsilon}\eta}{1-\varepsilon}\right]^2 + \left[\frac{6d\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{20\eta^2}{1-\varepsilon}\right] + \frac{d\lambda_1}{\varepsilon} \\
&\leq \frac{d\lambda_1}{\varepsilon} + \frac{\lambda_2}{\varepsilon} + \frac{12d\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{44\eta^2}{(1-\varepsilon)^2\varepsilon} \\
&\leq \frac{13d\tilde{\delta}^2}{(1-\varepsilon)\varepsilon^2} + \frac{45\tilde{\eta}^2}{(1-\varepsilon)^2\varepsilon}.
\end{aligned}
$$

Next, take v to be the unit vector in the direction of $\mu_{P'} - \mu_R$. Then, applying a similar argument as above, we bound

$$
\begin{aligned}
\lambda_2 &\geq v^\top(\Sigma_Q - \Sigma_{P'} - \lambda_1 I_d)v \\
&= v^\top(\Sigma_{\bar{P}} - \Sigma_{P'})v + 2\varepsilon v^\top \Sigma_R v + \varepsilon(1-\varepsilon)\|\mu_{\bar{P}} - \mu_R\|_2^2 - \varepsilon v^\top \Sigma_{\bar{P}} v - \lambda_1\|v\|_2^2 \\
&\geq v^\top(\Sigma_{\bar{P}} - \Sigma_{P'})v + \varepsilon(1-\varepsilon)\|\mu_{\bar{P}} - \mu_R\|_2^2 - \varepsilon v^\top \Sigma_{\bar{P}} v - \lambda_1
\end{aligned}
$$

Rearranging and applying Lemma 18 to $v_\sharp^\top P'$, we bound $\varepsilon(1-\varepsilon)\|\mu_{\bar{P}} - \mu_R\|_2^2$ by

$$
\begin{aligned}
&\lambda_2 + v^\top(\Sigma_{P'} - \Sigma_{\bar{P}})v + \varepsilon v^\top \Sigma_{\bar{P}} v + \lambda_1 \\
&\leq \lambda_2 + v^\top(\Sigma_{P'} - \Sigma_{\bar{P}}(\mu_{P'}))v + \|\mu_{P'} - \mu_{\bar{P}}\|_2^2 + \varepsilon v^\top \Sigma_{\bar{P}} v + \lambda_1 \\
&\leq \lambda_2 + \left(\frac{4\delta^2}{(1-\varepsilon)\varepsilon} + \frac{16\eta^2}{1-\varepsilon}\right) + \left(\delta + \frac{2\sqrt{\varepsilon}\eta}{1-\varepsilon}\right)^2 + \varepsilon\left(\frac{6\delta^2}{(1-\varepsilon)\varepsilon^2} + \frac{20\eta^2}{1-\varepsilon}\right) + \lambda_1 \\
&\leq \lambda_1 + \lambda_2 + \frac{12\delta^2}{(1-\varepsilon)\varepsilon} + \frac{44\eta^2}{(1-\varepsilon)^2} \\
&\leq \frac{13\tilde{\delta}^2}{(1-\varepsilon)\varepsilon} + \frac{45\tilde{\eta}^2}{(1-\varepsilon)^2}
\end{aligned}
$$

We thus have

$$
\|\mu_{\bar{P}} - \mu_R\|_2^2 \leq \frac{13\tilde{\delta}^2}{(1-\varepsilon)^2\varepsilon^2} + \frac{45\tilde{\eta}^2}{(1-\varepsilon)^3\varepsilon}
$$

Combining with an application of Lemma 18 to P', we bound

$$
\begin{aligned}
\|\mu_R - \mu_{P'}\|_2^2 &\leq 2\|\mu_R - \mu_{\bar{P}}\|_2^2 + 2\|\mu_{\bar{P}} - \mu_{P'}\|_2^2 \\
&\leq \frac{26\tilde{\delta}^2}{(1-\varepsilon)^2\varepsilon^2} + \frac{90\tilde{\eta}^2}{(1-\varepsilon)^3\varepsilon} + 2\left(\delta + \frac{2\sqrt{\varepsilon}\eta}{1-\varepsilon}\right)^2 \\
&\leq \frac{30\tilde{\delta}^2}{(1-\varepsilon)^2\varepsilon^2} + \frac{98\tilde{\eta}^2}{(1-\varepsilon)^3\varepsilon}.
\end{aligned}
$$

Next, we apply Lemma 19 to bound $\mathsf{W}_q(Q, P')$ by

$$
\begin{aligned}
&\frac{7\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\eta}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\operatorname{tr}(\Sigma_R) + \|\mu_R - \mu_{P'}\|_2^2} \\
&\leq \frac{7\varepsilon^{\frac{1}{q}-1}\delta\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\eta}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\frac{13d\tilde{\delta}^2}{(1-\varepsilon)\varepsilon^2} + \frac{45\tilde{\eta}^2}{(1-\varepsilon)^2\varepsilon} + \frac{30\tilde{\delta}^2}{(1-\varepsilon)^2\varepsilon^2} + \frac{98\tilde{\eta}^2}{(1-\varepsilon)^3\varepsilon}} \\
&\leq \frac{7\varepsilon^{\frac{1}{q}-1}\tilde{\delta}\sqrt{d}}{\sqrt{1-\varepsilon}} + \frac{12\varepsilon^{\frac{1}{q}-\frac{1}{2}}\tilde{\eta}}{\sqrt{1-\varepsilon}} + 2\varepsilon^{\frac{1}{q}}\sqrt{\frac{43\tilde{\delta}^2}{(1-\varepsilon)^2\varepsilon^2} + \frac{143\tilde{\eta}^2}{(1-\varepsilon)^3\varepsilon}} \\
&\leq \frac{21\varepsilon^{\frac{1}{q}-1}\tilde{\delta}\sqrt{d}}{1-\varepsilon} + \frac{36\varepsilon^{\frac{1}{q}-\frac{1}{2}}\tilde{\eta}}{(1-\varepsilon)^{3/2}}
\end{aligned}
$$

as desired. As usual, for $k < d$, we note that the analysis above applies to all k-dimensional orthogonal projections of the input measures, with the substitution $d \leftarrow k$. ■

B.4. Proof of Corollary 9

Here, we simply apply Proposition 8 of Niles-Weed and Rigollet (2022) for distributions satisfying a $T_1(\sigma)$ equality and note that a $O(1)$-sub-Gaussian distribution in $\mathbb{R}^d$ satisfies a $T_1(O(d))$ inequality (as noted therein). Since this empirical convergence guarantee is in expectation, and we seek a uniform bound over $k \in [d]$, we must multiply the resulting bound by d to apply Markov's inequality and deduce the desired error bound with high constant probability. ■

B.5. Proof of Proposition 11

We have in this case that $\Sigma_{P'} = \Sigma_P$ (since $\rho = 0$). Then, if $\|\Sigma_{\hat{P}}\|_{\mathrm{op}} \leq 1 + C\delta^2/\varepsilon$, we have by stability of P that

$$
\operatorname{tr}(\Sigma_{\hat{P}} - \Sigma_P - (1 + (C-1)\tfrac{\delta^2}{\varepsilon})I_d)_+ \leq \operatorname{tr}(\Sigma_{\hat{P}} - (1 + C\tfrac{\delta^2}{\varepsilon})I_d)_+ \leq 0,
$$

at which point Lemma 28 with $p = q = 1$ gives the Proposition.