Robust Estimation for Non-parametric Families via Generative Adversarial Networks

Banghua Zhu
Department of EECS
University of California, Berkeley
California, USA
Email: banghua@berkeley.edu

Jiantao Jiao and Michael I. Jordan
Departments of EECS and Statistics
University of California, Berkeley
California, USA
Email: {jiantao, michael_jordan}@berkeley.edu

Abstract—We provide a general framework for designing Generative Adversarial Networks (GANs) to solve high-dimensional robust statistics problems, which aim at estimating unknown parameter of the true distribution given adversarially corrupted samples. Prior work [1], [2] focus on the problem of robust mean and covariance estimation when the true distribution lies in the family of Gaussian distributions or elliptical distributions, and analyze depth or scoring rule based GAN losses for the problem. Our work extend these to robust mean estimation, second-moment estimation, and robust linear regression when the true distribution only has bounded Orlicz norms, which includes the broad family of sub-Gaussian, sub-exponential and bounded moment distributions. We also provide a different set of sufficient conditions for the GAN loss to work: we only require its induced distance function to be a cumulative density function of some light-tailed distribution, which is easily satisfied by neural networks with sigmoid activation. In terms of techniques, our proposed GAN losses can be viewed as a smoothed and generalized Kolmogorov-Smirnov distance, which overcomes the computational intractability of the original Kolmogorov-Smirnov distance used in the weaken the distance approach in [3].

I. INTRODUCTION

High-dimensional robust statistics focuses on estimating unknown parameters of high-dimensional distributions given corrupted samples. Formally, assume that the true distribution $p^\star$ with dimension d is altered to some corrupted distribution p such that $\mathsf{TV}(p^\star, p) \leq \epsilon$, where TV is the total variation distance. One takes n samples from p to form the empirical distribution $\hat{p}_n$. Given the empirical distribution and a loss function L, we aim to output a parameter $\hat{\theta}(\hat{p}_n)$ such that $L(p^\star, \hat{\theta}(\hat{p}_n))$ is small. As a concrete example in the setting of robust mean estimation, one might take the loss function to be $L(p^\star, \hat{\theta}(\hat{p}_n)) = \|\mathbb{E}_{p^\star}[X] - \hat{\theta}(\hat{p}_n)\|_2$, which is the ℓ_2 distance between the true mean and an algorithm's output.

To achieve robustness, it is necessary to impose nontrivial assumptions on the true distribution $p^\star$. Classical asymptotic statistical theory focused on minimum-distance functionals for Gaussian or elliptical distributions [2], [4]–[10]. Recent work has instead focused on computationally efficient algorithms based on weaker assumptions regarding the tail behavior of distributions [3], [11]–[20]. Another recent line of work focuses on generative adversarial networks (GANs) as a promising framework for developing robust estimators. Theoretical work has shown that when the true distribution lies in the family of Gaussian or elliptical distributions, one can design specific GAN losses to achieve theoretical guarantee on robust mean estimation and robust covariance estimation [1], [2].

In this work, we take the theoretical study of GANs for robust estimation further by connecting it to the minimum-distance functionals, moving beyond strong Gaussian assumptions and providing concrete sufficient conditions for robustness. Specifically, our main contributions are as follows:

- We extend the design of GANs for robust estimation from Gaussian and elliptical distribution in [1], [2] to a rich family of nonparametric classes of distributions, including the family of sub-Gaussian, sub-exponential and bounded moment distributions.
- We identify sufficient conditions for robust estimation which pinpoint a broad family of GANs that provably succeed in robust mean estimation, robust covariance estimation and robust linear regression.
- We provide computable methods for the "weaken the distance" approach in [3] by smoothing the Kolmogorov-Smirnov distance, which appears to be difficult to optimize.

II. PRELIMINARIES

We begin by introducing basic notation. A function $\psi : [0, +\infty) \mapsto [0, +\infty)$ is called an *Orlicz function* if ψ is convex, non-decreasing, and satisfies $\psi(0) = 0$, $\psi(x) \to \infty$ as $x \to \infty$. For a given Orlicz function ψ, the Orlicz norm of a random variable X is defined as $\|X\|_\psi \triangleq \inf\{t > 0 : \mathbb{E}_p[\psi(|X|/t)] \leq 1\}$. We use $\|A\|_2 = \sup_{v \neq 0} \|Av\|_2 / \|v\|_2$ to denote the spectral norm of the matrix A. For univariate random variables X and Y, we say that Y *stochastically dominates* X (to first order) if $\mathbb{P}(X \leq t) \geq \mathbb{P}(Y \leq t)$ for all $t \in \mathbb{R}$ [21]. We define the generalized inverse of a non-decreasing function ψ as $\psi^{-1}(y) = \inf\{x \mid \psi(x) > y\}$. We say a distribution r is an ϵ-deletion of another distribution p if $r \leq \frac{p}{1-\epsilon}$. The inequality can be formally understood as $\frac{dr}{dp} \leq \frac{1}{1-\epsilon}$, where $\frac{dr}{dp}$ is the Radon-Nikodym derivative, which can also be understood as $r(A) \leq \frac{p(A)}{1-\epsilon}$ for any set A; an equivalent characterization is that r can be obtained from p by conditioning on an event E of probability at least $1 - \epsilon$.

A. Tasks considered in Robust Estimation

We focus on three tasks in robust estimation—robust mean estimation, robust second-moment estimation, and robust linear regression—throughout this paper. For robust mean estimation, we take the loss function to be $L(p^\star, \hat{\theta}(\hat{p}_n)) = \|\mathbb{E}_{p^\star}[X] - \hat{\theta}(\hat{p}_n)\|_2$; for second-moment estimation, we use the loss function $L(p^\star, \hat{\theta}(\hat{p}_n)) = \|\mathbb{E}_{p^\star}[XX^\top] - \hat{\theta}(\hat{p}_n)\|_2$, where the algorithm output $\hat{\theta}(\hat{p}_n)$ is a matrix; and for linear regression, we take the loss function to be the excess predictive loss, $L(p^\star, \hat{\theta}(\hat{p}_n)) = \mathbb{E}_{p^\star}[(Y - X^\top\hat{\theta}(\hat{p}_n))^2 - (Y - X^\top\theta^\star(p^\star))^2]$, where $\theta^\star(p^\star) = \arg\min_\theta \mathbb{E}_{p^\star}[(Y - X^\top\theta)^2]$.

When the output $\hat{\theta}$ is a parameter of a distribution, the losses can alternatively be written as distances between distributions. For mean estimation, the loss can be written as $L(p^\star, q) = \|\mathbb{E}_{p^\star}[X] - \mathbb{E}_q[X]\|_2$; for second-moment estimation, the loss can be written as $L(p^\star, q) = \|\mathbb{E}_{p^\star}[XX^\top] - \mathbb{E}_q[XX^\top]\|_2$; and for linear regression, the loss function can be written as $L(p^\star, q) = \mathbb{E}_{p^\star}[(Y - X^\top\theta^\star(q))^2 - (Y - X^\top\theta^\star(p^\star))^2]$.

Note that the losses for mean and second-moment estimation admit a pseudonorm representation, $W_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}}(\mathbb{E}_p[f(X)] - \mathbb{E}_q[f(X)])$. In the case of mean estimation, we have $\mathcal{F}_{\text{mean}} = \{f(x) = v^\top x \mid v \in \mathbb{R}^d, \|v\|_2 \le 1\}$, and for second-moment estimation, we have $\mathcal{F}_{\text{sec}} = \{f(x) = \xi(v^\top x)^2 \mid v \in \mathbb{R}^d, \|v\|_2 \le 1, \xi \in \{\pm 1\}\}$. When we state our main theorem we will assume that the loss admits this pseudonorm representation, and showcase how to deal with linear regression as an exception.

B. Generalized Resilience and Minimum Distance Functional

It is shown in [3, Section 3.2] that when the true distribution lies in a "generalized resilience" family, it is possible to design algorithms with robustness guarantees. In particular, when the loss function takes the form of a pseudonorm $W_{\mathcal{F}}(p, q)$, the generalized resilience set can be expressed simply as:

$$\mathcal{G}_{\mathcal{F}}(\rho) = \{p \mid \mathbb{E}_r[f(X)] - \mathbb{E}_p[f(X)] \le \rho(\epsilon), \forall \epsilon \in [0, 1),$$
$$\forall r \le \frac{p}{1 - \epsilon}, f \in \mathcal{F}\}. \quad (1)$$

For convenience of notation, we use $\mathcal{G}_{\text{mean}}$ to denote $\mathcal{G}_{\mathcal{F}_{\text{mean}}}$, and $\mathcal{G}_{\text{sec}}$ to denote $\mathcal{G}_{\mathcal{F}_{\text{sec}}}$. It is shown in [3] that $\mathcal{G}_{\text{mean}}(\epsilon\psi^{-1}(\epsilon))$ is a superset of the set of bounded Orlicz norm distributions, $\{X \sim p \mid \sup_{\|v\|_2 \le 1} \|v^\top(X - \mathbb{E}[X])\|_\psi \le 1\}$, which implies a bounded k-th moment when $\psi(x) = x^k$, sub-exponentiality when $\psi(x) = \exp(x)$, and sub-Gaussianity when $\psi = \exp(x^2)$. Similarly, $\mathcal{G}_{\text{sec}}(\epsilon\psi^{-1}(\epsilon))$ is a superset of $\{X \sim p \mid \sup_{\|v\|_2 \le 1} \|(v^\top X)^2\|_\psi \le 1\}$.

The generalized resilience set takes a slightly more complicated form for linear regression. However, it is also a superset of some bounded Orlicz norm family. We use $\mathcal{G}_{\text{reg}}(\psi)$ to denote the following family:

$$\mathcal{G}_{\text{reg}}(\psi) = \left\{ p \mid \mathbb{E}_p\left[\psi\left(\frac{(v^\top X)^2}{\sigma_1^2 \mathbb{E}_p[(v^\top X)^2]}\right)\right] \le 1, \forall v \in \mathbb{R}^d, \right.$$
$$\left. \text{and } \mathbb{E}_p\left[\psi\left(\frac{(Y - X^\top\theta^*(p))^2}{\sigma_2^2}\right)\right] \le 1 \right\}. \quad (2)$$

Working under the restrictive assumption that the true distribution lies in $\mathcal{G}$, [3] study a projection algorithm and show that it achieves polynomial sample complexity for robust estimation. The algorithm projects the corrupted empirical distribution $\hat{p}_n$ onto $\mathcal{G}$ under the *generalized Kolmogorov-Smirnov distance* $\mathsf{TV}_{\mathcal{H}}$, defined as

$$\mathsf{TV}_{\mathcal{H}}(p, q) \triangleq \sup_{f \in \mathcal{H}, t \in \mathbb{R}} |\mathbb{P}_p(f(X) \ge t) - \mathbb{P}_q(f(X) \ge t)|. \quad (3)$$

The distance $\mathsf{TV}_{\mathcal{H}}$ is smaller than the total variation distance TV because it takes a supremum over only the events defined by threshold functions in $\mathcal{H}$, while TV takes the same supremum over all measurable events. The projection algorithm is analyzed for Guassian mean estimation in [22], with $\mathcal{H} = \{v^\top x \mid v \in \mathbb{R}^d\}$, and extended to bounded Orlicz norm family for robust second-moment estimation, robust linear regression in [3]. Unfortunately, however, it is difficult to compute the $\mathsf{TV}_{\mathcal{H}}$ projection in practice due to the lack of differentiability of commonly used losses. We show how to circumvent this issue in the following section.

III. A UNIFIED FRAMEWORK FOR DESIGN OF GANs

In this section we show that a large family of smoothed loss function can be used as projection functionals and achieve favorable finite sample bounds. We define the *smoothed generalized Kolmogorov-Smirnov distance* as

$$\widetilde{\mathsf{TV}}_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}, t \in \mathbb{R}} |\mathbb{E}_p[T(f(X) + t)] - \mathbb{E}_q[T(f(X) + t)]|, \quad (4)$$

where we assume that $T(x)$, after an affine transformation $aT(x) + b$, can be written as the cumulative distribution function (CDF) of some random variable Z; i.e., $aT(-\infty) + b = 0, aT(+\infty) + b = 1$ and T is right-continuous (and thus $aT(x + b) = \mathbb{P}(Z \le x)$ for some random variable Z). An example is the sigmoid function $T(x) = 1/(1 + e^{-x})$. When $Z = 0$ almost surely, the smoothed distance $\widetilde{\mathsf{TV}}_{\mathcal{F}}$ reduces to the case of generalized KS distance. The key observation is there is a generalization of the *mean cross* lemma in [3, Lemma 3.3] that works for the *smoothed generalized Kolmogorov-Smirnov distance*.

Lemma 1 (Closeness in smoothed KS distance implies mean cross). *Assume for two distribution p, q,*

$$\sup_{t \in \mathbb{R}} |\mathbb{E}_p[T(X + t)] - \mathbb{E}_q[T(X + t)]| \le \epsilon, \quad (5)$$

where we assume that there exist $a, b \in \mathbb{R}$ such that $aT(x) + b$ can be written as the CDF of some random variable Z. Assume the distribution of Z is inside the resilience family of $\mathcal{G}_{\text{mean}}(\rho_Z)$. Then there exist $r_p \le \frac{p}{1 - |a|\epsilon}$ and $r_q \le \frac{q}{1 - |a|\epsilon}$ such that

$$\mathbb{E}_{r_q}[X] - \mathbb{E}_{r_p}[X] \le 2\rho_Z(|a|\epsilon). \quad (6)$$

For the generalized KS distance $\mathsf{TV}_{\mathcal{F}}$, $Z \equiv 0$ almost surely, and $\rho_Z = 0$, which corresponds to Lemma [3, Lemma 3.3].

Proof of Lemma 1. For a fixed x, $aT(x+t) + b = \mathbb{P}(Z \le x+t) = \mathbb{P}(x - Z \ge -t)$. Thus we know

$$\sup_{t \in \mathbb{R}} |\mathbb{P}_{q \times \mathbb{P}_Z}(X - Z \ge -t) - \mathbb{P}_{p \times \mathbb{P}_Z}(X - Z \ge -t)| \le |a|\epsilon, \tag{7}$$

where X and Z are independent. Denote $\tilde{p}$ as the distribution of $X - Z$ when $X \sim p$, and $\tilde{q}$ as the distribution of $X - Z$ when $X \sim q$. Starting from $\tilde{p}, \tilde{q}$, we delete probability mass of size ϵ corresponding to the largest points in $\tilde{p}$ to get $\tilde{r}_p$, and delete probability mass of size $|a|\epsilon$ corresponding to the smallest points $\tilde{q}$ to get $\tilde{r}_q$. Equation (7) implies that $\mathbb{P}_{Y \sim \tilde{r}_p}(Y \ge t) \le \mathbb{P}_{Y \sim \tilde{r}_q}(Y \ge t)$ holds for all $t \in \mathbb{R}$. Hence, $\tilde{r}_q$ stochastically dominates r_p and $\mathbb{E}_{\tilde{r}_p}[Y] \le \mathbb{E}_{\tilde{r}_q}[Y]$. Thus we know that there exist $\tilde{r}_p \le \frac{\tilde{p}}{1 - |a|\epsilon}$ and $\tilde{r}_q \le \frac{\tilde{q}}{1 - |a|\epsilon}$ such that

$$\mathbb{E}_{\tilde{r}_q}[X - Z] \le \mathbb{E}_{\tilde{r}_p}[X - Z].$$

Denote the original distribution of Z as p_Z. Note that the deletion process that yields $\tilde{r}_p$ and $\tilde{r}_q$ can be viewed as operating on the joint distribution of X, Z, such that the marginal distributions of $\tilde{r}_p$ and $\tilde{r}_q$ are obtained as deletions of p_Z. Rearranging, we obtain

$$\begin{aligned}\mathbb{E}_{\tilde{r}_q}[X] - \mathbb{E}_{\tilde{r}_p}[X] &\le \mathbb{E}_{\tilde{r}_q}[Z] - \mathbb{E}_{\tilde{r}_p}[Z] \\ &\le |\mathbb{E}_{\tilde{r}_q}[Z] - \mathbb{E}_{p_Z}[Z]| + |\mathbb{E}_{\tilde{r}_p}[Z] - \mathbb{E}_{p_Z}[Z]|. \\ &\le 2\rho_Z(|a|\epsilon).\end{aligned}$$

$\square$

We now show that given the mean cross lemma for smoothed KS distance, the smoothed distance can be used to deliver a small statistical error for robust estimation under a pseudonorm-based loss function $W_{\mathcal{F}}$. We will show later how the analysis extends beyond pseudonorm loss functions by considering the special case of linear regression.

Theorem 1. *For any $p^* \in \mathcal{G}_{\mathcal{F}}(\rho)$ in Equation (1), denote by p a corrupted distribution such that $\mathsf{TV}(p^*, p) \le \epsilon$. Let the observed empirical distribution of p be $\hat{p}_n$. Define*

$$A(p, q) = \sup_{(d_1, d_2) \in \mathcal{D}} \mathbb{E}_p[d_1(X)] + \mathbb{E}_q[d_2(X)]. \tag{8}$$

Here $\mathcal{D}$ is some family of discriminator function pairs (d_1, d_2). For some function $T(x)$, let $aT(x) + b$ be the CDF of some random variable Z. Assume the following conditions:

1) For all $(d_1, d_2) \in \mathcal{D}$, $x \in \mathbb{R}$, we have $|d_2(x)| \le 1/2$.

2) For any distribution pair $p, q \in \mathcal{G}_{\mathcal{F}}(\rho)$, we have

$$\begin{aligned}&A(q, p) - A(p, p) \le \epsilon \Rightarrow \\ &\sup_{f \in \mathcal{F}, t \in \mathbb{R}} |\mathbb{E}_q[T(f(X) + t)] - \mathbb{E}_p[T(f(X) + t)]| \le C\epsilon\end{aligned} \tag{9}$$

for some constant $C > 0$ and any $\epsilon < 1/C|a|$.

3) The distribution of Z is inside the resilient set $\mathcal{G}_{\mathsf{mean}}(\rho_Z)$,

Then the projection algorithm $q = \arg\min_{q \in \mathcal{G}_{\mathcal{F}}(\rho)} A(q, \hat{p}_n)$ guarantees

$$W_{\mathcal{F}}(p^*, q) \le 2\rho(C\tilde{\epsilon}) + 2\rho_Z(C\tilde{\epsilon}), \tag{10}$$

where $\tilde{\epsilon} = 2\epsilon + 2\bar{A}(p, \hat{p}_n)$, and

$$\bar{A}(p, q) = \sup_{(d_1, d_2) \in \mathcal{D}} |\mathbb{E}_p[d_2(X)] - \mathbb{E}_q[d_2(X)]|. \tag{11}$$

Remark 1. *The three conditions are weak enough to be satisfied for a large family of neural networks. The first condition on the magnitude of d_2 can be easily satisfied by setting the output of the neural network designed for d_2 to pass through a bounded activation function. The second condition is satisfied when the set $\mathcal{D}_1 = \{T(f(\cdot) + t) | f \in \mathcal{F}, t \in \mathbb{R}\}$. The third condition requires a careful design of T, and thus d_1, which can also be easily satisfied when the network has a bounded activation function. For example, a simple sigmoid or tanh function gives a CDF whose induced distribution is sub-exponential, and a ramp function gives a CDF whose induced distribution is sub-Gaussian. In [2], it was shown that under mild conditions, one can produce valid (d_1, d_2) using proper scoring rules and appropriate neural network architectures to ensure the three conditions. Our result, combined with [2], extends the results for mean and second-moment estimation in [2] to general resilient sets including sub-Gaussian and bounded k-th moments, while [2] gives guarantee for elliptical distributions as semi-parametric classes.*

Proof of Theorem 1. The proof mainly focuses on verifying two important properties in [3].

1) **Robustness to perturbation:** For any distribution p_1, p_2, p_3, we have

$$\begin{aligned}&|A(p_1, p_2) - A(p_1, p_3)| \\ =\ &|\sup_{(d_1, d_2) \in \mathcal{D}} (\mathbb{E}_{p_1}[d_1(X)] + \mathbb{E}_{p_2}[d_2(X)]) \\ &- \sup_{(d_1, d_2) \in \mathcal{D}} (\mathbb{E}_{p_1}[d_1(X)] + \mathbb{E}_{p_3}[d_2(X)])| \\ \le\ &\sup_{(d_1, d_2) \in \mathcal{D}} |\mathbb{E}_{p_1}[d_1(X)] + \mathbb{E}_{p_2}[d_2(X)] \\ &- (\mathbb{E}_{p_1}[d_1(X)] + \mathbb{E}_{p_3}[d_2(X)])| \\ \le\ &\sup_{(d_1, d_2) \in \mathcal{D}} |\mathbb{E}_{p_2}[d_2(X)] - \mathbb{E}_{p_3}[d_2(X)]| \\ =\ &\bar{A}(p_2, p_3) \\ \le\ &\sup_{\sup_x |d_2(x)| \le \frac{1}{2}} |\mathbb{E}_{p_2}[d_2(X)] - \mathbb{E}_{p_3}[d_2(X)]| \\ =\ &\mathsf{TV}(p_2, p_3). \end{aligned} \tag{12}$$

Combining Equation (12), the fact $q = \arg\min_{q \in \mathcal{G}_{\mathcal{F}}} A(q, \hat{p}_n)$ and $p^* \in \mathcal{G}_{\mathcal{F}}$, one has

$$\begin{aligned}&A(q, p^*) - A(p^*, p^*) \\ \le\ &A(q, p) + \epsilon - A(p^*, p^*) \\ \le\ &A(q, \hat{p}_n) + \epsilon + \bar{A}(p, \hat{p}_n) - A(p^*, p^*) \\ \le\ &A(p^*, p_n) + \epsilon + \bar{A}(p, \hat{p}_n) - A(p^*, p^*) \\ \le\ &A(p^*, p) + \epsilon + 2\bar{A}(p, \hat{p}_n) - A(p^*, p^*) \\ \le\ &2\epsilon + 2\bar{A}(p, \hat{p}_n) = \tilde{\epsilon}.\end{aligned}$$

2) **Generalized modulus of continuity:** For any $p, q \in \mathcal{G}_{\mathcal{F}}$, from $A(q, p) - A(p, p) \le \tilde{\epsilon}$ and condition 2) in Equa-

tion (9), we have

$$\sup_{f\in\mathcal{F},t\in\mathbb{R}} |\mathbb{E}_q[T(f(X)+t)] - \mathbb{E}_{p^\star}[T(f(X)+t)]| \le C\tilde{\epsilon}.$$

From Lemma 1, we know that for any fixed $f^* \in \arg\max_{f\in\mathcal{F}} \mathbb{E}_{p^\star}[f(X)] - \mathbb{E}_q[f(X)]$, there exist $r_{p^\star} \le \frac{p^\star}{1-C|a|\tilde{\epsilon}}, r_q \le \frac{q}{1-C|a|\tilde{\epsilon}}$, such that

$$\mathbb{E}_{\tilde{r}_{p^\star}} f^*(X) \le \mathbb{E}_{\tilde{r}_q} f^*(X) + 2\rho_Z(C|a|\tilde{\epsilon}).$$

From $q \in \mathcal{G}_\mathcal{F}(\rho)$, we have

$$\forall r \le \frac{q}{1-C|a|\tilde{\epsilon}}, W_\mathcal{F}(r,q) \le \rho(C|a|\tilde{\epsilon}). \qquad (13)$$

Therefore

$$\begin{aligned}
W_\mathcal{F}(p^\star, q) &= \mathbb{E}_{p^\star}[f^*(X)] - \mathbb{E}_q[f^*(X)] \\
&\le \mathbb{E}_{p^\star}[f^*(X)] - \mathbb{E}_{\tilde{r}_q}[f^*(X)] + \rho(C|a|\tilde{\epsilon}) \\
&\le \mathbb{E}_{p^\star}[f^*(X)] - \mathbb{E}_{\tilde{r}_{p^\star}}[f^*(X)] + 2\rho_Z(C|a|\tilde{\epsilon}) \\
&\quad + \rho(C|a|\tilde{\epsilon}) \\
&\le 2\rho(C|a|\tilde{\epsilon}) + 2\rho_Z(C|a|\tilde{\epsilon}), \qquad (14)
\end{aligned}$$

which finishes the proof.

$\square$

IV. APPLICATION OF THEOREM 1 FOR ROBUST ESTIMATION

Now we demonstrate how Theorem 1 leads to concrete designs of GANs for robust mean estimation, second-moment estimation and robust linear regression.

A. Mean estimation

We start with the problem of mean estimation. Recall the choice of function set $\mathcal{F}_{\mathsf{mean}} = \{f(x) = v^\top X | v \in \mathbb{R}^d, \|v\|_2 \le 1\}$. We show that different choices of $A(p,q)$ give similar performance guarantees in terms of robustness. For the convenience of notation, let

$$g_1(v,t,X) = \mathsf{sigmoid}(v^\top X + t),$$

$$g_2(w, \{v_j\}, \{t_j\}, X) = \mathsf{sigmoid}\left(\sum_{j\le l} w_j g_1(v_j, t_j, X)\right)$$

which represent one- and two-layer neural networks, respectively, with sigmoid activation functions. Here v_j, v are d-dimensional vectors, $t, t_j \in \mathbb{R}$, and w is a l-dimensional vector. We consider the following design of distances $A(p,q)$:

$$A_1(p,q) = \sup_{\|v\|_2\le 1, t\in\mathbb{R}} |\mathbb{E}_p[g_1(v,t,X)] - \mathbb{E}_q[g_1(v,t,X)]|,$$

$$(15)$$

$$\begin{aligned}
A_2(p,q) = \sup_{\|w\|_1\le 1, \|v_j\|_2\le 1, t_j\in\mathbb{R}} &|\mathbb{E}_p[g_2(w, \{v_j\}, \{t_j\}, X)] \\
&- \mathbb{E}_q[g_2(w, \{v_j\}, \{t_j\}, X)]|, \qquad (16)
\end{aligned}$$

$$\begin{aligned}
A_3(p,q) = \sup_{\substack{\|w\|_1\le 1, \|v_j\|_2\le 1, \\ t_j\in\mathbb{R}}} &\mathbb{E}_p[\log(g_2(w, \{v_j\}, \{t_j\}, X))] \\
&+ \mathbb{E}_q[\log(1 - g_2(w, \{v_j\}, \{t_j\}, X))], \qquad (17)
\end{aligned}$$

where A_1 can be viewed as a simple one-layer discriminator, A_2 is a two-layer discriminator, and A_3 is a two-layer discriminator with loss chosen as log score, which is analyzed for Gaussian and elliptical distributions in [2].

Corollary 1. *Assume the true distribution $p^\star \in \mathcal{G}_{\mathsf{mean}}(\rho)$. Let the projection algorithm $q_1 = \arg\min_{q\in\mathcal{G}_{\mathsf{mean}}} A_1(q, \hat{p}_n)$, and q_2, q_3 be the projection algorithm that projects under distance A_2, A_3. For any $i \in \{1,2,3\}$, with probability at least $1-\delta$,*

$$\|\mathbb{E}_{p^\star}[X] - \mathbb{E}_{q_i}[X]\|_2 \le C_1 \cdot (\rho(\tilde{\epsilon}) + \tilde{\epsilon}\log(1/\tilde{\epsilon})),$$

where $\tilde{\epsilon} = C_2(\epsilon + \sqrt{d/n} + \sqrt{\log(1/\delta)/n})$.

Remark 2. *Corollary 1 recovers [3, Theorem 3.2] up to an additive term $\tilde{\epsilon}\log(1/\tilde{\epsilon})$. As a direct result of Corollary 1, one can see that if $p^\star$ is inside the sub-Gaussian family, the projection algorithm guarantees a rate of $C\cdot\tilde{\epsilon}\log(1/\tilde{\epsilon})$, which is nearly optimal up to logarithmic factors compared to the result for Gaussian [2]. When $p^\star$ has bounded covariance, the rate becomes $(\epsilon + \sqrt{d/n} + \sqrt{\log(1/\delta)/n})^{1/2}$, which is optimal in terms of the dependence on ϵ, but sub-optimal in dependence with respect to d, n [2], [3], [23]–[27].*

Compared to the interpretation of proper scoring rules and appropriate network structures in [2], our analysis provides a different view in which robustness is achieved by controlling the tail of the induced distribution of $T(x)$. The results in Corollary 1 also generalize to deeper neural networks where multiple layers with sigmoid activation enables better bound due to the concentration of Z considered in Theorem 1.

Although projection to $\mathcal{G}$ is not exactly computable for general $\mathcal{G}$, in practice one may use a subset of Orlicz-norm bounded distributions which can be parameterized using neural networks, or turn the constrained optimization problem into a regularized optimization problem by adding the Orlicz norm as a regularizer in the loss function [28].

Proof of Corollary 1. Since scaling the distance does not change the final projection q_i, it suffices to verify the three assumptions in Theorem 1 for scaled distances $A_1/2, A_2/2, A_3/2$. For the first assumption, one can see that the three distances all satisfy $|d_2(x)| \le 1/2$ due to the fact that $|g_1| \le 1, g_2 \in [1/2, e/(e+1)], \log(1-g_2) \in [-0.7, -0.3]$ when $\|w\|_1 \le 1, \|v_j\|_2 \le 1$.

Now we verify the second assumption. Taking $w_1 = 1, w_j = 0, \forall j > 1$ in g_2 and applying triangle inequality, we have

$$\begin{aligned}
&A_1(q,p) - A_1(p,p) \le \epsilon \Rightarrow \\
&\sup_{\|v\|_2\le 1, t\in\mathbb{R}} |\mathbb{E}_p[g_1(v,t,X)] - \mathbb{E}_q[g_1(v,t,X)]| \le \epsilon, \\
&A_2(q,p) - A_2(p,p) \le \epsilon \Rightarrow \\
&\sup_{\|v\|_2\le 1, t\in\mathbb{R}} |\mathbb{E}_p[\mathsf{sigmoid}(g_1(v,t,X))] - \\
&\quad \mathbb{E}_q[\mathsf{sigmoid}(g_1(v,t,X))]| \le \epsilon, \\
&A_3(q,p) - A_3(p,p) \le \epsilon \Rightarrow \\
&\sup_{\|v\|_2\le 1, t\in\mathbb{R}} |\mathbb{E}_p[\log(\mathsf{sigmoid}(g_1(v,t,X)))] - \\
&\quad \mathbb{E}_q[\log(\mathsf{sigmoid}(g_1(v,t,X)))]| \le \epsilon.
\end{aligned}$$

These imply that A_1, A_2, A_3 satisfy assumption 2) in Theorem 1 with particular choices of T: $T_1(x) = \text{sigmoid}(x)$, $T_2 = \text{sigmoid}(T_1(x))$, $T_3 = \log(T_2(x))$.

We now verify the third assumption. For CDF $T_1(x) = \frac{1}{1+\exp(-x)}$, the corresponding random variable Z is subexponential since $\max(\mathbb{P}(X \leq -t), \mathbb{P}(X \geq t)) \leq \exp(-t)/2$ for any $t \geq 0$. For T_2, T_3, we can verify that after scaling the CDF, $\mathbb{E}[\exp(|Z|/10)] \leq 1$. This shows the induced random variables from T_2, T_3 are also sub-exponential. From [3] we know the distribution of Z lies in the set of $\mathcal{G}_{\text{mean}}(\epsilon \log(1/\epsilon))$.

Furthermore, from [2, Lemma 7.3] and [1, Lemma 8.1], we know that $\bar{A}_1(\hat{p}_n, p), \bar{A}_2(\hat{p}_n, p), \bar{A}_3(\hat{p}_n, p)$ are all bounded by $C(\sqrt{d/n} + \sqrt{\log(1/\delta)/n})$ with probability at least $1 - \delta$. $\quad\square$

B. Second-moment estimation

Using the same technique as mean estimation, we can show similar results for second-moment estimation by changing $\mathcal{F}_{\text{mean}}$ to $\mathcal{F}_{\text{sec}}$ in $\widetilde{\text{TV}}_{\mathcal{F}}$. Let A_1, A_2, A_3 be the same as mean estimation in Equation (15), (16), (17) except that we set

$$g_1(v, t, X) = \text{sigmoid}((v^\top X)^2 + t),$$

$$g_2(w, \{v_j\}, \{t_j\}, X) = \text{sigmoid}\left(\sum_{j \leq l} w_j g_1(v_j, t_j, X)\right).$$

$$= \text{sigmoid}\left(\sum_{j \leq l} w_j \text{sigmoid}((v_j^\top X)^2 + t)\right).$$

Following the same proof as Corollary 1, we have[1]

Corollary 2. *Assume the true distribution $p^\star \in \mathcal{G}_{\text{sec}}(\rho)$. Let the projection algorithm $q_1 = \arg\min_{q \in \mathcal{G}_{\text{sec}}} A_1(q, \hat{p}_n)$, and q_2, q_3 be the projection algorithm that projects under distance A_2, A_3. For any $i \in \{1, 2, 3\}$, with probability at least $1 - \delta$,*

$$\|\mathbb{E}_{p^\star}[XX^\top] - \mathbb{E}_{q_i}[XX^\top]\|_2 \leq C_1 \cdot (\rho(\tilde{\epsilon}) + \tilde{\epsilon}\log(1/\tilde{\epsilon})),$$

where $\tilde{\epsilon} = C_2(\epsilon + \sqrt{d/n} + \sqrt{\log(1/\delta)/n})$.

Remark 3. *When the true distribution lies in the set of sub-Gaussian distributions, the projection algorithm guarantees a rate of $C \cdot \tilde{\epsilon}\log(1/\tilde{\epsilon})$, which is near optimal up to logarithmic factors compared to the result for Gaussian [2].*

C. Linear Regression

For linear regression, we adopt a difference family of $\mathcal{F}$ in $\widetilde{\text{TV}}_{\mathcal{F}}$, which takes the form of $\{(Y - v_1^\top X)^2 - (Y - v_2^\top X)^2 \mid v_1, v_2 \in \mathbb{R}^d\}$. Let A_1, A_2, A_3 be the same as mean estimation in Equation (15), (16), (17) except that we set

$$g_1(v_1, v_2, t, X) = \text{sigmoid}((Y - v_1^\top X)^2$$
$$- (Y - v_2^\top X)^2 + t),$$

$$g_2(w, \{v_j^{(1)}\}, \{v_j^{(2)}\}, \{t_j\}, X)$$

$$= \text{sigmoid}\left(\sum_{j \leq l} w_j g_1(v_j^{(1)}, v_j^{(2)}, t_j, X)\right).$$

Although the loss for linear regression is not pseudonorm, and thus Theorem 1 is not directly applicable, we show below that under slight modification of the analysis in Theorem 1 that invokes the general property of generalized resilience in [3], the projection algorithm under the distances A_1, A_2, A_3 still guarantees robust regression.

Corollary 3. *Assume the true distribution $p^\star \in \mathcal{G}_{\text{reg}}(\psi)$ in Equation (2). Let the projection algorithm $q_1 = \arg\min_{q \in \mathcal{G}_{\text{reg}}} A_1(q, \hat{p}_n)$, and q_2, q_3 be the projection algorithm that projects under distance A_2, A_3. For any $i \in \{1, 2, 3\}$, with probability at least $1 - \delta$,*

$$\mathbb{E}_{p^\star}[(\theta^\star(q_i)^\top X - Y)^2 - (\theta^\star(p)^\top X - Y)^2]$$
$$\leq C_1 \cdot (\tilde{\epsilon}\psi^{-1}(1/\tilde{\epsilon}) + \tilde{\epsilon}\log(1/\tilde{\epsilon})),$$

where $\tilde{\epsilon} = C_2(\epsilon + \sqrt{d/n} + \sqrt{\log(1/\delta)/n})$.

Remark 4. *Corollary 3 recovers [3, Theorem 3.3] up to an additive term $\tilde{\epsilon}\log(1/\tilde{\epsilon})$. When X and $Y - X^\top\theta^\star(p^\star)$ are both sub-Gaussian, our dependence on ϵ is the same as the optimal rate in Gaussian example in [9] up to a log factor while the sample complexity matches [9] exactly. [29] guarantee an optimal parameter error $\|\hat{\theta} - \theta^*\|_2 \lesssim (\epsilon\log(1/\epsilon))^2$ given $O(d/\epsilon^2)$ samples when X is isotropic Gaussian and $Y - X^\top\theta^\star(p^\star)$ is Gaussian, while we can improve the results to $(\epsilon\sqrt{\log(1/\epsilon)})^2$, as is achieved in [30], by taking $\psi(x) = \exp(x^2)$ and using it as the generalized resilience set.*

Proof of Corollary 3. We slightly modify the last part of the proof in Theorem 1 for linear regression. Following the same proof as Theorem 1, we know for fixed $f = (Y - X^\top\theta^\star(q_i))^2 - (Y - X^\top\theta^\star(p^\star))^2$, one can find deletions $r_{p^\star}, r_{q_i}$ which delete at most $\tilde{\epsilon}$-fraction of the distributions $p^\star, q_i$, such that

$$\mathbb{E}_{r_{p^\star}}[(Y - X^\top\theta^\star(q_i))^2 - (Y - X^\top\theta^\star(p^\star))^2]$$
$$\leq \mathbb{E}_{r_q}[(Y - X^\top\theta^\star(q_i))^2 - (Y - X^\top\theta^\star(p^\star))^2] + 2\rho_Z(C\tilde{\epsilon})$$
$$\leq \mathbb{E}_{r_q}[(Y - X^\top\theta^\star(q_i))^2 - (Y - X^\top\theta^\star(r_{q_i}))^2] + 2\rho_Z(C\tilde{\epsilon})$$
$$\leq C\tilde{\epsilon}\psi^{-1}(1/\tilde{\epsilon}) + 2\rho_Z(C\tilde{\epsilon}). \qquad (18)$$

Here the last inequality comes from [3, Lemma F.3]. Invoking [3, Lemma F.3] again, we know that Equation (18) implies that

$$\mathbb{E}_{p^\star}[(Y - X^\top\theta^\star(q_i))^2 - (Y - X^\top\theta^\star(p^\star))^2]$$
$$\leq C_1(\tilde{\epsilon}\psi^{-1}(1/\tilde{\epsilon}) + \rho_Z(C_2\tilde{\epsilon})).$$

The rest of the proof follows directly from that of Corollary 1 by verifying the three assumptions in Theorem 1. $\quad\square$

V. ACKNOWLEDGEMENT

---

[1]Besides a different choice of $g_1, g_2, \mathcal{F}$, the only difference in the proof is the concentration of $\bar{A}(\hat{p}_n, p)$. One can follow a similar analysis in [2, Lemma 7.3] and [1, Lemma 8.1] to derive the concentration.

REFERENCES

[1] C. Gao, J. Liu, Y. Yao, and W. Zhu, "Robust estimation and generative adversarial nets," *arXiv preprint arXiv:1810.02030*, 2018.

[2] C. Gao, Y. Yao, and W. Zhu, "Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective," *arXiv preprint arXiv:1903.01944*, 2019.

[3] B. Zhu, J. Jiao, and J. Steinhardt, "Generalized resilience and robust statistics," *arXiv preprint arXiv:1909.08755*, 2019.

[4] P. J. Huber, "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, vol. 1, no. 5, pp. 799–821, 1973.

[5] D. L. Donoho and R. C. Liu, "The 'automatic' robustness of minimum distance functionals," *The Annals of Statistics*, vol. 16, no. 2, pp. 552–586, 1988.

[6] J. Adrover and V. Yohai, "Projection estimates of multivariate location," *The Annals of Statistics*, vol. 30, no. 6, pp. 1760–1781, 2002.

[7] P. J. Huber, *Robust Statistics*.   Springer, 2011.

[8] Z. Chen, D. E. Tyler *et al.*, "The influence function and maximum bias of Tukey's median," *The Annals of Statistics*, vol. 30, no. 6, pp. 1737–1759, 2002.

[9] C. Gao, "Robust regression via mutivariate regression depth," *arXiv preprint arXiv:1702.04656*, 2017.

[10] B. Zhu, J. Jiao, and J. Steinhardt, "When does the Tukey median work?" in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 1201–1206.

[11] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Being robust (in high dimensions) can be practical," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 999–1008.

[12] J. Steinhardt, M. Charikar, and G. Valiant, "Resilience: A criterion for learning in the presence of arbitrary outliers," *arXiv preprint arXiv:1703.04940*, 2017.

[13] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in Neural Information Processing Systems*, 2017, pp. 3517–3529.

[14] I. Diakonikolas, D. M. Kane, and A. Stewart, "Learning geometric concepts with nasty noise," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 1061–1073.

[15] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart, "Sever: A robust meta-algorithm for stochastic optimization," *arXiv preprint arXiv:1803.02815*, 2018.

[16] L. Liu, Y. Shen, T. Li, and C. Caramanis, "High dimensional robust sparse regression," *arXiv preprint arXiv:1805.11643*, 2018.

[17] M. Chen, C. Gao, and Z. Ren, "Robust covariance and scatter matrix estimation under Huber's contamination model," *The Annals of Statistics*, vol. 46, no. 5, pp. 1932–1960, 2018.

[18] A.-H. Bateni and A. S. Dalalyan, "Minimax rates in outlier-robust estimation of discrete models," *arXiv preprint arXiv:1902.04650*, 2019.

[19] G. Lecué and J. Depersin, "Robust subgaussian estimation of a mean vector in nearly linear time," *arXiv preprint arXiv:1906.03058*, 2019.

[20] B. Zhu, J. Jiao, and D. Tse, "Deconstructing generative adversarial networks," *arXiv preprint arXiv:1901.09465*, 2019.

[21] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*.   Springer, 1979, vol. 143.

[22] D. L. Donoho and R. C. Liu, "Pathologies of some minimum distance estimators," *The Annals of Statistics*, pp. 587–608, 1988.

[23] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2016, pp. 655–664.

[24] I. Diakonikolas, D. M. Kane, and A. Pensia, "Outlier robust mean estimation with subgaussian rates via stability," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1830–1840, 2020.

[25] G. Lugosi and S. Mendelson, "Robust multivariate mean estimation: the optimality of trimmed mean," *The Annals of Statistics*, vol. 49, no. 1, pp. 393–410, 2021.

[26] J. Depersin and G. Lecué, "Robust sub-gaussian estimation of a mean vector in nearly linear time," *The Annals of Statistics*, vol. 50, no. 1, pp. 511–536, 2022.

[27] G. Lugosi and S. Mendelson, "Mean estimation and regression under heavy-tailed distributions: A survey," *Foundations of Computational Mathematics*, vol. 19, no. 5, pp. 1145–1190, 2019.

[28] J.-L. Wu, K. Kashinath, A. Albert, D. Chirila, H. Xiao *et al.*, "Enforcing statistical constraints in generative adversarial networks for modeling chaotic dynamical systems," *Journal of Computational Physics*, vol. 406, p. 109209, 2020.

[29] I. Diakonikolas, W. Kong, and A. Stewart, "Efficient algorithms and lower bounds for robust linear regression," in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2019, pp. 2745–2754.

[30] A. Pensia, V. Jog, and P.-L. Loh, "Robust regression with covariate filtering: Heavy tails and adversarial contamination," *arXiv preprint arXiv:2009.12976*, 2020.