On Minimax Optimality of GANs for Robust Mean Estimation

Kaiwen Wu[1,2] **Gavin Weiguang Ding**[3] **Ruitong Huang**[3] **Yaoliang Yu**[1,2]
University of Waterloo[1] Vector Institute[2] Borealis AI[3]

Abstract

Generative adversarial networks (GANs) have become one of the most popular generative modeling techniques in machine learning. In this work, we study the statistical and robust properties of GANs for Gaussian mean estimation under Huber's contamination model, where an ϵ proportion of training data may be arbitrarily corrupted. We prove that f-GAN, when equipped with appropriate discriminators, achieve optimal minimax rate, hence extending the recent result of Gao et al. (2019a). In contrast, we show that other GAN variants such as MMD-GAN (with Gaussian kernel) and W-GAN may fail to achieve minimax optimality. We further adapt f-GAN to the sparse and the unknown covariance settings. We perform numerical simulations to confirm our theoretical findings and reveal new insights on the importance of discriminators.

1 Introduction

Robust estimation under Huber's contamination model (Huber, 1964) has been an important problem in statistics. Under this model, data are sampled *i.i.d.* from a contaminated distribution as follows:

$$X_i \sim (1-\epsilon)\mathbb{P}_\theta + \epsilon\mathbb{H},\ i = 1, \dots, n, \tag{1}$$

and the goal is to estimate θ given the observations $\{X_1, X_2, \cdots X_n\}$, where $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is the distribution class of interest, $\mathbb{H}$ is an arbitrary contamination distribution, and ϵ is the contamination proportion. In this work, we mostly focus on the case where $\mathbb{P}_\theta$ is a p-dimensional standard Gaussian distribution $\mathcal{N}(\theta, I_p)$, and the goal is to estimate its mean θ.

Proceedings of the 23[rd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Despite its simplicity, robust mean estimation in high dimensions remains a challenging task. Sample efficient estimators based on clean data (*i.e.* $\epsilon = 0$), such as the sample average, do not necessarily work well in the presence of contamination, while robust estimators designed in low dimensions, such as the coordinate-wise median and geometric median, may suffer the curse of dimensionality and result in suboptimal error bound in high dimensions. On the other hand, the well-known minimax optimal robust estimator, Tukey's median (Tukey, 1975), is computationally intractable (Amenta et al., 2000).

Only in recent years, polynomial-time computable robust estimators with minimax optimal error bounds start to appear (Diakonikolas et al., 2016; Lai et al., 2016), based on analysis using higher order moments. Since then, various efficient robust estimators, as well as adaptations and applications to other settings, have emerged (*e.g.* Balakrishnan et al., 2017; Diakonikolas et al., 2017, 2019a,b). Meanwhile, Gao et al. (2019a) have recently established a deep connection between generative adversarial networks (Goodfellow et al., 2014, GAN) and depth-based robust estimators (*e.g.* Tukey's median), leading them to study robust mean estimators using GANs and establish minimax optimal error bounds, provided that suitable architectures are chosen for the discriminator.

Although GANs have been extremely popular for generative modeling, their robustness properties are nevertheless not well-studied, even in restricted settings as we consider here. In this work, building on the results of Gao et al. (2019a), we provide further analysis of robust estimators based on different GAN variants under Huber's contamination model. Our analysis not only provides insightful guidance on developing robust estimators using GANs, but also reveals the subtle difference of different GAN variants in the presence of contamination. Unlike previous works that study the generalization of GANs on clean data (Arora et al., 2017; Feizi et al., 2017; Zhang et al., 2018; Thanh-Tung et al., 2019), we study generalization on clean data and robustness against contamination simultaneously. As we will see later, this subtle difference may

lead to different conclusions in certain cases. For instance, it might be easy for a weak discriminator to generalize well on clean data, *e.g.* MMD-GAN, but it may not guarantee robustness against contamination.

To summarize, we make the following contributions:

- We generalize previous results in (Gao et al., 2019a) and prove minimax optimal rate for most f-GANs equipped with suitable discriminators;
- For MMD-GAN with Gaussian kernel, we prove a finite sample bound $O(\sqrt{\frac{p}{n}} \vee \sqrt{p}\epsilon)$ and a matching lower bound, indicating that MMD-GAN may not achieve the minimax rate;
- We prove that the estimator for Wasserstein GAN in 1-dimension is robust to contamination, and empirically verify that Wasserstein GAN suffers $\Omega(\sqrt{p}\epsilon)$ error in high dimensions;
- We demonstrate that estimators based on f-GANs are flexible: they can be easily adapted to other learning settings, *e.g.* sparse means or unknown covariance matrix, and they again achieve the minimax rate with minor modifications.

Proofs are deferred to the appendix and code is available at https://github.com/watml/robust-gan.

2 Preliminary

In this section we first define the robust distribution estimation problem and its Gaussian mean specialization. Then, we recall some existing results. Our notations are mostly standard: we use $a \vee b$ to denote $\max\{a, b\}$; $a_n \lesssim b_n$ denotes $a_n \le Cb_n$ for some absolute constant C and any sufficiently large n; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$; we use $\|\cdot\|$ to denote the Euclidean norm and $\|\cdot\|_2$ the induced spectral norm.

Let $\mathbb{P} \in \mathcal{P}$ be an unknown distribution on $\mathcal{X} \subseteq \mathbb{R}^p$ that we are interested in estimating, and $\mathbb{Q} \in \mathcal{Q}_{\mathbb{P}}$ some perturbation of $\mathbb{P}$. Given an *i.i.d.* sample $\mathbf{X}_{1:n} = \rbag X_1, \dots, X_n \lbag \sim \mathbb{Q}$, our goal is to construct an estimator $\hat{\mathbb{P}}_n : \mathcal{X}^n \to \mathcal{P}$ that achieves small maximal risk under some (semi)metric $\mathsf{d} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$:

$$\mathcal{R}_n(\hat{\mathbb{P}}_n) := \sup_{\mathbb{P}\in\mathcal{P}} \sup_{\mathbb{Q}\in\mathcal{Q}_{\mathbb{P}}} \mathbf{E}[\mathsf{d}(\hat{\mathbb{P}}_n(\mathbf{X}_{1:n}), \mathbb{P})]. \tag{2}$$

We define the minimax risk as the infimum over all (measurable) estimators:

$$\mathcal{R}_n^* = \inf_{\hat{\mathbb{P}}_n : \mathcal{X}^n \to \mathcal{P}} \mathcal{R}_n(\hat{\mathbb{P}}_n), \tag{3}$$

and we call an estimator $\hat{\mathbb{P}}_n$ minimax optimal if it achieves the minimax risk asymptotically, *i.e.*, $\mathcal{R}_n(\hat{\mathbb{P}}_n) \asymp \mathcal{R}_n^*$. Note that $\mathcal{Q}_{\mathbb{P}} = \{\mathbb{P}\}$ recovers the usual notion of minimax optimality (*e.g.* Tsybakov, 2009).

We will focus on Huber's ϵ-contamination model which is defined below (Huber, 1964):

$$\mathcal{Q}_{\mathbb{P}} = \mathcal{Q}_{\mathbb{P},\epsilon} := \{(1-\epsilon)\mathbb{P} + \epsilon\mathbb{H} : \mathbb{H} \text{ any distribution on } \mathcal{X}\},$$

i.e., with probability $1-\epsilon$ a sample X comes from the clean distribution $\mathbb{P}$ while with probability ϵ it comes from an arbitrary contamination distribution $\mathbb{H}$. Equivalently (see Álvarez-Esteban et al. (2011)),

$$\mathbb{Q} \in \mathcal{Q}_{\mathbb{P}} \iff \mathbb{Q} \ge (1-\epsilon)\mathbb{P} \iff \tfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \le \tfrac{1}{1-\epsilon}.$$

In fact, the ϵ-contamination model is a special case of total variation perturbation, since

$$\mathbb{Q} \in \mathcal{Q}_{\mathbb{P}} \Rightarrow \mathsf{TV}(\mathbb{P}, \mathbb{Q}) \le \epsilon,$$

where recall that $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) := \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|$.

Our main goal is to understand when the minimum variational discrepancy estimator, a.k.a. generative adversarial networks (GAN),

$$\hat{\mathbb{P}}_n(\mathbf{X}_{1:n}) := \operatorname*{argmin}_{\mathbb{P}\in\mathcal{P}} \mathsf{D}(\hat{\mathbb{Q}}_n, \mathbb{P}), \ \hat{\mathbb{Q}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i} \tag{4}$$

is minimax optimal, where

$$\mathsf{D}(\mathbb{Q}, \mathbb{P}) := \sup_{T\in\mathcal{T}} \mathbf{E}_{\mathbb{Q}}[T(X)] - \mathbf{E}_{\mathbb{P}}[s(T(Y))], \tag{5}$$

is some variational discrepancy measure between two distributions, $s : \mathbb{R} \to \mathbb{R}$ is a fixed function, and $\mathcal{T}$ is a chosen set of real-valued test functions on $\mathcal{X}$. Note that the discrepancy measure D that we use to construct the estimator $\hat{\mathbb{P}}_n$ is usually different from the distance d that we employ to evaluate $\hat{\mathbb{P}}_n$. Crucially, the estimator $\hat{\mathbb{P}}_n$ above does not need to know the contamination proportion ϵ hence is adaptive. Different choices of $\mathcal{T}$ and s lead to different variants of GAN (examples will follow), and as we shall see, the choice of test functions $\mathcal{T}$ (also known as discriminators) plays a decisive role in achieving minimax (sub)optimality.

The robust Gaussian mean estimation problem refers to the special case:

$$\mathcal{P} = \{\mathcal{N}(\theta, I_p) : \theta \in \mathbb{R}^p\}, \tag{6}$$

where for simplicity we assume the covariance matrix is identity. This problem has been widely studied. In his seminal work, Huber (1964) considered the univariate case and discovered the "optimal" estimator in terms of minimizing the asymptotic variance. Let

$$\mathsf{d}(\mathcal{N}(\hat{\theta}, I_p), \mathcal{N}(\theta, I_p)) = \|\hat{\theta} - \theta\| \tag{7}$$

be the Euclidean distance. Chen et al. (2018) proved that the minimax risk (see (3)) has the following order:

$$\mathcal{R}_n^* \asymp \sqrt{\frac{p}{n}} \vee \epsilon, \tag{8}$$

where the first term $\sqrt{\frac{p}{n}}$ describes the sample efficiency of the estimator while the second term ϵ describes its robustness. Note that the latter ϵ term cannot be avoided even when sample size grows to infinity. Standard (and efficient) estimators such as sample average has infinite risk while more robust ones such as coordinate-wise median or geometric median achieve suboptimal risk ($\sqrt{\frac{p}{n}} \vee \sqrt{p}\epsilon$, a factor of $\sqrt{p}$ off). Chen et al. (2018) also proved that Tukey's median (Tukey, 1975), although NP-hard to compute (Amenta et al., 2000), does achieve minimax optimality. Around the same time, Diakonikolas et al. (2016); Lai et al. (2016) discovered estimators that are both minimax-optimal and polynomial-time computable. However, this latter estimator is based on high-order moments which do not fit into the GAN framework in (4).

Very recently, Gao et al. (2019a) revealed a deep connection between robust estimation and GANs. Their main result confirmed that total variation GAN (TV-GAN), where $s(t) = t\mathbf{1}_{t\in[0,1]}$ in (5), and Jensen-Shannon GAN (JS-GAN), where $s(t) = -\log(2 - \exp(t))$ in (5), both achieve the minimax rate if we equip them with suitable discriminators $\mathcal{T}$. Our first main contribution is to further substantiate the results of (Gao et al., 2019a) in two aspects: (a) we prove that the entire f-GAN family in (Nowozin et al., 2016) is minimax optimal, once equipped with appropriate discriminators; (b) in contrast, other popular GAN variants, such as the MMD-GAN (Li et al., 2015; Dziugaite et al., 2015; Li et al., 2017) and Wasserstein GAN (Arjovsky et al., 2017), may not achieve the minimax rate.

From now on, we will restrict to the robust Gaussian mean estimation problem, with (6) and (7) always kept in mind. Extensions to other settings will be discussed in Section 6.

3 f-GAN

We first consider using f-GAN (Nowozin et al., 2016) for the robust Gaussian mean estimation problem. f-GAN is based on minimizing the f-divergence between two probability distributions. Assuming that both $\mathbb{P}$ and $\mathbb{Q}$ have probability density functions $p(x)$ and $q(x)$ respectively, the f-divergence between $\mathbb{Q}$ and $\mathbb{P}$ is defined as follows:

$$\mathcal{D}_f(\mathbb{Q}||\mathbb{P}) = \int f\left(\frac{q(x)}{p(x)}\right) p(x)\,\mathrm{d}x, \tag{9}$$

where $f : \mathbb{R}_+ \to \mathbb{R}$ is a (strictly) convex function satisfying $f(1) = 0$. The f-divergence (9) has the following equivalent dual representation:

$$\mathcal{D}_f(\mathbb{Q}||\mathbb{P}) = \sup_{T\in\mathcal{T}} \mathbf{E}_{\mathbb{Q}}[T(X)] - \mathbf{E}_{\mathbb{P}}[f^\star(T(X))],$$

where $f^\star(t) = \sup_x xt - f(x)$ is the convex conjugate of f, and $\mathcal{T}$ is the class of all measurable functions, which is typically represented with a neural network. It is clear that choosing $s(t) = f^\star(t)$ recovers the general formulation (5). Some examples of common f-divergences and their conjugate functions can be found in Table 1. Keeping (6) in mind, the estimator produced by f-GAN is

$$\hat{\theta}_n = \operatorname*{argmin}_{\eta\in\mathbb{R}^p} \sup_{T\in\mathcal{T}} \mathbf{E}_{\hat{\mathbb{Q}}_n} T(X) - \mathbf{E}_{\mathcal{N}(\eta,I)} f^\star(T(Y)). \tag{10}$$

Gao et al. (2019a) proved that TV-GAN and JS-GAN both achieve the minimax rate, which we first generalize to the entire f-GAN family, under mild assumptions on f and the test function class $\mathcal{T}$:

Assumption 1. *The discriminator $\mathcal{T}$ is parameterized by a two-layer network:*

$$\mathcal{T} = \left\{ g\left(\sum_{i=1}^{l} w_i \sigma(u_i^\top x + b_i)\right) : \|w\|_1 \le \kappa \right\},$$

where σ is the sigmoid function, κ is a constant, ℓ is the number of hidden neurons.

Assumption 2. *The functions f and g satisfy the following properties: 1) both $f^\star$ and g are twice continuously differentiable; 2) $g' > 0$; 3) $g(0) \in \partial f(1)$.*

We note that the assumptions on f and g are mild. The twice continuously differentiable assumption is satisfied by all common f-divergences. The strict monotonicity of the activation function g is satisfied in practice and is suggested in (Nowozin et al., 2016) such that the network output encodes the confidence about whether a given input comes from the training data or the generator distribution. The third assumption can always be satisfied by simply shifting g.

We have the following minimax optimal convergence rate guarantee for f-GAN. [1]

Theorem 1. *Let $\hat{\theta}$ be the estimator defined in* (10)*, where f and g satisfy Assumption 2 and $\mathcal{V}$ satisfies Assumption 1. Assuming that $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon \le c$ for some sufficiently small constant c, then with probability at least $1 - \delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{\frac{p}{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}. \tag{11}$$

[1] Note that this bound is stated in probability. Rigorously speaking, it is not in the same form as the minimax rate (8), which is in expectation. However, one can transform the bound (11), by integrating the tail distribution of $\|\hat{\theta}_n - \theta\|$, into a bound in expectation that matches the minimax rate (8). For simplicity, we always state the bounds in probability in the following.

Table 1: Examples of common f-divergences and the corresponding activation functions.

Divergence	$f(x)$	$f^\star(t)$	$g(v)$	$g(0)$
Kullback-Leibler (KL)	$x\log x$	$\exp(t-1)$	$v+1$	1
Reverse KL (RKL)	$-\log x$	$-1-\log(-t)$	$-\exp(-v)$	-1
Squared Hellinger (SH)	$(\sqrt{x}-1)^2$	$\frac{t}{1-t}\quad(t<1)$	$1-\exp(-v)$	0
Jensen Shannon (JS)	$-(x+1)\log\frac{1+x}{2}+x\log x$	$-\log(2-\exp(t))$	$\log(2)-\log(1+\exp(-v))$	0
Total Variation (TV)	$\max\{x-1,0\}$	$t\quad(0\le t\le 1)$	$\frac{1}{1+\exp(-v)}$	$\frac{1}{2}$

Table 2: A comparison of the estimation error between f-GANs and the filtering method (Diakonikolas et al., 2016, 2017). We set $n=50000$, $\epsilon=0.2$, $p=100$. The target distribution is $\mathcal{N}(\mathbf{0},I_p)$. Numbers are averaged over 5 runs with standard deviations shown in parenthesis.

Contamination	KL-GAN	RKL-GAN	SH-GAN	JS-GAN	TV-GAN	Filtering
$\mathcal{N}(0.05\cdot\mathbf{1},I_p)$	0.1093 (0.0007)	0.1115 (0.0024)	0.1095 (0.0014)	**0.1092** (0.0013)	0.1106 (0.0020)	0.1122 (0.0025)
$\mathcal{N}(0.1\cdot\mathbf{1},I_p)$	0.2053 (0.0018)	0.2057 (0.0018)	0.2059 (0.0030)	**0.2040** (0.0016)	0.2055 (0.0016)	0.2062 (0.0031)
$\mathcal{N}(0.2\cdot\mathbf{1},I_p)$	0.4030 (0.0037)	0.4047 (0.0037)	0.4040 (0.0033)	0.4032 (0.0027)	0.4034 (0.0026)	**0.3986** (0.0029)
$\mathcal{N}(0.5\cdot\mathbf{1},I_p)$	0.9166 (0.1589)	**0.0786** (0.0046)	0.1027 (0.0058)	0.0851 (0.0050)	1.4163 (0.0166)	0.1436 (0.0074)
$\mathcal{N}(1\cdot\mathbf{1},I_p)$	1.0806 (0.3101)	**0.0559** (0.0021)	0.0599 (0.0025)	0.0683 (0.0022)	8.4945 (0.0081)	0.1417 (0.0091)
$\mathcal{N}(2\cdot\mathbf{1},I_p)$	1.9081 (0.4246)	**0.0552** (0.0042)	0.0613 (0.0026)	0.0736 (0.0018)	9.5560 (0.1185)	0.1443 (0.0063)
$\mathcal{N}(3\cdot\mathbf{1},I_p)$	2.0420 (0.4291)	**0.0559** (0.0046)	0.0596 (0.0029)	0.0696 (0.0022)	7.3902 (0.1156)	0.1471 (0.0043)

We list common f-divergences along with their corresponding activation functions g in Table 1. Theorem 1 includes all previous attempts as special cases. In particular, it shows that the choice of f-divergence does not matter statistically, as long as one chooses the right activation function and the appropriate discriminator. This is probably not too surprising, given the fact that TV-GAN and JS-GAN can already achieve the minimax rate and that we parametrize the discriminator all in a similar way in Assumption 1. However, the choice of f still plays an important role when it comes to optimization. For instance, we empirically observed that TV-GAN and KL-GAN are much harder to train than SH-GAN, RKL-GAN and JS-GAN. The subtle training difficulty of TV-GAN is also observed in (Gao et al., 2019a).

We present numerical simulation results in Table 2 to compare the estimation error of different f-GANs and that of the filtering method in (Diakonikolas et al., 2016, 2017), which is a polynomial-time algorithm with provable error bounds. We notice that KL-GAN and TV-GAN can diverge easily in some cases. However, SH-GAN, RKL-GAN and JS-GAN typically converge steadily and often achieve comparable performance against the filtering method.

4 MMD-GAN

Surprisingly, unlike f-GAN, MMD-GAN (with Gaussian kernel) cannot achieve the minimax rate, as we show in this section. In particular, we prove a convergence rate of $\sqrt{\frac{p}{n}}\vee\sqrt{p}\epsilon$ for MMD-GAN, which we then confirm is tight but is also a factor of $\sqrt{p}$ off.

MMD-GAN (Li et al., 2015; Dziugaite et al., 2015; Li et al., 2017) is developed based on the maximum mean discrepancy (Gretton et al., 2012, MMD), whose discriminator is the unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ induced by some kernel k:

$$\mathcal{T}=\{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\le 1\}. \tag{12}$$

For simplicity, we only consider the Gaussian kernel

$$k(x,x')=\exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right), \tag{13}$$

which is popular in practice (Bikowski et al., 2018). We set $s(t)=t$ in (5) and obtain the MMD-GAN estimator formally as:

$$\hat{\theta}_n=\operatorname*{argmin}_{\eta\in\mathbb{R}^p}\sup_{f\in\mathcal{T}}\mathbf{E}_{\hat{\mathbb{Q}}_n}f(X)-\mathbf{E}_{\mathcal{N}(\eta,I_p)}f(Y). \tag{14}$$

Our first result is a finite sample upper bound for the estimator via MMD-GAN.

Theorem 2. *Let $\mathcal{T}$ be the RKHS unit ball induced by the Gaussian kernel with bandwidth σ. For the estimator defined in* (14)*, with probability at least $1-\delta$,*

$$\|\hat{\theta}_n-\theta\|\lesssim(2+\sigma^2)^{\frac{1}{2}}(1+\frac{2}{\sigma^2})^{\frac{p}{4}}\left(\frac{1}{\sqrt{n}}\vee\epsilon+\sqrt{\frac{\log 1/\delta}{n}}\right).$$

Note that for any fixed σ, the upper bound in Theorem 2 is exponential in p, which is even worse than classical estimators such as coordinate-wise median.

However, it is possible to achieve a far better convergence rate, by choosing σ adaptively according to the dimension p. Indeed, our next result shows that when $\sigma = \sqrt{p}$, the bound can be significantly improved into linear dependence on $\sqrt{p}$.

Corollary 1. *Let $\mathcal{F}$ be the RKHS unit ball induced by the Gaussian kernel with bandwidth $\sigma = \sqrt{p}$, then with probability at least $1 - \delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{p}\left(\frac{1}{\sqrt{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}\right). \tag{15}$$

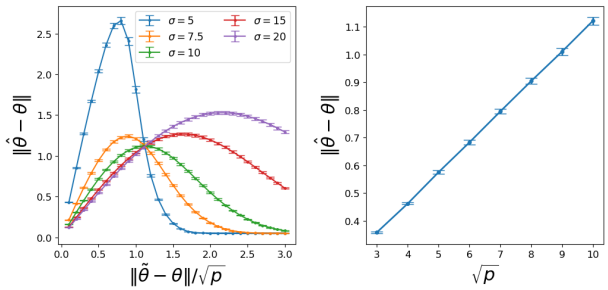
The robust error (15) now grows linearly w.r.t. the square root of dimension. Nevertheless, it still fails to match the minimax optimal rate, and this dimensional dependence appears to be unavoidable in our experiments. Indeed, we next prove that this square root dependence on the dimension is tight even for the population limit of MMD-GAN, *i.e.*, even when we have access to infinitely many samples, the robust error still grows linearly w.r.t. $\sqrt{p}$.

Theorem 3. *Consider the population limit of $\hat{\theta}$ given by MMD-GAN. For any $\sigma > 0$, there always exists a contaminated distribution $\mathbb{Q}$ such that*

$$\|\hat{\theta} - \theta\| \gtrsim \sqrt{p}\epsilon. \tag{16}$$

Our technique to prove Theorem 3 is to construct the least favorable contamination *explicitly*, and analyze the resulting *nonconvex* landscape. Contrary to one might expect, the strongest contamination $\mathbb{H}$ is not a distribution that is very far away from θ. For instance, if one pick the contamination $\mathbb{H}$ to be a Dirac measure at infinity, MMD-GAN will simply give up in minimizing the error caused by contamination, and consequently it recovers the true parameter θ accurately. In turn, we find that one strong contamination $\mathbb{H}$ is a Dirac measure $\delta_{\tilde{\theta}}$ that is $\sqrt{p}$ close to the center of the target distribution, *i.e.*, $\|\theta - \tilde{\theta}\| \approx \sqrt{p}$.

Theorem 2 and Theorem 3 together establish a suboptimal convergence rate of MMD-GAN. This is in stark contrast to the no contamination setting ($\epsilon = 0$) where MMD-GAN and many other standard estimators all achieve the $\sqrt{\frac{p}{n}}$ minimax rate. Moreover, MMD discriminators in (12) are known to metricize weak convergence of distributions and enjoy small sample complexity. Indeed, the discriminator function class in (12) is relatively small: its Rademacher complexity is of order $\frac{1}{\sqrt{n}}$. In contrast, the Rademacher complexity of the f-GAN discriminators (see (1)) is of order $\sqrt{\frac{p}{n}}$. While a smaller complexity of the function class ensures a smaller sample complexity, and thus a more accurate estimation of the underlying discriminator, it does not guarantee better estimation of the true distribution. Even worse, a smaller complexity of the function class may come with the price of a higher robustness error in the presence of contamination, as we showed for MMD-GAN.



(a) different σ and $\delta_{\tilde{\theta}}$ (b) different dimension p

Figure 1: Numerical simulation of MMD-GAN. Results are averaged over 3 runs with the standard deviations indicated by the error bars. We set $\epsilon = 0.1$ and $n = 50000$. We always pick $\mathbb{H}$ to be a Dirac measure $\delta_{\tilde{\theta}}$ and tune $\tilde{\theta}$ accordingly. **Left:** The estimation error using different σ against different contamination $\mathbb{H}$. When $\sigma = \sqrt{p} = 10$, MMD-GAN achieves the minimum worst-case error. **Right:** The estimation error w.r.t. the square root of dimension. The error grows linearly as the dimension increases.

To verify our theory, we present some numerical simulation results in Figure 1. In Figure 1a, we plot the estimation error against different contamination distribution $\mathbb{H}$ when using different bandwidth σ. As shown in the proof of Theorem 3, one strong contamination is a Dirac measure $\delta_{\tilde{\theta}}$ satisfying $\|\theta - \tilde{\theta}\| \approx \sqrt{p}$. Thus, we always pick $\mathbb{H} = \delta_{\tilde{\theta}}$ and tune $\tilde{\theta}$ accordingly. We set $p = 100$ in this experiment. It is clear that MMD-GAN achieves the minimum worst-case error when $\sigma = \sqrt{p} = 10$, matching the result of Corollary 1. Meanwhile, for $\sigma = 10$, the worst-case error is achieved when $\|\theta - \tilde{\theta}\| \approx \sqrt{p}$, which matches the result in Theorem 3. Moreover, it also confirms that large $\tilde{\theta}$ does not necessarily lead to large estimation error. In fact, the estimation error starts to decrease if the contamination distribution $\mathbb{H}$ is too far away from the true distribution $\mathbb{P}_\theta$. For example, when $\sigma = 5$, the estimation error is almost zero for $\|\theta - \tilde{\theta}\| \geq 2\sqrt{p}$.

In Figure 1b, we plot the estimation error w.r.t. $\sqrt{p}$. Again, we choose the contamination $\mathbb{H} = \delta_{\tilde{\theta}}$ according to the proof of Theorem 3. We tune a few $\tilde{\theta}$ around $\|\theta - \tilde{\theta}\| \approx \sqrt{p}$ and take the worst case error among them. The bandwidth σ is set to $\sqrt{p}$, as this is usually the best parameter, verified both theoretically and empirically. As we can see, the estimation error grows linearly w.r.t. the square root of dimension.

5 Wasserstein GAN

In this section we consider Wasserstein GAN (with the Euclidean norm as the ground cost), whose discriminator is the set of all Lipschitz continuous functions,

$$\mathcal{T} = \{f : |f(x) - f(y)| \le \|x - y\|, \forall x, y \in \mathcal{X}\}, \quad (17)$$

and its associated Wasserstein GAN (W-GAN) estimator (Arjovsky et al., 2017):

$$\hat{\theta}_n = \operatorname*{argmin}_{\eta \in \mathbb{R}^p} \sup_{f \in \mathcal{T}} \mathbf{E}_{\hat{\mathbb{Q}}_n} f(X) - \mathbf{E}_{\mathcal{N}(\eta, I_p)} f(Y), \quad (18)$$

which arises from (5) with $s(t) = t$. It is well-known that the discrepancy (5) in this case corresponds to the dual form of Wasserstein distance, with the Euclidean norm as the ground cost. The estimator (18) minimizes the Wasserstein distance $\mathcal{W}_1(\hat{\mathbb{Q}}_n, \mathcal{N}(\eta, I_p))$ between training data and the generator distribution.

On the one hand, the sample complexity of estimating the Wasserstein distance suffers an exponential dependence w.r.t. p, $O(n^{-\frac{1}{p}})$ (Peyré et al., 2019), requiring exponentially many samples to achieve an accurate estimate. Therefore, the sample efficiency of the estimator in (18) will not match the $O(\sqrt{\frac{p}{n}})$ term in the minimax rate under Huber's contamination model. On the other hand, the estimate of Wasserstein distance itself is not robust either. For instance, $\mathcal{W}\left(\mathbb{P}_\theta, (1-\epsilon)\mathbb{P}_\theta + \epsilon\delta_{\tilde{\theta}}\right)$ diverges to infinity as $\tilde{\theta} \to \infty$, since the Wasserstein distance between two distributions is lower bounded by the Euclidean distance between their means. Namely, a small fraction of outliers can make the estimate of Wasserstein distance arbitrarily large.

Nevertheless, the minimizer of Wasserstein distance, *i.e.* the estimator (18), may still exhibit some robustness in low dimensions. We demonstrate this for $p = 1$ and show that the estimator in (18) is robust by studying its population risk.

Theorem 4. *Consider W-GAN with $p = 1$. Let the contamination distribution $\mathbb{H} = \delta_{\tilde{\theta}}$. Suppose ϵ is sufficiently small, then $|\theta - \hat{\theta}| \lesssim \epsilon$. Further, there exists a contamination distribution such that $|\theta - \hat{\theta}| \gtrsim \epsilon$.*

However, in high dimensional spaces, W-GAN may not achieve the minimax rate. We demonstrate this by empirically verifying that the estimation error of W-GAN is $O(\sqrt{p}\epsilon)$.

Since W-GAN essentially minimizes the Wasserstein distance between the model and the sample empirical distribution, we directly minimize this quantity as opposed to using a neural network to approximate the set of all Lipschitz functions, in the following way:

$$\hat{\theta}_n = \operatorname*{argmin}_{\eta \in \mathbb{R}^p} \mathcal{W}\left(\hat{\mathbb{Q}}_n, \mathcal{N}(\eta, I_p)\right), \quad (19)$$

Algorithm 1: Minimize the Wasserstein Distance

Input: $\mathbf{a} = \frac{1}{n}\mathbf{1}_n$, $\mathbf{b} = \frac{1}{m}\mathbf{1}_m$, $\lambda > 0$, η_1

for $k = 1, 2, 3 \cdots$ **do**
- Sample $\{x_i\}_{i=1}^n$ from the training set.
- Sample $\{y_j\}_{j=1}^m$ from $\mathcal{N}(\eta_k, I_p)$.
- Compute $C_{ij} = \|x_i - y_j\|$, $K_{ij} = \exp(-C_{ij}/\lambda)$.
- **for** $l = 1, 2, \cdots L$ **do**
  - $\mathbf{f}^{(l+1)} = \lambda \log \mathbf{a} - \lambda \log\left(\mathbf{K} e^{\mathbf{g}^{(l)}/\lambda}\right)$
  - $\mathbf{g}^{(l+1)} = \lambda \log \mathbf{b} - \lambda \log\left(\mathbf{K} e^{\mathbf{f}^{(l+1)}/\lambda}\right)$
- $\eta_{k+1} = \eta_k - \alpha_k \nabla_\eta \langle e^{\mathbf{f}^{(L)}} K e^{\mathbf{g}^{(L)}}, C\rangle$

which is essentially the primal form of (18). Our motivation is to avoid the difficulty in min-max optimization and to ensure that we are minimizing the true Wasserstein distance. Since the primal problem of computing the Wasserstein distance is a linear program, it can be computed very accurately. To speed up computation, we use Sinkhorn iteration (Peyré et al., 2019) to solve an entropic regularized version of the Wasserstein distance. Recall that the entropic regularized optimal transport between two empirical distributions is formulated as follows:

$$\begin{aligned} \underset{\Pi \ge 0}{\text{minimize}} \;\; & \langle \Pi, C\rangle + \lambda \sum_{i=1}^n \sum_{j=1}^m \Pi_{ij} \log \Pi_{ij} \\ \text{s.t.} \;\; & \Pi \mathbf{1}_m = \mathbf{a}, \Pi^\top \mathbf{1}_n = \mathbf{b}, \end{aligned}$$

whose dual problem is

$$\underset{\mathbf{f}, \mathbf{g}}{\text{maximize}} \;\; \langle \mathbf{f}, \mathbf{a}\rangle + \langle \mathbf{g}, \mathbf{b}\rangle - \lambda \langle e^{\mathbf{f}/\lambda}, \mathbf{K} e^{\mathbf{g}/\lambda}\rangle,$$

where $\mathbf{K}_{ij} = \exp(-C_{ij}/\lambda)$. Alternating maximization on the dual variables $\mathbf{f}$ and $\mathbf{g}$ yields the Sinkhorn iteration algorithm. We use gradient descent to optimize (19) and call Sinkhorn iteration to evaluate the objective and its gradient. More specifically, we run the Sinkhorn iteration until converge, and then differentiate its output w.r.t. η directly. The full procedure is shown in Algorithm 1.

Incorporating entropic regularization for computing the Wasserstein distance has become popular since the work of Cuturi (2013), and has been previously applied to training generative models (Genevay et al., 2018). The approximation error of entropic regularization decays exponentially fast (Cominetti and Martín, 1994; Weed, 2018), thus it does not change the estimation result much.

We present the simulation results for W-GAN in Figure 2. In Figure 2a, we show the one dimensional result. We use excessive number of samples $n = 50000$ and small dimensions so that the sampling error is

negligible when compared to the contamination error $\epsilon = 0.1$. The estimator is always bounded for various contaminations, which include a Gaussian distribution with unit variance and a Dirac distribution. The estimation error increases as the contamination moves further away from the target distribution, but eventually stops increasing after reaching its peak. However, in higher dimensions, we can see that the estimation error grows linearly w.r.t. $\sqrt{p}$, as shown in Figure 2b. In addition, we observe that the entropic regularization constant λ does not change the results much, since we pick λ to be a small constant such that it does not affect the global minimizer of (19) noticably.

It remains an open question for high dimensions, *i.e.* $p > 1$. Although technical difficulties have so far defied us to develop theoretical guarantees for the high dimensional case, our empirical results suggest that the estimation error of Wasserstein GAN is $O(\sqrt{p}\epsilon)$, even with sufficiently many samples.

While MMD-GAN may not achieve minimax rate because of its small discriminator class, Wasserstein GAN still appears to be suboptimal even though its class of Lipschitz functions is much larger in complexity. We hypothesize the reason to be the unboundedness of Lipschitz functions, and we believe discriminators with bounded range is a crucial property for inducing robust estimators.

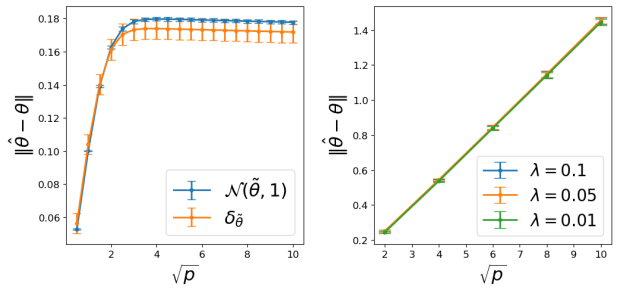
Notice that we focus specifically on the class of Lipschitz functions (w.r.t. Euclidean norm) in this section. The statistical property of a subset of Lipschitz functions might be very different. For example, one can easily show that the solution to (18) is the sample average if the discriminator (17) is restricted to the set of linear Lipschitz-1 functions. In practice, the discriminator is usually parametrized by a neural network, with gradient penalty to enforce Lipschitz constraint (Gulrajani et al., 2017). We trained three such variants of W-GANs (with ReLU and sigmoid activations) in this manner and still find they are not robust in high dimensions (see appendix).

6 Adaptation to Other Settings

We extend our results on f-GAN to more general settings in this section, with minor modifications on the discriminator and the generator.

6.1 Adaptation to Sparsity

We consider the problem of estimating a sparse Gaussian mean in this section. Assume that the target distribution is $\mathcal{N}(\theta, I_p)$ where the mean θ has at most s nonzero entries, in notation $\|\theta\|_0 \le s$. The learning goal is to estimate θ with the prior knowledge that



(a) WGAN in 1 dimension (b) WGAN in p dimension

Figure 2: **Left:** The estimation error w.r.t. the distance between the centers of $\mathbb{H}$ and $\mathbb{P}_\theta$. The estimation error remains bounded for various $\mathbb{H}$. **Right:** The estimation error w.r.t. the square root of dimension. The error grows linearly as the dimension increases.

at most s entries in θ are nonzero. Given n samples $\mathbf{X}_{1:n} \sim \mathbb{Q}$, our estimator is defined as

$$\hat{\theta}_n = \operatorname*{argmin}_{\|\eta\|_0 \le s} \sup_{T \in \mathcal{T}} \mathbf{E}_{\hat{\mathbb{Q}}_n} T(X) - \mathbf{E}_{\mathcal{N}(\eta, I_p)} f^\star(T(Y)), \quad (20)$$

where the discriminator function class $\mathcal{T}$ is defined as

$$\mathcal{T} = \Big\{ g\Big(\sum_{i=1}^{l} w_i \sigma(u_i^\top x + b_i) \Big) : \|w\|_1 \le \kappa,\ \|u\|_0 \le 2s \Big\}.$$

Note that the only difference between the above estimator and the one of f-GAN in Section 3 is the extra constraints to incorporate the prior knowledge of sparsity. The next theorem shows that the sample efficiency of the estimator is improved by incorporating the sparsity information.

Theorem 5. *Assuming that* $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon \le c$ *for some sufficiently small constant* c*, with probability at least* $1 - \delta$*, the estimator defined in* (20) *satisfies*

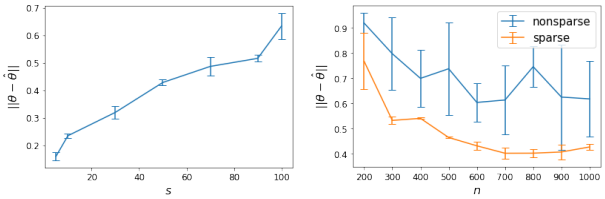
$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{\frac{s \log \frac{ep}{s}}{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}. \quad (21)$$

Notice that $s = p$ recovers the $\sqrt{\frac{p}{n}}$ convergence rate in the nonsparse setting. Generally, the convergence rate is better when $s < p$. Further, we demonstrate that the above convergence rate (21) is tight.

Theorem 6. *Let* $\Theta = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s\}$ *and* $\mathbb{P}_\theta = \mathcal{N}(\theta, I_p)$*. There exist absolute constants* c_1 *and* c_2*, such that for any estimator* $\hat{\theta}$*,*

$$\sup_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{Q}_\theta} \mathbb{Q}\left(\|\hat{\theta} - \theta\| \ge c_1 \left(\sqrt{\frac{s \log ep/s}{n}} \vee \epsilon \right) \right) \ge c_2.$$

Theorem 6, together with Theorem 5, establishes the minimax rate of sparse Gaussian mean estimation under Huber's contamination model.



(a) varying s (b) sparse vs. nonsparse

Figure 3: **Left:** The estimation error w.r.t. different s. Smaller s has smaller estimation error. **Right:** The estimation errors of the sparse and nonsparse estimators w.r.t. a true sparse Gaussian mean. The sparse estimator has smaller estimation error as well as smaller variance than its nonsparse counterpart.

We present numerical simulation results in Figure 3. In our experiments, we compute the estimator (20) by alternating gradient ascent and descent. We find that using the projection method[2] to handle the ℓ_0 norm constraint gives reasonable estimation results. We choose JS-GAN to implement the estimator (20) since it is easier to optimize empirically. In Figure 3a, we show the estimation error w.r.t. different sparsity levels. We set $p = 100$, $n = 500$ and $\epsilon = 0.05$ with various s. Notice that $\frac{s \log \frac{ep}{s}}{n} \gg \epsilon^2$, thus the convergence rate (21) is dominated by $\frac{s \log \frac{ep}{s}}{n}$. As expected, smaller s leads to a smaller estimation error. Figure 3b shows a comparison between the sparse estimator (20) and the nonsparse estimator (10). We fix $p = 100$, $s = 50$ and $\epsilon = 0.2$, with various sample size n. As sample size n increases, the estimation error of both estimators decreases. The sparse estimator always has smaller estimation error than its nonsparse counterpart, as well as smaller variance, since the former has a much smaller sample complexity.

A few recent works (*e.g.* Balakrishnan et al., 2017; Diakonikolas et al., 2019b) have developed polynomial-time algorithms for sparse mean estimation. However, their sample complexity is not minimax optimal. On the other hand, we focus on establishing the statistical properties of GANs and leave its computational complexity as future work.

6.2 Adaptation to Unknown Covariance

We further extend our f-GAN result to the setting where the covariance matrix Σ is unknown, and the learning goal is to estimate the mean without knowing the covariance matrix. Assuming that the target distribution is $\mathcal{N}(\theta, \Sigma)$ whose covariance matrix has spectral norm bounded by M, our estimator is defined as

$$(\hat{\theta}_n, \hat{\Sigma}) = \operatorname*{argmin}_{\eta \in \mathbb{R}^p, \|\Gamma\|_2 \le M} \sup_{T \in \mathcal{T}} \mathbf{E}_{\hat{\mathbb{Q}}_n} T(X) - \mathbf{E}_{\mathcal{N}(\eta,\Gamma)} f^\star(T(Y)), \tag{22}$$

where $\mathcal{T}$ is the discriminator function class satisfying Assumption 1; f and g satisfy Assumption 2; Γ is a positive definite matrix with spectral norm bounded by M. We show that the estimator $\hat{\theta}_n$ can still achieve the minimax rate.

Theorem 7. *Let $\hat{\theta}_n$ be the estimator defined in (22). Assuming that $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon \le c$ for some sufficiently small constant c, then with probability at least $1 - \delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{\frac{p}{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}.$$

Estimating Gaussian mean and covariance simultaneously via f-GAN has also been studied in Gao et al. (2019b). In their work, the GAN estimator is designed through proper scoring rules, thus their result is only applicable to a subset of f-GANs, while Theorem 7 covers the entire f-GAN family. On the other hand, their result is stronger than Theorem 7 as it, besides mean estimation, also provides guarantee for covariance estimation.

7 Concluding Remarks

We have studied the statistical and robust properties of several popular GAN variants in the setting of Gaussian mean estimation under Huber's contamination model. We showed that f-GAN equipped with an appropriate discriminator can provably achieve the minimax rate, while MMD-GAN and Wasserstein GAN may not. Extensions to the sparse setting and the unknown covariance setting were also discussed. We mention two future directions:

Computational Complexity. In this work, we only focus on the statistical properties. To make these estimators work in practice, we need optimization algorithms that guarantee to find reasonable solutions in polynomial time. In fact, as shown in our experiments, certain types of f-GAN are particularly difficult to train, although they can achieve the minimax optimal rate in theory. It would be interesting to characterize the computational complexity of f-GAN estimators, either by developing polynomial-time algorithm, or by proving hardness results on some f-GAN training.

Minimax Optimal Function Class. We have shown a few positive as well as negative examples of GANs in the contamination setting. It would be interesting to completely characterize minimax optimal function classes, *i.e.* a necessary and sufficient condition on the discriminator class for achieving minimax optimality.

[2]Keeping the top s entries with largest absolute value and setting others to zero.

Acknowledgment

We thank the reviewers for critical comments that improved the final presentation. This work is supported by NSERC, Mitacs, and the Waterloo-Huawei Joint Innovation Lab.

References

P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Annales de l'I.H.P. Probabilités et statistiques*, 47:358–375, 2011.

N. Amenta, M. Bern, D. Eppstein, and S.-H. Teng. Regression depth and center points. *Discrete & Computational Geometry*, 23:305–323, 2000.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223, 2017.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and Equilibrium in Generative Adversarial Nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning*, pages 224–232, 2017.

S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally Efficient Robust Sparse Estimation in High Dimensions. In *Proceedings of the 2017 Conference on Learning Theory*, pages 169–212, 2017.

M. Bikowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Hubers contamination model. *Ann. Statist.*, 46:1932–1960, 2018.

R. Cominetti and J. S. Martín. Asymptotic Analysis of the Exponential Penalty Trajectory in Linear Programming. *Mathematical Programming*, 67:169–187, 1994.

M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

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)*, pages 655–664, 2016.

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*, pages 999–1008, 2017.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart. Sever: A Robust Meta-Algorithm for Stochastic Optimization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1596–1606, 2019a.

I. Diakonikolas, D. Kane, S. Karmalkar, E. Price, and A. Stewart. Outlier-Robust High-Dimensional Sparse Estimation via Iterative Filtering. In *Advances in Neural Information Processing Systems 32*, pages 10689–10700, 2019b.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training Generative Neural Networks via Maximum Mean Discrepancy Optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, page 258267, 2015.

S. Feizi, F. Farnia, T. Ginart, and D. Tse. Understanding GANs: the LQG setting, 2017. arXiv preprint arXiv:1710.10793.

C. Gao, J. Liu, Y. Yao, and W. Zhu. Robust Estimation via Generative Adversarial Networks. In *International Conference on Learning Representations*, 2019a.

C. Gao, Y. Yao, and W. Zhu. Generative Adversarial Nets for Robust Scatter Estimation: A Proper Scoring Rule Perspective, 2019b. arXiv preprint arXiv:1903.01944.

A. Genevay, G. Peyre, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13: 723–773, 2012.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pages 5767–5777, 2017.

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

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic Estimation of Mean and Covariance. In *2016 IEEE*

57th Annual Symposium on Foundations of Computer Science (FOCS), pages 665–674, 2016.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In *Advances in Neural Information Processing Systems 30*, pages 2203–2213, 2017.

Y. Li, K. Swersky, and R. Zemel. Generative Moment Matching Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1718–1727, 2015.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems 29*, pages 271–279, 2016.

G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11:355–607, 2019.

H. Thanh-Tung, T. Tran, and S. Venkatesh. Improving Generalization and Stability of Generative Adversarial Networks. In *International Conference on Learning Representations*, 2019.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

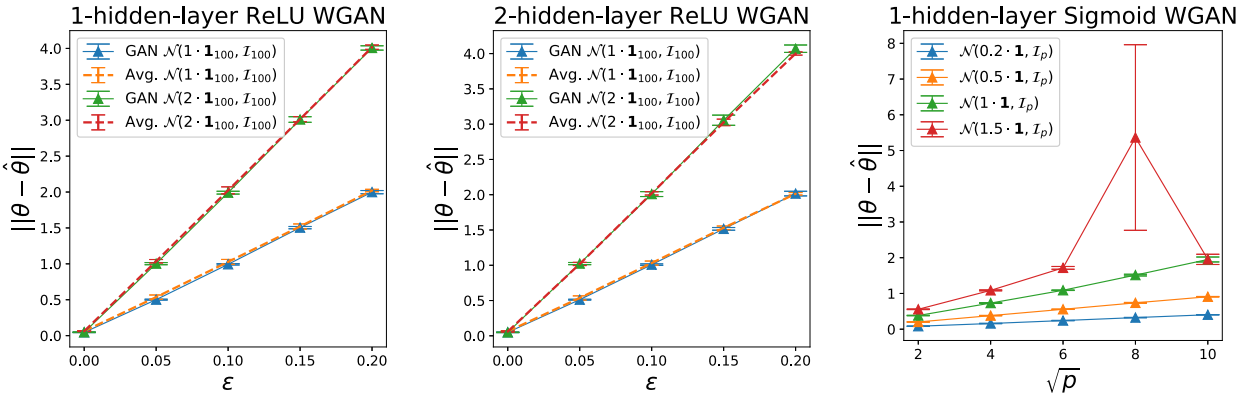
J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, 1975.

J. Weed. An explicit analysis of the entropic penalty in linear programming. In *Proceedings of the 31st Conference On Learning Theory*, pages 1841–1855, 2018.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the Discrimination-Generalization Tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

A Additional Experiments

In this section, we present additional experiments on W-GAN architectures in practice.



(a) error *vs.* ϵ (b) error *vs.* ϵ (c) error *vs.* $\sqrt{p}$

Figure 4: Estimation error of W-GAN, when the discriminator is parametrized by a neural network. The optimization is unstable for the sigmoid network in some cases.

We train three Wasserstein GANs: a one-hidden-layer ReLU network, a two-hidden-layer ReLU network and a one-hidden-layer sigmoid network. We use gradient penalty to enforce the Lipschitz constraint on the discriminator. The results are shown in Figure 4.

As mentioned in Section 5, statistical properties of different subsets of Lipschitz functions may be very different. Here, we also observe the difference for networks with different activation functions. With ReLU activation, the solution of Wasserstein GAN is very close to sample average, whose error is plotted in Figure 4 for comparison. The Wasserstein GAN with sigmoid activation is slightly more robust than that with ReLU network. But still, the estimation error grows as the dimension increases.

B Technical Lemmas

Definition 1. *The f-divergence with a restricted function class $\mathcal{V}$ is defined as*

$$\mathcal{D}_{\mathcal{V}}(\mathbb{P}||\mathbb{Q}) = \sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}} g(V(X)) - \mathbf{E}_{\mathbb{Q}} f^\star(g(V(X))).$$

Lemma 1 (Minimizer of $\mathcal{D}_{\mathcal{V}}$)**.** *Assume f is convex and $f(1) = 0$, f and g satisfy Assumption 2, and the discriminator class $\mathcal{V}$ satisfies Assumption 1. Then, for any distribution $\mathbb{P}$ and $\mathbb{Q}$,*

$$\mathcal{D}_{\mathcal{V}}(\mathbb{P}||\mathbb{Q}) \geq 0.$$

In addition,

$$\mathcal{D}_{\mathcal{V}}(\mathbb{P}||\mathbb{P}) = 0.$$

Proof. Since $f^\star$ is the convex conjugate function of f, we have

$$f^\star(t) + f(x) = xt \Leftrightarrow t \in \partial f(x).$$

In particular, since $f(1) = 0$, we have

$$f^\star(t) = t \Leftrightarrow t \in \partial f(1).$$

According to Assumption 2, $g(0) \in \partial f(1)$, thus

$$f^\star(g(0)) = g(0).$$

For any $\mathbb{P}$ and $\mathbb{Q}$, let the discriminator $V(x)$ be the function $x \mapsto 0$ (by setting all weights to zeros), then

$$\mathbf{E}_{\mathbb{P}} g(V(X)) - \mathbf{E}_{\mathbb{Q}} f^\star(g(V(X))) = 0.$$

Hence, the supremum over $\mathcal{V}$, namely $\mathcal{D}_{\mathcal{V}}$, is nonnegative.

To show $\mathcal{D}_{\mathcal{V}}(\mathbb{P}||\mathbb{P}) = 0$, it is sufficient to show that $\mathcal{D}_{\mathcal{V}}(\mathbb{P}||\mathbb{P}) \le 0$. Notice that for all t, we have

$$\begin{aligned} f^\star(t) &= \sup_x xt - f(x) \\ &\ge 1 \cdot t - f(1) \\ &= t. \end{aligned}$$

Hence,

$$\begin{aligned} \mathcal{D}_{\mathcal{V}}(\mathbb{P}||\mathbb{P}) &= \sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}} g(V(X)) - \mathbf{E}_{\mathbb{P}} f^\star(g(V(X))) \\ &\le \sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}} g(V(X)) - \mathbf{E}_{\mathbb{P}} g(V(X)) \\ &= 0, \end{aligned}$$

which finishes the proof. □

Lemma 2. *For any distribution $\mathbb{P}_1$, $\mathbb{P}_2$ and $\mathbb{P}_3$, we have*

$$|\mathcal{D}_{\mathcal{V}}((1-\epsilon)\mathbb{P}_1 + \epsilon\mathbb{P}_2||\mathbb{P}_3) - \mathcal{D}_{\mathcal{V}}(\mathbb{P}_1||\mathbb{P}_3)| \le 2\kappa\epsilon L_g,$$

where L_g is the Lipschitz constant of g in $[-\kappa, \kappa]$.

Proof. First, notice that $|V(x)| \le \|w\|_1 \le \kappa$. Expand $\mathcal{D}_{\mathcal{V}}$, we have

$$\begin{aligned} |\mathcal{D}_{\mathcal{V}}((1-\epsilon)\mathbb{P}_1 + \epsilon\mathbb{P}_2||\mathbb{P}_3) - \mathcal{D}_{\mathcal{V}}(\mathbb{P}_1||\mathbb{P}_3)| =& \Big| \left(\sup_{V \in \mathcal{V}} \mathbf{E}_{(1-\epsilon)\mathbb{P}_1 + \epsilon\mathbb{P}_2} g(V(X)) - \mathbf{E}_{\mathbb{P}_3} f^\star(g(V(X))) \right) && (23) \\ & - \left(\sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}_1} g(V(X)) - \mathbf{E}_{\mathbb{P}_3} f^\star(g(V(X))) \right) \Big| && (24) \\ \le& \sup_{V \in \mathcal{V}} \left| \mathbf{E}_{(1-\epsilon)\mathbb{P}_1 + \epsilon\mathbb{P}_2} g(V(X)) - \mathbf{E}_{\mathbb{P}_1} g(V(X)) \right| && (25) \\ =& \epsilon \sup_{V \in \mathcal{V}} |\mathbf{E}_{\mathbb{P}_2} g(V(X)) - \mathbf{E}_{\mathbb{P}_1} g(V(X))| && (26) \\ \le& \epsilon \sup_{V \in \mathcal{V}} |\mathbf{E}_{\mathbb{P}_2} [g(V(X)) - g(0)] - \mathbf{E}_{\mathbb{P}_1} [g(V(X)) - g(0)]| && (27) \\ \le& \epsilon \left(\sup_{V \in \mathcal{V}} |\mathbf{E}_{\mathbb{P}_2} [g(V(X)) - g(0)]| + \sup_{V \in \mathcal{V}} |\mathbf{E}_{\mathbb{P}_1} [g(V(X)) - g(0)]| \right) && (28) \\ \le& \epsilon \left(\sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}_2} |g(V(X)) - g(0)| + \sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}_1} |g(V(X)) - g(0)| \right) && (29) \\ \le& \epsilon L_g \left(\sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}_2} |V(X)| + \sup_{V \in \mathcal{V}} \mathbf{E}_{\mathbb{P}_1} |V(X)| \right) && (30) \\ \le& 2\kappa\epsilon L_g, && (31) \end{aligned}$$

where (25) uses the inequality $|\sup f_1 - \sup f_2| \le \sup |f_1 - f_2|$; (30) uses Lipschitz continuity of g on $[-\kappa, \kappa]$ (recall that g is twice continuously differentiable). □

Lemma 3. *Consider the discriminator function class in Assumption 1. For any distribution $\mathbb{P}$, the* i.i.d. *samples $X_1, X_2, \cdots X_n \sim \mathbb{P}$ satisfy*

$$\sup_{V \in \mathcal{V}} \left| \frac{1}{n} \sum_{i=1}^n g(V(X_i)) - \mathbf{E}_{\mathbb{P}} g(V(X_i)) \right| \le C \left(2\kappa L_g \sqrt{\frac{p}{n}} + 2\kappa L_g \sqrt{\frac{\log 1/\delta}{n}} \right),$$

with probability at least $1 - \delta$ for some constant C, where L_g is the Lipschitz constant of g in $[-\kappa, \kappa]$.

Proof. One can first verify the function class $g \circ \mathcal{V}$ satisfies the condition of bounded difference inequality, since

$$|g(x) - g(y)| \le |g(\kappa) - g(-\kappa)| \le 2\kappa L_g,$$

where we use the assumption on g that it is increasing and Lipschitz (since g has continuous second order derivative). The rest of the proof aims for proving the Rademacher complexity of $g \circ \mathcal{V}$ is bounded by $\kappa L_g \sqrt{\frac{p}{n}}$.

Since g is a Lipschitz function on $[-\kappa, \kappa]$, by contraction lemma,

$$\mathfrak{R}(g(\mathcal{V})) \le L_g \mathfrak{R}(\mathcal{V}).$$

In addition, we have

$$\begin{aligned}
\mathfrak{R}(\mathcal{V}) &= \mathbf{E}_\xi \sup_{V \in \mathcal{V}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i V(X_i) \right| \\
&= \mathbf{E}_\xi \sup_{w_j, u_i, b_i} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \sum_{j \ge 1} w_j \sigma(u_j^\top X_i + b_j) \right| \\
&= \mathbf{E}_\xi \sup_{w_j, u_j, b_j} \left| \frac{1}{n} \sum_{j \ge 1} w_j \sum_{i=1}^n \xi_i \sigma(u_j^\top X_i + b_j) \right| \\
&= \kappa \mathbf{E}_\xi \sup_{u, b} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \sigma(u^\top X_i + b) \right| \\
&\lesssim \kappa \sqrt{\frac{p}{n}},
\end{aligned}$$

where ξ_i are independent Rademacher random variables. We use Cauchy inequality in the second last step and the last inequality is because the Rademacher complexity of $\left\{\sigma(u^\top x + b) : u \in \mathbb{R}^p, b \in \mathbb{R}\right\}$ is $O(\sqrt{\frac{p}{n}})$ (Gao et al., 2019a). □

Lemma 4. *Suppose $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}$ is the unit ball in the RKHS induced by a kernel $k(\cdot, \cdot)$ satisfying $\sup_x k(x, x) \le 1$ (*e.g. *a Gaussian kernel). For any distribution $\mathbb{P}$, the* i.i.d. *samples $X_1, X_2, \cdots X_n \sim \mathbb{P}$ satisfy*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbf{E}_{\mathbb{P}} f(X) \right| \le \frac{2}{\sqrt{n}} + 2\sqrt{\frac{\log 2/\delta}{2n}}$$

with probability at least $1 - \delta$.

Proof. It is well known that the Rademacher complexity of $\mathcal{F}$ is upper bounded by $\frac{1}{\sqrt{n}}$. By standard concentration inequality we can obtain the above result. □

Lemma 5. *Consider the function class $\mathcal{V}$ defined in* (20). *For any distribution $\mathbb{P}$, the* i.i.d. *samples $X_1, X_2, \cdots X_n \sim \mathbb{P}$ satisfy*

$$\sup_{V \in \mathcal{V}} \left| \frac{1}{n} \sum_{i=1}^n g(V(X_i)) - \mathbf{E}_{\mathbb{P}} g(V(x)) \right| \le C\kappa L_g \left(\sqrt{\frac{s \log \frac{ep}{s}}{n}} + \sqrt{\frac{\log 1/\delta}{n}} \right)$$

with probability at least $1 - \delta$, where C is an absolute constant.

Proof. The proof follows the similar steps of Lemma 3, except that in the last step we have a better bound on the function class $\mathcal{F} = \left\{\sigma(u^\top x + b) : u \in \mathbb{R}^p, \|u\|_0 \le 2s, b \in \mathbb{R}\right\}$.

We decompose $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \cdots \cup \mathcal{F}_{\binom{p}{2s}}$, where each $\mathcal{F}_j$ denotes a subset of $\mathcal{F}$ with distinct sparsity pattern. It is not hard to see that each $\mathcal{F}_j$ has Rademacher complexity $\sqrt{\frac{2s}{n}}$. Thus for each fixed $\mathcal{F}_j$, we can use Rademacher

complexity to prove

$$\sup_{f\in\mathcal{F}_j}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbf{E}_{\mathbb{P}}f(X)\right|\le C\left(\sqrt{\frac{s}{n}}+\sqrt{\frac{\log\binom{p}{2s}/\delta}{n}}\right)$$

holds with probability at least $1-\delta/\binom{p}{2s}$. Using union bound over all $\mathcal{F}_j$, with probability at least $1-\delta$, we have

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbf{E}_{\mathbb{P}}f(X)\right|&\le C\left(\sqrt{\frac{s}{n}}+\sqrt{\frac{\log\binom{p}{2s}/\delta}{n}}\right)\\&\le C\left(\sqrt{\frac{s}{n}}+\sqrt{\frac{\log\left(\frac{ep}{s}\right)^{2s}/\delta}{n}}\right)\\&\le C\left(\sqrt{\frac{s}{n}}+\sqrt{\frac{2s\log\frac{ep}{s}+\log 1/\delta}{n}}\right)\\&\le C'\left(\sqrt{\frac{2s\log\frac{ep}{s}}{n}}+\sqrt{\frac{\log 1/\delta}{n}}\right),\end{aligned}$$

which finishes the proof. □

Lemma 6. *Let Φ be the CDF of the standard Gaussian distribution. For any $\eta\in\mathbb{R}$, there uniquely exists a τ, such that*

$$\Phi(\tau-\eta)=(1-\epsilon)\Phi(\tau).$$

Moreover, $(\tau(\eta)-\eta)\left(\eta-\Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)\right)>0$.

Proof. On the one hand,

$$\lim_{t\to+\infty}\frac{\Phi(t-\eta)}{\Phi(t)}=\frac{1}{1-\epsilon}>1.$$

On the other hand,

$$\lim_{t\to-\infty}\frac{\Phi(t-\eta)}{\Phi(t)}=\lim_{t\to-\infty}\frac{\phi(t-\eta)}{\phi(t)}=\lim_{t\to-\infty}\exp\left(\frac{1}{2}\eta(2t-\eta)\right)=0.$$

Since both $\Phi(t-\eta)$ and $\Phi(t)$ are continuous, τ exists. Denote $t_0=\frac{1}{\eta}\log(1-\epsilon)+\frac{1}{2}\eta$. It is easy to check that $\tau\in(t_0,+\infty)$, in which the function $\Phi(t-\eta)-(1-\epsilon)\Phi(t)$ is monotonic. Thus τ is unique.

Since τ uniquely exists for every η, $\tau(\eta)$ is a function of η. Now we characterize the properties of $\tau(\eta)$.

Differentiate w.r.t. η on both sides of

$$\Phi(\tau-\eta)=(1-\epsilon)\Phi(\tau),$$

we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\eta}(\tau(\eta)-\eta)=&\frac{\phi(\tau-\eta)}{\phi(\tau-\eta)-(1-\epsilon)\phi(\tau)}-1\\=&\frac{(1-\epsilon)\phi(\eta)}{\phi(\tau-\eta)-(1-\epsilon)\phi(\tau)},\end{aligned}$$

where ϕ is the density of the standard Gaussian distribution. It can be verified that the denominator is strictly positive. Thus $\tau(\eta)-\eta$ is an increasing function w.r.t. η.

One can verify that

$$\tau = \eta = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$$

satisfies $\Phi(\tau - \eta) = (1-\epsilon)\Phi(\tau)$, hence a root of $\tau(\eta) - \eta = 0$. Since $\tau(\eta) - \eta$ is increasing, the root is unique, which concludes the proof.

□

C f-GAN

Theorem 1. *Let $\hat{\theta}$ be the estimator defined in (10), where f and g satisfy Assumption 2 and $\mathcal{V}$ satisfies Assumption 1. Assuming that $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon \leq c$ for some sufficiently small constant c, then with probability at least $1-\delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{\frac{p}{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}. \tag{11}$$

Proof. We start with bounding the distance between $\mathcal{N}(\theta, I_p)$ and $\mathcal{N}(\hat{\theta}, I_p)$ in terms of $\mathcal{D}_{\mathcal{V}}$. With probability at least $1-2\delta$, we have

$$\begin{aligned}
\mathcal{D}_{\mathcal{V}}(\mathcal{N}(\theta, I_p)||\mathcal{N}(\hat{\theta}, I_p)) \leq & \mathcal{D}_{\mathcal{V}}((1-\epsilon)\mathcal{N}(\theta, I_p) + \epsilon\mathbb{H}||\mathcal{N}(\hat{\theta}, I_p)) + 2\kappa\epsilon L_g && (32)\\
\leq & \mathcal{D}_{\mathcal{V}}(\hat{\mathbb{Q}}_n||\mathcal{N}(\hat{\theta}, I_p)) + 2\kappa\epsilon L_g + 2\kappa L_g\sqrt{\frac{p}{n}} + 2\kappa L_g\sqrt{\frac{\log 1/\delta}{n}} && (33)\\
\leq & \mathcal{D}_{\mathcal{V}}(\hat{\mathbb{Q}}_n||\mathcal{N}(\theta, I_p)) + 2\kappa\epsilon L_g + 2\kappa L_g\sqrt{\frac{p}{n}} + 2\kappa L_g\sqrt{\frac{\log 1/\delta}{n}} && (34)\\
\leq & \mathcal{D}_{\mathcal{V}}((1-\epsilon)\mathcal{N}(\theta, I_p) + \epsilon\mathbb{H}||\mathcal{N}(\theta, I_p)) + 2\kappa\epsilon L_g + 4\kappa L_g\sqrt{\frac{p}{n}} + 4\kappa L_g\sqrt{\frac{\log 1/\delta}{n}} && (35)\\
\leq & \mathcal{D}_{\mathcal{V}}(\mathcal{N}(\theta, I_p)||\mathcal{N}(\theta, I_p)) + 4\kappa\epsilon L_g + 4\kappa L_g\sqrt{\frac{p}{n}} + 4\kappa L_g\sqrt{\frac{\log 1/\delta}{n}} && (36)\\
\leq & 4\kappa\epsilon L_g + 4\kappa L_g\sqrt{\frac{p}{n}} + 4\kappa L_g\sqrt{\frac{\log 1/\delta}{n}}, && (37)
\end{aligned}$$

where (32) and (36) use Lemma 2; (33) and (35) use Lemma 3; (34) follows by the fact that $\hat{\theta}$ minimizes $\mathcal{D}_{\mathcal{V}}$; (37) follows from Lemma 1. The bound holds for the supremum over $\mathcal{V}$. In particular, it holds for any $V \in \mathcal{V}$. Pick $w_1 = \kappa$, $u_1 = u$ with $\|u\| = 1$ and $b_1 = -u^\top\hat{\theta}$, and let

$$\psi_\xi(t) = \mathbf{E}_{z\sim\mathcal{N}(0,1)}\left[g\left(t\sigma(z+\xi)\right) - f^\star(g(t\sigma(z)))\right],$$

then

$$\psi_{u^\top(\hat{\theta}-\theta)}(\kappa) \lesssim 4\kappa\epsilon L_g + 4\kappa L_g\sqrt{\frac{p}{n}} + 4\kappa L_g\sqrt{\frac{\log 1/\delta}{n}}$$

holds for every u and κ with probability at least $1-2\delta$. Since g and $f^\star$ are twice continuously differentiable, ψ'' is continuous in $[0,\kappa]$ and $|\psi''|$ can be bounded by some constant $M(\kappa)$. A key observation is that $\psi_\xi(t) + M(\kappa)t^2$ is convex in $[0,\kappa]$. Thus, by subgradient inequality,

$$\psi_\xi(\kappa) + M(\kappa)\kappa^2 \geq \kappa\psi'_\xi(0),$$

where we recall $\psi_\xi(0) = 0$ since $g(0) = f^\star(g(0))$. This is because by Frechel inequality

$$f(x) + f^\star(y) = xy \Leftrightarrow y \in \partial f(x)$$

and by Assumption 2 $f(1) = 0$ and $g(0) \in \partial f(1)$.

We have

$$\psi'_\xi(0) = g'(0)\left(h(\xi) - h(0)\right),$$

where

$$h(\xi) = \mathbf{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z+\xi)].$$

Since h is increasing and $h'(0)$ is strictly positive, there exist constants $c > 0$ and $c' > 0$, such that any ξ satisfying $|h(\xi) - h(0)| < c'$ has $|h(\xi) - h(0)| \geq c\xi$.

Thus

$$\begin{aligned}
\|\hat{\theta} - \theta\| &= \sup_{\|u\|=1} u^\top(\theta - \hat{\theta}) \\
&\leq \sup_{\|u\|=1} \frac{1}{c}\left(h(u^\top(\theta-\hat{\theta})) - h(0)\right) \\
&\leq \sup_{\|u\|=1} \frac{1}{c\cdot g'(0)} \cdot \psi'_{u^\top(\theta-\hat{\theta})}(0) \\
&\leq \sup_{\|u\|=1} \frac{1}{c\cdot g'(0)}\left(\psi_{u^\top(\theta-\hat{\theta})}(\kappa) + M(\kappa)\kappa^2\right)/\kappa \\
&\leq \frac{1}{c\cdot g'(0)}\left(4\epsilon L_g + 4L_g\sqrt{\frac{p}{n}} + 4L_g\sqrt{\frac{\log 1/\delta}{n}} + M(\kappa)\kappa\right),
\end{aligned}$$

where the first inequality holds when $\sqrt{\frac{p}{n}} + \epsilon$ is sufficiently small such that $\left|h(u^\top(\theta-\hat{\theta})) - h(0)\right| \leq c'$. Note that $\lim_{\kappa\to 0} M(\kappa)\kappa = 0$, since $M(\kappa)$ is monotonically decreasing w.r.t. κ. We can pick κ sufficiently small such that

$$M(\kappa)\kappa \leq 4\epsilon L_g + 4L_g\sqrt{\frac{p}{n}}.$$

Thus

$$\|\hat{\theta} - \theta\| \lesssim \epsilon + \sqrt{\frac{p}{n}} + \sqrt{\frac{\log 1/\delta}{n}}$$

holds with probability at least $1-\delta$. □

D MMD GAN

Theorem 2. *Let $\mathcal{T}$ be the RKHS unit ball induced by the Gaussian kernel with bandwidth σ. For the estimator defined in* (14)*, with probability at least $1-\delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim (2+\sigma^2)^{\frac{1}{2}}(1+\frac{2}{\sigma^2})^{\frac{p}{4}}\left(\frac{1}{\sqrt{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}\right).$$

Proof. First, since every $f \in \mathcal{T}$ has bounded range:

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|k(\cdot,x)\|_{\mathcal{H}} = \sqrt{k(x,x)} \leq 1,$$

we can show that the contamination can only change the MMD distance by a constant factor of ϵ:

$$\begin{aligned}
\mathrm{MMD}\left[(1-\epsilon)\mathbb{P}_\theta + \epsilon\mathbb{H}, \mathbb{P}\right] &= \sup_{f\in\mathcal{T}} \mathbf{E}_{(1-\epsilon)\mathbb{P}_\theta+\epsilon\mathbb{H}} f(X) - \mathbf{E}_{\mathbb{P}} f(X) \\
&\leq \sup_{f\in\mathcal{T}} \mathbf{E}_{\mathbb{P}_\theta} f(X) - \mathbf{E}_{\mathbb{P}} f(X) + \epsilon\mathbf{E}_{\mathbb{H}} f(X) - \epsilon\mathbf{E}_{\mathbb{P}_\theta} f(X) \\
&\leq \sup_{f\in\mathcal{T}} \left(\mathbf{E}_{\mathbb{P}_\theta} f(X) - \mathbf{E}_{\mathbb{P}} f(X)\right) + \epsilon \sup_{f\in\mathcal{T}} \left(\mathbf{E}_{\mathbb{H}} f(X) - \mathbf{E}_{\mathbb{P}_\theta} f(X)\right) \\
&\leq \mathrm{MMD}\left[\mathbb{P}_\theta, \mathbb{P}\right] + 2\epsilon,
\end{aligned}$$

where $\mathbb{P}$ is an arbitrary distribution. The reverse direction also holds by a similar argument. Follow the similar steps in Theorem 1, using Lemma 4, we can show

$$\sup_{f\in\mathcal{T}} \mathbf{E}_{\mathbb{P}_\theta} f(X) - \mathbf{E}_{\mathbb{P}_{\hat{\theta}}} f(X) \le 2\epsilon + \frac{4}{\sqrt{n}} + 4\sqrt{\frac{\log(2/\delta)}{2n}},$$

holds with probability at least $1-\delta$. Recall that the MMD between two distributions is the distance of the mean embedding in a RKHS (Gretton et al., 2012)

$$\sup_{f\in\mathcal{T}} \mathbf{E}_{\mathbb{P}_\theta} f(X) - \mathbf{E}_{\mathbb{P}_{\hat{\theta}}} f(X) = \|\mu_{\mathbb{P}_\theta} - \mu_{\mathbb{P}_{\hat{\theta}}}\|_{\mathcal{H}}.$$

When $\mathbb{P}_\theta$ and $\mathbb{P}_{\hat{\theta}}$ are both Gaussian distributions, the right hand side can be computed in a closed form:

$$\begin{aligned}
\|\mu_{\mathbb{P}_\theta} - \mu_{\mathbb{P}_{\hat{\theta}}}\|_{\mathcal{H}}^2 &= \mathbf{E}_{x,x'\sim\mathbb{P}_\theta}[k(x,x')] - 2\mathbf{E}_{x\sim\mathbb{P}_\theta,x'\sim\mathbb{P}_{\hat{\theta}}}[k(x,x')] + \mathbf{E}_{x,x'\sim\mathbb{P}_{\hat{\theta}}}[k(x,x')] \\
&= 2\mathbf{E}_{x\sim N(0,2I_p)}\left[\exp\left(-\frac{x^Tx}{2\sigma^2}\right)\right] - 2\mathbf{E}_{x\sim N(\theta-\hat{\theta},2I_p)}\left[\exp\left(-\frac{x^Tx}{2\sigma^2}\right)\right] \\
&= \sqrt{\left(\frac{\sigma^2}{2+\sigma^2}\right)^p}\left(1-\exp\left(-\frac{1}{2(2+\sigma^2)}\|\hat{\theta}-\theta\|^2\right)\right).
\end{aligned}$$

Assuming that $\frac{1}{n}$ and ϵ are sufficiently small thus $\|\mu_{\mathbb{P}_\theta} - \mu_{\mathbb{P}_{\hat{\theta}}}\|_{\mathcal{H}}$ is sufficiently small, such that

$$1-\exp\left(-\frac{1}{2(2+\sigma^2)}\|\hat{\theta}-\theta\|^2\right) \le \frac{1}{2},$$

then by the inequality $\frac{1}{2}x \le 1-\exp(-x)$,

$$\frac{1}{2}\cdot\frac{\|\hat{\theta}-\theta\|^2}{2(2+\sigma^2)} \le 1-\exp\left(-\frac{1}{2(2+\sigma^2)}\|\hat{\theta}-\theta\|^2\right).$$

Combining all of the above, we have proven that

$$\begin{aligned}
\|\hat{\theta}-\theta\| &\le 2\sqrt{2+\sigma^2}\sqrt{1-\exp\left(-\frac{1}{2(2+\sigma^2)}\|\hat{\theta}-\theta\|^2\right)} \\
&\le 2\sqrt{2+\sigma^2}\left(1+\frac{2}{\sigma^2}\right)^{\frac{p}{4}}\|\mu_{\mathbb{P}_{\hat{\theta}}} - \mu_{\mathbb{P}_\theta}\|_{\mathcal{H}} \\
&\le 2\sqrt{2+\sigma^2}\left(1+\frac{2}{\sigma^2}\right)^{\frac{p}{4}}\left(2\epsilon + \frac{4}{\sqrt{n}} + 4\sqrt{\frac{\log(2/\delta)}{2n}}\right),
\end{aligned}$$

holds with probability at least $1-\delta$. □

Corollary 1. *Let $\mathcal{F}$ be the RKHS unit ball induced by the Gaussian kernel with bandwidth $\sigma = \sqrt{p}$, then with probability at least $1-\delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{p}\left(\frac{1}{\sqrt{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}\right). \tag{15}$$

Proof. We optimize the bound in Theorem 2 by choosing appropriate bandwidth σ according to the dimension p. Consider the coefficient $(2+\sigma^2)(1+\frac{2}{\sigma^2})^{\frac{p}{2}}$ in Theorem 2. It achieves its minimum value at $\sigma = \sqrt{p}$, which turns out to be $(2+p)(1+\frac{2}{p})^{\frac{p}{2}} \le 2ep \lesssim p$. Plugging in the choice of σ finishes the proof. □

Theorem 3. *Consider the population limit of $\hat{\theta}$ given by MMD-GAN. For any $\sigma > 0$, there always exists a contaminated distribution $\mathbb{Q}$ such that*

$$\|\hat{\theta}-\theta\| \gtrsim \sqrt{p}\epsilon. \tag{16}$$

Proof. Consider a Dirac contamination $\mathbb{H} = \delta_{\tilde{\theta}}$.

$$\hat{\theta} = \underset{\eta \in \mathbb{R}^p}{\text{minimize}}\, \text{MMD}^2[(1-\epsilon)\mathbb{P}_\theta + \epsilon\delta_{\tilde{\theta}}, \mathbb{P}_\eta]. \tag{38}$$

Since MMD between mixture of Gaussian has a closed form solution, it is easy to show that (38) is equivalent to

$$\underset{\eta \in \mathbb{R}^p}{\text{minimize}} -(1-\epsilon)\exp\left(-\frac{\|\theta-\eta\|^2}{2(2+\sigma^2)}\right) - \epsilon\left(\frac{2+\sigma^2}{1+\sigma^2}\right)^{\frac{p}{2}}\exp\left(-\frac{\|\tilde{\theta}-\eta\|^2}{2(1+\sigma^2)}\right).$$

Although we have a closed form solution for MMD, the objective function is still nonconvex w.r.t. η. However, a key observation is that the global minimizer must lie in the line segment between θ and $\tilde{\theta}$. If not, a projection onto this line segment has strictly smaller objective value. This observation allows us to parametrize $\eta = \theta + t(\tilde{\theta}-\theta)$, where $0 \le t \le 1$.

$$\underset{0 \le t \le 1}{\text{minimize}} -(1-\epsilon)\exp\left(-\frac{\|\theta-\tilde{\theta}\|^2}{2(2+\sigma^2)}t^2\right) - \epsilon\left(\frac{2+\sigma^2}{1+\sigma^2}\right)^{\frac{p}{2}}\exp\left(-\frac{\|\theta-\tilde{\theta}\|^2}{2(1+\sigma^2)}(t-1)^2\right).$$

We first prove the following claim.

Claim: for any $\sigma > 0$, as long as $\|\theta-\tilde{\theta}\|^2 = p(1+\sigma^2)\log\frac{2+\sigma^2}{1+\sigma^2}$, then $t^\star \ge \epsilon$.

If the claim holds, then

$$\begin{aligned}
\|\hat{\theta}-\theta\| &= \|\eta^\star - \theta\| \\
&= t^\star\|\theta-\tilde{\theta}\| \\
&\ge \epsilon\sqrt{p(1+\sigma^2)\log\frac{2+\sigma^2}{1+\sigma^2}} \\
&\ge \epsilon\sqrt{p\log 2},
\end{aligned}$$

where the last inequality is because $(1+\sigma^2)\log\frac{2+\sigma^2}{1+\sigma^2} \ge \log 2$, which finishes the proof. The rest of the proof is dedicated to proving the claim.

It is sufficient to prove the gradient w.r.t. t is negative in $[0,\epsilon]$, which is equivalent to prove

$$(1-\epsilon)\exp\left(-\frac{\|\theta-\tilde{\theta}\|^2}{2(2+\sigma^2)}t^2\right)\frac{\|\theta-\tilde{\theta}\|^2}{2+\sigma^2}t \le \epsilon\left(\frac{2+\sigma^2}{1+\sigma^2}\right)^{\frac{p}{2}}\exp\left(-\frac{\|\theta-\tilde{\theta}\|^2}{2(1+\sigma^2)}(t-1)^2\right)\frac{\|\theta-\tilde{\theta}\|^2}{1+\sigma^2}(1-t)$$

holds for any $t \le \epsilon$. Taking logarithm on both sides, it is equivalent to show

$$\log\frac{\epsilon}{1-\epsilon} + \log\frac{1-t}{t} + \left(\frac{p}{2}+1\right)\log\frac{2+\sigma^2}{1+\sigma^2} + \frac{\|\theta-\tilde{\theta}\|^2}{2(2+\sigma^2)}t^2 - \frac{\|\theta-\tilde{\theta}\|^2}{2(1+\sigma^2)}(t-1)^2 \ge 0 \tag{39}$$

for any $0 \le t \le \epsilon$. It is easy to see that for $t \le \epsilon$, we have

$$\log\frac{\epsilon}{1-\epsilon} + \log\frac{1-t}{t} \ge 0.$$

Further,

$$\left(\frac{p}{2}+1\right)\log\frac{2+\sigma^2}{1+\sigma^2} + \frac{\|\theta-\tilde{\theta}\|^2}{2(2+\sigma^2)}t^2 - \frac{\|\theta-\tilde{\theta}\|^2}{2(1+\sigma^2)}(t-1)^2$$

is a quadratic function w.r.t. t, and is monotonic increasing when $0 \le t \le 1$. Thus its minimum value is achieved at $t = 0$, which is

$$\begin{aligned}
\left(\frac{p}{2}+1\right)\log\frac{2+\sigma^2}{1+\sigma^2} - \frac{\|\theta-\tilde{\theta}\|^2}{2(1+\sigma^2)} &= \left(\frac{p}{2}+1\right)\log\frac{2+\sigma^2}{1+\sigma^2} - \frac{p}{2}\log\frac{2+\sigma^2}{1+\sigma^2} \\
&\ge 0,
\end{aligned}$$

where the first inequality is because the specific choice of $\tilde{\theta}$ in the claim. Thus the left hand side of (39) is positive, which finishes the proof. □

E Wasserstein GAN

Theorem 4. *Consider W-GAN with $p = 1$. Let the contamination distribution $\mathbb{H} = \delta_{\tilde{\theta}}$. Suppose ϵ is sufficiently small, then $|\theta - \hat{\theta}| \lesssim \epsilon$. Further, there exists a contamination distribution such that $|\theta - \hat{\theta}| \gtrsim \epsilon$.*

Proof. Without loss of generality, we assume that $\theta = 0$ and $\tilde{\theta} > 0$. Recall that the Wasserstein distance with Euclidean distance as ground cost in one dimension has a closed-form expression (Peyré et al., 2019) as follows:

$$\underset{\eta \in \mathbb{R}}{\text{minimize}} \int_{-\infty}^{+\infty} \left| \Phi(t-\eta) - (1-\epsilon)\Phi(t) - \epsilon \mathbf{1}_{t \geq \tilde{\theta}} \right| \mathrm{d}t, \tag{40}$$

where Φ is the CDF of the standard Gaussian distribution. It is clear that the minimizer $\eta^\star \geq 0$.

Let L be the objective in (40) and let $\eta_0 = \Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)$. We show that if $\eta > \eta_0$ then $\frac{\mathrm{d}L}{\mathrm{d}\eta} > 0$, hence the solution $\eta^\star \leq \eta_0$. By Lemma 6, if $\eta > \eta_0$, then $\tau(\eta) > \eta$, where $\tau(\eta)$ (uniquely) satisfies $\Phi(\tau - \eta) = (1-\epsilon)\Phi(\tau)$. Given a fixed $\eta > \eta_0$, we discuss two cases.

Case 1: $\tilde{\theta} \leq \tau(\eta)$

Decompose (40) into two terms:

$$\begin{aligned} L &= \int_{-\infty}^{\tilde{\theta}} + \int_{\tilde{\theta}}^{+\infty} \left| \Phi(t-\eta) - (1-\epsilon)\Phi(t) - \epsilon \mathbf{1}_{t \geq \tilde{\theta}} \right| \mathrm{d}t \\ &= \int_{-\infty}^{\tilde{\theta}} \left(-\Phi(t-\eta) + (1-\epsilon)\Phi(t) \right) \mathrm{d}t + \int_{\tilde{\theta}}^{+\infty} \left(-\Phi(t-\eta) + (1-\epsilon)\Phi(t) + \epsilon \right) \mathrm{d}t. \end{aligned}$$

Taking the derivative of the objective function w.r.t. η, we get

$$\frac{\mathrm{d}L}{\mathrm{d}\eta} = \int_{-\infty}^{\tilde{\theta}} \phi(t-\eta)\,\mathrm{d}t + \int_{\tilde{\theta}}^{+\infty} \phi(t-\eta)\,\mathrm{d}t > 0,$$

where ϕ is the density of the standard Gaussian distribution.

Case 2: $\tilde{\theta} \geq \tau(\eta)$

Decompose (40) into three terms:

$$\begin{aligned} L &= \int_{-\infty}^{\tau(\eta)} + \int_{\tau(\eta)}^{\tilde{\theta}} + \int_{\tilde{\theta}}^{+\infty} \left| \Phi(t-\eta) - (1-\epsilon)\Phi(t) - \epsilon \mathbf{1}_{t \geq \tilde{\theta}} \right| \mathrm{d}t \\ &= \int_{-\infty}^{\tau} -\Phi(t-\eta) + (1-\epsilon)\Phi(t)\,\mathrm{d}t + \int_{\tau}^{\tilde{\theta}} \Phi(t-\eta) - (1-\epsilon)\Phi(t)\,\mathrm{d}t + \int_{\tilde{\theta}}^{+\infty} -\Phi(t-\eta) + (1-\epsilon)\Phi(t) + \epsilon\,\mathrm{d}t. \end{aligned}$$

Taking the derivative of the objective function w.r.t. η, we get

$$\begin{aligned} \frac{\mathrm{d}L}{\mathrm{d}\eta} &= \int_{-\infty}^{\tau(\eta)} \phi(t-\eta)\,\mathrm{d}t - \int_{\tau(\eta)}^{\tilde{\theta}} \phi(t-\eta)\,\mathrm{d}t + \int_{\tilde{\theta}}^{+\infty} \phi(t-\eta)\,\mathrm{d}t \\ &> \int_{-\infty}^{\tau(\eta)-\eta} \phi(t)\,\mathrm{d}t - \int_{\tau(\eta)-\eta}^{+\infty} \phi(t)\,\mathrm{d}t \\ &> 0, \end{aligned}$$

where we recall that $\tau(\eta) - \eta > 0$.

To sum up, in both cases $\frac{\mathrm{d}L}{\mathrm{d}\eta}$ is positive, thus any $\eta > \eta_0$ cannot be the solution to (40). Lastly, we roughly estimate η_0.

$$\lim_{\epsilon \to 0} \frac{\eta_0}{\epsilon} = \lim_{\epsilon \to 0} \frac{\Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)}{\epsilon} = \lim_{\epsilon \to 0} \frac{1}{\phi\left(\Phi^{-1}\left(\frac{1}{2(1-\epsilon)}\right)\right)} \cdot \frac{1}{2} = \sqrt{\frac{\pi}{2}}.$$

Therefore, when ϵ is sufficiently small, $\eta^\star$ behaves like a linear function of ϵ, *i.e.* $|\hat{\theta} - \theta| \le \eta_0 \lesssim \epsilon$.

For the lower bound, consider a contamination $\delta_{\tilde{\theta}}$ with $\tilde{\theta} \to +\infty$. We prove that any $\eta < \eta_0$, cannot be the solution either. Decompose L into three terms:

$$L = \int_{-\infty}^{\tau(\eta)} + \int_{\tau(\eta)}^{\tilde{\theta}} + \int_{\tilde{\theta}}^{+\infty} \left| \Phi(t-\eta) - (1-\epsilon)\Phi(t) - \epsilon \mathbf{1}_{t \ge \tilde{\theta}} \right| \mathrm{d}t.$$

Taking the derivative of the objective function w.r.t. η, we get

$$\begin{aligned}
\frac{\mathrm{d}L}{\mathrm{d}\eta} &= \int_{-\infty}^{\tau(\eta)} \phi(t-\eta)\,\mathrm{d}t - \int_{\tau(\eta)}^{\tilde{\theta}} \phi(t-\eta)\,\mathrm{d}t + \int_{\tilde{\theta}}^{+\infty} \phi(t-\eta)\,\mathrm{d}t \\
&= \int_{-\infty}^{\tau(\eta)-\eta} \phi(t)\,\mathrm{d}t - \int_{\tau(\eta)-\eta}^{0} \phi(t)\,\mathrm{d}t - \int_{0}^{\tilde{\theta}-\eta} \phi(t)\,\mathrm{d}t + \int_{\tilde{\theta}-\eta}^{+\infty} \phi(t-\eta)\,\mathrm{d}t.
\end{aligned}$$

As $\tilde{\theta}$ goes to infinity, the forth term goes to zero, and the third term will become larger than the first term (recall that $\tau(\eta) - \eta < 0$ since $\eta < \eta_0$). Thus

$$\lim_{\tilde{\theta} \to +\infty} \frac{\mathrm{d}L}{\mathrm{d}\eta} = -\int_{\tau(\eta)-\eta}^{0} \phi(t)\,\mathrm{d}t < 0,$$

which indicates that any $\eta < \eta_0$ cannot be the solution, *i.e.* $|\hat{\theta} - \theta| \ge \eta_0 \gtrsim \epsilon$.

□

F Adaptation

Theorem 5. *Assuming that $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon \le c$ for some sufficiently small constant c, with probability at least $1-\delta$, the estimator defined in* (20) *satisfies*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{\frac{s \log \frac{ep}{s}}{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}. \tag{21}$$

Proof. The proof follows the same idea in the proof of Theorem 1. The only difference is that in the sparse setting, we can use Lemma 5 to get a better sample complexity.

First, by *Lemma* 5 and following similar steps to the proof of Theorem 1, we can show that

$$\mathcal{D}_{\mathcal{V}}(\mathcal{N}(\theta, I_p), \mathcal{N}(\hat{\theta}, I_p) \lesssim \kappa\epsilon L_g + \kappa L_g \sqrt{\frac{s \log \frac{ep}{s}}{n}} + \kappa L_g \sqrt{\frac{\log 1/\delta}{n}}$$

holds with probability at least $1-\delta$. Next, we can prove the following improved bound of the Euclidean distance in a similar way to Theorem 1:

$$\begin{aligned}
\|\hat{\theta} - \theta\| &\le \sup_{\|u\|_0 \le 2s} \left| u^T \left(\theta - \hat{\theta} \right) \right| \\
&\le \mathcal{D}_{\mathcal{V}}(\mathcal{N}(\theta, I_p), \mathcal{N}(\hat{\theta}, I_p) \\
&\lesssim \epsilon + \sqrt{\frac{s \log \frac{ep}{s}}{n}} + \sqrt{\frac{\log 1/\delta}{n}},
\end{aligned}$$

whenever κ and $\epsilon + \sqrt{\frac{s \log \frac{ep}{s}}{n}}$ is sufficiently small, which finishes the proof. □

Theorem 6. *Let $\Theta = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le s\}$ and $\mathbb{P}_\theta = \mathcal{N}(\theta, I_p)$. There exist absolute constants c_1 and c_2, such that for any estimator $\hat{\theta}$,*

$$\sup_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{Q}_\theta} \mathbb{Q}\left(\|\hat{\theta} - \theta\| \ge c_1 \left(\sqrt{\frac{s \log ep/s}{n}} \vee \epsilon \right) \right) \ge c_2.$$

Proof. When $\epsilon = 0$, it is well known that there exist absolute constants c_1 and c_2 such that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta \left(\|\hat{\theta} - \theta\|^2 \geq c_1 \cdot \frac{s \log ep/s}{n} \right) \geq c_2.$$

In addition, the modulus of continuity for sparse Gaussian mean estimation is

$$\begin{aligned} \omega(\epsilon, \Theta) &= \sup \left\{ \|\theta_1 - \theta_2\|^2 : \mathrm{TV}(\mathcal{N}(\theta_1, I_p), \mathcal{N}(\theta_2, I_p)) \leq \frac{\epsilon}{1-\epsilon}, \theta_1, \theta_2 \in \Theta \right\} \\ &\gtrsim \epsilon^2. \end{aligned}$$

Thus, by Chen et al. (2018, Theorem 5.1)

$$\begin{aligned} \inf_{\hat{\theta}} \sup_{\theta \in \Theta, \mathbb{Q}} \mathbf{E}_{(1-\epsilon)\mathbb{P}_\theta + \epsilon \mathbb{Q}} \|\hat{\theta} - \theta\|^2 &\gtrsim \frac{s \log ep/s}{n} \vee \omega(\epsilon, \Theta) \\ &\gtrsim \frac{s \log ep/s}{n} \vee \epsilon^2 \end{aligned}$$

□

Theorem 7. *Let $\hat{\theta}_n$ be the estimator defined in (22). Assuming that $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon \leq c$ for some sufficiently small constant c, then with probability at least $1 - \delta$,*

$$\|\hat{\theta}_n - \theta\| \lesssim \sqrt{\frac{p}{n}} \vee \epsilon + \sqrt{\frac{\log 1/\delta}{n}}.$$

Proof. Follow in the similar argument as the proof of Theorem 1, we can show that

$$\sup_{V \in \mathcal{V}} \mathbf{E}_{\mathcal{N}(\theta, \Sigma)} g(V(X)) - \mathbf{E}_{\mathcal{N}(\hat{\theta}, \hat{\Sigma})} f^\star(g(V(X))) \lesssim 4\kappa\epsilon L_g + 4\kappa L_g \sqrt{\frac{p}{n}} + 4\kappa L_g \sqrt{\frac{\log 1/\delta}{n}}$$

holds with probability at least $1 - 2\delta$. Pick $w_1 = \kappa$, $w_j = 0$ for $j > 1$, $u_1 = \frac{u}{\sqrt{u^\top \Sigma u}}$, where $\|u\| = 1$, and $b_1 = -u_1^\top \hat{\theta}$. We have

$$\begin{aligned} &\sup_{\|u\|=1} \mathbf{E}_{x \sim \mathcal{N}(\theta, \Sigma)} g\left(\kappa\sigma \left(\frac{1}{\sqrt{u^\top \Sigma u}} u^\top \left(x - \hat{\theta} \right) \right) \right) - \mathbf{E}_{x \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma})} f^\star \circ g \left(\kappa\sigma \left(\frac{1}{\sqrt{u^\top \Sigma u}} u^\top \left(x - \hat{\theta} \right) \right) \right) \\ =& \sup_{\|u\|=1} \mathbf{E}_{z \sim \mathcal{N}(0,1)} g\left(\kappa\sigma \left(z + \frac{1}{\sqrt{u^\top \Sigma u}} (\theta - \hat{\theta}) \right) \right) - \mathbf{E}_{z \sim \mathcal{N}(0,1)} f^\star \circ g \left(\kappa\sigma \left(\frac{\sqrt{u^\top \hat{\Sigma} u}}{\sqrt{u^\top \Sigma u}} z \right) \right) \\ \leq& 4\kappa\epsilon L_g + 4\kappa L_g \sqrt{\frac{p}{n}} + 4\kappa L_g \sqrt{\frac{\log 1/\delta}{n}} \end{aligned}$$

Define

$$\psi_\xi(t) = \mathbf{E}_{z \sim \mathcal{N}(0,1)} g\left(t\sigma \left(z + \frac{1}{\sqrt{u^\top \Sigma u}} (\theta - \hat{\theta}) \right) \right) - \mathbf{E}_{z \sim \mathcal{N}(0,1)} f^\star \circ g \left(t\sigma \left(\frac{\sqrt{u^\top \hat{\Sigma} u}}{\sqrt{u^\top \Sigma u}} z \right) \right).$$

Then with probability at least $1 - 2\delta$, we have

$$\phi_{u^\top(\theta - \hat{\theta})}(t) \lesssim 4\kappa\epsilon L_g + 4\kappa L_g \sqrt{\frac{p}{n}} + 4\kappa L_g \sqrt{\frac{\log 1/\delta}{n}}.$$

By subgradient inequality of $\psi_\xi(t) + M(\kappa)\kappa^2$, we have

$$\phi_\xi(\kappa) + M(\kappa)\kappa^2 \geq \kappa \phi_\xi'(0),$$

where $M(\kappa)$ is the bound on the second order derivative of ϕ in $[0, \kappa]$ and $\psi_\xi(0) = 0$ by a similar argument as the proof of Theorem 1. Next, we upper bound $\|\hat{\theta} - \theta\|$ using $\phi'_\xi(0)$. A simple observation is that

$$\mathbf{E}_{z\sim\mathcal{N}(0,1)}\sigma(z) = \mathbf{E}_{z\sim\mathcal{N}(0,1)}\sigma\left(\frac{\sqrt{u^\top\hat{\Sigma}u}}{\sqrt{u^\top\Sigma u}}z\right) = \frac{1}{2}.$$

Thus (recall that $\partial f^\star(g(0)) = 1$)

$$\begin{aligned}\phi'_\xi(0) &= \mathbf{E}_{z\sim\mathcal{N}(0,1)}g'(0)\sigma\left(z + \frac{1}{\sqrt{u^\top\Sigma u}}\xi\right) - \mathbf{E}_{z\sim\mathcal{N}(0,1)}g'(0)\sigma\left(\frac{\sqrt{u^\top\hat{\Sigma}u}}{\sqrt{u^\top\Sigma u}}z\right)\\ &= \mathbf{E}_{z\sim\mathcal{N}(0,1)}\left[g'(0)\sigma\left(z + \frac{1}{\sqrt{u^\top\Sigma u}}\xi\right) - \mathbf{E}_{z\sim\mathcal{N}(0,1)}g'(0)\sigma\left(z\right)\right],\end{aligned}$$

which is exactly $\psi'_{\frac{\xi}{\sqrt{u^\top\Sigma u}}}(0)$ defined in the proof of Theorem 1. Thus, following the same argument, we have

$$\frac{\|\hat{\theta} - \theta\|}{\sqrt{u^\top\Sigma u}} \lesssim \epsilon + \sqrt{\frac{p}{n}} + \sqrt{\frac{\log 1/\delta}{n}},$$

whenever $\kappa \lesssim \sqrt{\frac{p}{n}} + \epsilon$ and $\sqrt{\frac{p}{n}} + \epsilon$ is sufficiently small. Finally, notice that $\sqrt{u^\top\Sigma u}$ is upper bounded by some constant since Σ has bounded spectral norm, which finishes the proof. □