Generative Adversarial Nets for Robust Scatter Estimation: A Proper Scoring Rule Perspective

Chao Gao   CHAOGAO@GALTON.UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago IL 60637, USA

Yuan Yao   YUANY@UST.HK
Weizhi Zhu   WZHUAI@UST.HK
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong, China

Editor: Arthur Gretton

Abstract

Robust covariance matrix estimation is a fundamental task in statistics. The recent discovery on the connection between robust estimation and generative adversarial nets (GANs) by Gao et al. (2019) suggests that it is possible to compute depth-like robust estimators using similar techniques that optimize GANs. In this paper, we introduce a general *learning via classification* framework based on the notion of proper scoring rules. This framework allows us to understand both matrix depth function, a technique of rate-optimal robust estimation, and various GANs through the lens of variational approximations of f-divergences induced by proper scoring rules. We then propose a new class of robust covariance matrix estimators in this framework by carefully constructing discriminators with appropriate neural network structures. These estimators are proved to achieve the minimax rate of covariance matrix estimation under Huber's contamination model. The results are also extended to robust scatter estimation for elliptical distributions. Our numerical results demonstrate the good performance of the proposed procedures under various settings against competitors in the literature.

Keywords: robust statistics, neural networks, minimax rate, data depth, contamination model, GAN

1. Introduction

We study robust covariance matrix estimation under Huber's contamination model (Huber, 1964, 1965). In this setting, one has observations $X_1, ..., X_n \overset{iid}{\sim} (1-\epsilon)N(0, \Sigma) + \epsilon Q$ in $\mathbb{R}^p$, and the goal is to estimate the covariance matrix Σ using contaminated data without any assumption on the contamination distribution Q. Even though many robust covariance matrix estimators have been proposed and analyzed in the literature (Maronna, 1976; Tyler,

1987; Zuo and Serfling, 2000; Han and Liu, 2013; Mitra and Zhang, 2014; Wegkamp and Zhao, 2016), the problem of optimal covariance estimation under the contamination model has not been investigated until the recent work by Chen et al. (2018). It was shown in Chen et al. (2018) that the minimax rate with respect to the squared operator norm $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2$ is $\frac{p}{n} \vee \epsilon^2$. An important feature of the minimax rate is its *dimension-free* dependence on the contamination proportion ϵ through the second term ϵ^2. An estimator that can achieve the minimax rate is given by the maximizer of the covariance matrix depth function (Zhang, 2002; Chen et al., 2018; Paindaveine and Van Bever, 2018).

Despite its statistical optimality, the robust covariance matrix estimator that maximizes the depth function cannot be efficiently computed unless the dimension of the data is extremely low. This is the same weakness that is also shared by Tukey's halfspace depth (Tukey, 1975) and Rousseeuw and Hubert's regression depth (Rousseeuw and Hubert, 1999). In fact, even an approximate algorithm that computes these depth functions takes $O(e^{Cp})$ in time (Rousseeuw and Struyf, 1998; Amenta et al., 2000; Chan, 2004; Chen et al., 2018).

On the other hand, a recent connection between depth functions and *Generative Adversarial Nets* (GANs) was discovered by Gao et al. (2019). The GAN (Goodfellow et al., 2014) is a very popular technique in deep learning to learn complex distributions such as the generating process of images. In the formulation of GAN, there is a *generator* and a *discriminator*. The generator, modeled by a neural network, is trying to learn a distribution as close to the data as possible, while the discriminator, modeled by another neural network, is trying to distinguish samples from the generator and data. If the generator class and the discriminator class are chosen appropriately, this two-player game will reach its equilibrium when the discriminator cannot tell the difference between samples from the generator and the data, and that means the generator has successfully learned the underlying distribution of the data. Since GAN can be written as a minimax optimization problem, this suggests a mathematical resemblance to the robust estimators that are maximizers of depth functions, which are maximin optimization problems. Indeed, under the framework of f-Learning, it was shown by Gao et al. (2019) that both procedures are minimizers of variational lower bounds of f-divergence functions. While GAN minimizes the Jensen-Shannon divergence, the robust estimators induced by depth functions all minimize the total variation distance. An alternative perturbation view on the connection between GAN and robust estimation was later discussed by Zhu et al. (2019).

The connection between GAN, or more generally, f-GAN (Nowozin et al., 2016), and robust estimation opens a door of approximating these hard-to-compute depth functions by neural networks, and then standard techniques used routinely to train GANs can be applied to compute various robust estimators. Appropriate choices of neural network structures have been discussed in Gao et al. (2019), but only for optimal robust location estimation.

In this paper, our goal is to understand appropriate network structures for optimal robust covariance or scatter matrix estimation under the framework of *learning via classification*. Our main result shows that the network structures for optimal robust location estimation may not have sufficient discriminative power for optimal covariance matrix es-

timation. Therefore, we propose necessary modifications of the network structures so that optimal covariance matrix estimation under Huber's contamination model can be achieved.

The idea of *learning via classification* has longstanding roots in statistics and machine learning (Diggle and Gratton, 1984; Freund and Schapire, 1996; Buja et al., 2005; Hyvärinen, 2005; Gneiting and Raftery, 2007; Mohamed and Lakshminarayanan, 2016; Dawid, 2007; Devroye and Lugosi, 2012; Sutherland et al., 2016; Arjovsky et al., 2017; Gutmann et al., 2018; Bińkowski et al., 2018; Baraud and Birgé, 2018). In this paper, we further expand this scope of statistical learning by building a general estimation framework using classification with cost functions derived from proper scoring rules (Buja et al., 2005; Gneiting and Raftery, 2007; Dawid, 2007). Our framework is partly inspired by the discussion in Mohamed and Lakshminarayanan (2016). Connections between proper scoring rules and f-GANs are also observed in Tan et al. (2019). Using Savage representation (Savage, 1971), we identify a class of smooth objective functions that can be used for training optimal robust procedures under Huber's contamination model. The variational lower bounds of these objective functions cover important special cases including GANs and depth functions.

Main Contributions. We summarize our main contributions of the paper as follows.

- We formulate a general framework of learning via classification using the concept of proper scoring rules. We show that the class of learning procedures under this framework has a one-to-one correspondence to a class of symmetric f-divergences, thanks to the Savage representation of the proper scoring rules. As a result, it leads to various forms of GANs and depth functions that are suitable for robust estimation under Huber's contamination model.

- We propose appropriate neural network structures for the classifiers in the proper scoring rules in the context of optimal robust covariance matrix estimation. We show that symmetry and boundedness of the activation function are the keys to the optimal robust estimation of covariance matrix. In addition, width and depth of the network structure would also have an influence in practice. Depending on whether the intercept node is included or not, we show that the neural network is required to have at least one or two hidden layers for the robust covariance estimation task.

- We also study robust scatter matrix estimation under general elliptical distributions. We show that in such a *semiparametric learning* setting, one does not need to use a more complicated *discriminator*, and only the *generator* of the GAN needs to be modified.

Connections to the Literature. Our work is closely related to the recent developments on the statistical properties of GANs and the literature of robust covariance estimation under Huber's contamination model. For example, generalization bounds of GANs were established by Zhang et al. (2017). Nonparametric density estimation using GANs was studied by Liang

(2017). Provable guarantees of learning Gaussian distributions with quadratic discriminators were established by Feizi et al. (2017). Theoretical guarantees of learning Gaussian mixtures, exponential families and injective neural network generators were obtained by Bai et al. (2018). The connection between GANs and robust estimation was established by Gao et al. (2019) and also studied by Zhu et al. (2019). Polynomial-time algorithms for robust covariance matrix estimation under Huber's contamination model have been considered by Lai et al. (2016); Diakonikolas et al. (2016); Du et al. (2017); Diakonikolas et al. (2017) among others in the literature, with the main focus on error bounds in Frobenius norm and total variation loss.

Paper Organization. The rest of the paper is organized as follows. We develop an estimation framework of proper scoring rules in Section 2. In Section 3, we discuss robust covariance matrix estimation under this framework, and propose appropriate neural network classes for this task. An extension to simultaneous mean and covariance estimation is considered in Section 4. In Section 5, we consider general robust scatter matrix estimation under elliptical distributions. Our numerical results are given in Section 6. Section 7 collects all the proofs in the paper.

Notation. We close this section by introducing the notation used in the paper. For $a, b \in \mathbb{R}$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For an integer m, $[m]$ denotes the set $\{1, 2, ..., m\}$. Given a set S, $|S|$ denotes its cardinality, and $\mathbb{I}_S$ is the associated indicator function. For two positive sequences $\{a_n\}$ and $\{b_n\}$, the relation $a_n \lesssim b_n$ means that $a_n \leq C b_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For a vector $v \in \mathbb{R}^p$, $\|v\|$ denotes the ℓ_2 norm, $\|v\|_\infty$ the ℓ_∞ norm, and $\|v\|_1$ the ℓ_1 norm. For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we use $\|A\|_{\mathrm{op}}$ to denote its operator norm, which is its largest singular value. We use $\mathbb{P}$ and $\mathbb{E}$ to denote generic probability and expectation whose distribution is determined from the context. For two probability distributions P_1 and P_2, their total variation distance is $\mathsf{TV}(P_1, P_2) = \sup_B |P_1(B) - P_2(B)|$. The sigmoid function, the ramp function, and the rectified linear unit function (ReLU) are denoted by $\mathsf{sigmoid}(x) = \frac{1}{1+e^{-x}}$, $\mathsf{ramp}(x) = \max(\min(x + 1/2, 1), 0)$, and $\mathsf{ReLU}(x) = \max(x, 0)$.

2. An Estimation Framework of Proper Scoring Rules

The idea of learning via classification can be formulated as a two-player game. Given a probability distribution P, two players act with strategies $T \in \mathcal{T}$ and $Q \in \mathcal{Q}$ to optimize the cost function $D_T(P, Q)$. The first player chooses a classification rule T in the class $\mathcal{T}$ to distinguish samples generated by P from samples generated by Q. The second player then chooses a probability distribution Q in the class $\mathcal{Q}$ so that samples generated by Q cannot be distinguished from samples generated by P even when the first player uses the optimal classification rule. This minimax game can be formulated as

$$\min_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} D_T(P, Q). \tag{1}$$

4

The minimax strategy for the second player $\operatorname{argmin}_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} D_T(P, Q)$ can then be used to learn the distribution P. This is the principle behind the idea of GANs (Goodfellow et al., 2014; Nowozin et al., 2016) and many other statistical learning procedures in the literature (Diggle and Gratton, 1984; Freund and Schapire, 1996; Buja et al., 2005; Hyvärinen, 2005; Gneiting and Raftery, 2007; Mohamed and Lakshminarayanan, 2016; Dawid, 2007; Devroye and Lugosi, 2012; Sutherland et al., 2016; Arjovsky et al., 2017; Gutmann et al., 2018; Bińkowski et al., 2018; Baraud and Birgé, 2018). In this section, we discuss a class of cost functions $D_T(P, Q)$ induced by proper scoring rules. We show that the divergence function $D_\mathcal{T}(P, Q) = \max_{T \in \mathcal{T}} D_T(P, Q)$ can be viewed as a variational lower bound of some f-divergence, and the minimax strategy (1) used as a robust estimator has an interesting connection with the matrix depth function in the literature (Zhang, 2002; Chen et al., 2018; Paindaveine and Van Bever, 2018).

2.1 Proper Scoring Rules

Consider a binary event space $\Omega = \{0, 1\}$. A probabilistic forecast is a quoted probability $t \in [0, 1]$ for either 0 or 1 to occur. A scoring rule S is defined as a pair of functions $S(\cdot, 1)$ and $S(\cdot, 0)$. To be specific, $S(t, 1)$ is the forecaster's reward if he or she quotes t when the event 1 occurs, and $S(t, 0)$ is the reward when the event 0 occurs.

Suppose the event occurs with probability p. Then, the expected reward for the forecaster is given by the formula

$$S(t; p) = pS(t, 1) + (1 - p)S(t, 0).$$

It is called a proper scoring rule if

$$S(p; p) \geq S(t; p), \quad \text{for any } t \in [0, 1],$$

or equivalently $p \in \operatorname{argmax}_{t \in [0,1]} S(t; p)$. The scoring rule is strictly proper when the equality above holds if and only if $t = p$. In this paper, we restrict our discussion to binary proper scoring rules. Readers interested in more general definitions are referred to Buja et al. (2005); Gneiting and Raftery (2007); Dawid (2007).

2.2 The Savage Representation

A scoring rule S is *regular* if both $S(\cdot, 0)$ and $S(\cdot, 1)$ are real-valued, except possibly that $S(0, 1) = -\infty$ or $S(1, 0) = -\infty$. The celebrated Savage representation (Savage, 1971) asserts that a regular scoring rule S is proper if and only if there is a convex function $G(\cdot)$, such that

$$\begin{cases} S(t, 1) = G(t) + (1 - t)G'(t), \\ S(t, 0) = G(t) - tG'(t). \end{cases} \quad (2)$$

Here, $G'(t)$ is a subgradient of G at the point t. Moreover, the statement also holds for strictly proper scoring rules when convex is replaced by strictly convex.

For any regular proper scoring rule, the convex function $G(\cdot)$ can be determined by

$$G(t) = S(t; t) = tS(t, 1) + (1 - t)S(t, 0),$$

and the Savage representation simply says that $S(t; t)$ is a convex function in t.

2.3 Relation to f-Divergence

Given two probability distributions P and Q, a divergence function $D(P, Q)$ measures the difference between P and Q. It satisfies the following two properties:

1. For any P and Q, $D(P, Q) \geq 0$.

2. Whenever $P = Q$, $D(P, Q) = 0$.

Following the principle outlined in Mohamed and Lakshminarayanan (2016), we show that a general class of divergence functions can be induced from proper scoring rules. To motivate the derivation, we consider a classification problem by introducing a binary latent variable $y \in \{0, 1\}$. The conditional distribution of X given y is specified as $X|(y = 1) \sim P$ and $X|(y = 0) \sim Q$. We also assume that $\mathbb{P}(y = 1) = \frac{1}{2}$ so that the joint distribution (X, y) is fully specified. The classification problem is to find a function $T(X) \in [0, 1]$ that forecasts the probability of $y = 1$ given X. With a proper scoring rule $\{S(\cdot, 1), S(\cdot, 0)\}$, it is natural to consider the following cost function for the task,

$$\mathbb{E}\left[yS(T(X), 1) + (1 - y)S(T(X), 0)\right]$$
$$= \frac{1}{2}\mathbb{E}_{X \sim P}S(T(X), 1) + \frac{1}{2}\mathbb{E}_{X \sim Q}S(T(X), 0).$$

Then, one can find a good classification rule $T(\cdot)$ by maximizing the above objective over $T \in \mathcal{T}$. This leads to the following definition of a divergence function,

$$D_{\mathcal{T}}(P, Q) = \max_{T \in \mathcal{T}} \left[\frac{1}{2}\mathbb{E}_{X \sim P}S(T(X), 1) + \frac{1}{2}\mathbb{E}_{X \sim Q}S(T(X), 0)\right] - G(1/2), \qquad (3)$$

where $G(\cdot)$ is the convex function in the Savage representation of the proper scoring rule.

The definition (3) can be understood as a variational lower bound of some f-divergence. Given a convex function $f(\cdot)$ that satisfies $f(1) = 0$, recall that the definition of the f-divergence between P and Q is given by

$$D_f(P\|Q) = \int f\left(\frac{dP}{dQ}\right) dQ.$$

Proposition 1 *Given any regular proper scoring rule $\{S(\cdot, 1), S(\cdot, 0)\}$ and any class $\mathcal{T} \ni \{\frac{1}{2}\}$, $D_{\mathcal{T}}(P, Q)$ is a divergence function, and*

$$D_{\mathcal{T}}(P, Q) \leq D_f\left(P\middle\|\frac{1}{2}P + \frac{1}{2}Q\right), \qquad (4)$$

where $f(t) = G(t/2) - G(1/2)$. Moreover, whenever $\mathcal{T} \ni \frac{dP}{dP + dQ}$, the inequality above becomes an equality.

Proof Suppose $\mathcal{T} \ni \{\frac{1}{2}\}$, then $D_{\mathcal{T}}(P,Q) \geq \frac{1}{2}S(1/2,1) + \frac{1}{2}S(1/2,0) - G(1/2) = 0$. When $P = Q$, we have $D_{\mathcal{T}}(P,Q) \leq \max_{t \in [0,1]}\left[G(t) - G(1/2) - (t-1/2)G'(t)\right] \leq 0$ by the convexity of $G(\cdot)$, and therefore $D_{\mathcal{T}}(P,Q) = 0$, which implies it is a divergence function. Since $\{S(\cdot,1), S(\cdot,0)\}$ is a proper scoring rule, $p(x)S(T(x),1) + q(x)S(T(x),0)$ is maximized at $T(x) = \frac{p(x)}{p(x)+q(x)}$. Thus,

$$
\begin{aligned}
D_{\mathcal{T}}(P,Q) &\leq \frac{1}{2}\mathbb{E}_{X \sim P}S\left(\frac{dP}{dP+dQ}(X),1\right) + \frac{1}{2}\mathbb{E}_{X \sim Q}S\left(\frac{dP}{dP+dQ}(X),0\right) - G(1/2) \\
&= \frac{1}{2}\mathbb{E}_{X \sim P}G\left(\frac{dP}{dP+dQ}(X)\right) + \frac{1}{2}\mathbb{E}_{X \sim Q}G\left(\frac{dP}{dP+dQ}(X)\right) - G(1/2) \\
&= D_f\left(P\middle\|\frac{1}{2}P + \frac{1}{2}Q\right),
\end{aligned}
$$

and obviously the inequality above becomes an equality when $\mathcal{T} \ni \frac{dP}{dP+dQ}$. ∎

It is worth noting that $D_f\left(P\middle\|\frac{1}{2}P + \frac{1}{2}Q\right)$ is in general not symmetric with respect to P and Q. However, when the regular proper scoring rule is symmetric in the sense that $S(t,1) = S(1-t,0)$, we have $G(t) = G(1-t)$, or equivalently, $f(t) = f(2-t)$, in which case the corresponding f-divergence satisfies

$$
D_f\left(P\middle\|\frac{1}{2}P + \frac{1}{2}Q\right) = D_f\left(Q\middle\|\frac{1}{2}P + \frac{1}{2}Q\right),
$$

and is symmetric.

2.4 Variational Lower Bounds and GANs

The variational form of the divergence function makes it easy to define a sample version of (3). Replacing $\mathbb{E}_{X \sim P}$ in (3) by the empirical measure, we have a divergence function between $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ and Q, which is a useful objective function for statistical estimation. Given a class of probability measures $\mathcal{Q}$, the induced estimator of P is defined by

$$
\widehat{P} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \max_{T \in \mathcal{T}} \left[\frac{1}{n}\sum_{i=1}^{n}S(T(X_i),1) + \mathbb{E}_{X \sim Q}S(T(X),0)\right]. \tag{5}
$$

We drop the term $-G(1/2)$ in (3) because it is a constant that does not affect the definition of (5). The formula (5) has an interpretation of a minimax game between two players. The goal of the first player is to find the best discriminator T that learns whether a sample is from the empirical distribution or the model distribution Q. The second player is to find a model distribution Q as close to the empirical distribution as possible so that the first player cannot tell the difference. In the context of deep learning, both the discriminator class $\mathcal{T}$ and the generator class $\mathcal{Q}$ are modeled by neural networks, and (5) is recognized as the technique of *Generative Adversarial Nets* proposed by Goodfellow et al. (2014). The relation between GANs and proper scoring rules was discussed by Mohamed and Lakshminarayanan (2016) in the context of learning implicit models.

2.5 Examples

1. *Log Score.* The log score is perhaps the most commonly used rule because of its various intriguing properties (Jiao et al., 2015). The scoring rule with $S(t, 1) = \log t$ and $S(t, 0) = \log(1 - t)$ is regular and strictly proper. Its Savage representation is given by the convex function $G(t) = t \log t + (1 - t) \log(1 - t)$, which is interpreted as the negative Shannon entropy of Bernoulli(t). The corresponding divergence function $D_{\mathcal{T}}(P, Q)$, according to Proposition 1, is a variational lower bound of the Jensen-Shannon divergence

$$\mathsf{JS}(P, Q) = \frac{1}{2} \int \log \left(\frac{dP}{dP + dQ} \right) dP + \frac{1}{2} \int \log \left(\frac{dQ}{dP + dQ} \right) dQ + \log 2.$$

Its sample version (5) is the original GAN proposed by Goodfellow et al. (2014) that is widely used in learning distributions of images.

2. *Zero-One Score.* The zero-one score $S(t, 1) = 2\mathbb{I}\{t \geq 1/2\}$ and $S(t, 0) = 2\mathbb{I}\{t < 1/2\}$ is also known as the misclassification loss. This is a regular proper scoring rule but not strictly proper. The induced divergence function $D_{\mathcal{T}}(P, Q)$ is a variational lower bound of the total variation distance

$$\mathsf{TV}(P, Q) = P \left(\frac{dP}{dQ} \geq 1 \right) - Q \left(\frac{dP}{dQ} \geq 1 \right) = \frac{1}{2} \int |dP - dQ|.$$

The sample version (5) is recognized as the TV-GAN that was extensively studied by Gao et al. (2019) in the context of robust estimation.

3. *Quadratic Score.* Also known as the Brier score (Brier, 1950), the definition is given by $S(t, 1) = -(1 - t)^2$ and $S(t, 0) = -t^2$. The corresponding convex function in the Savage representation is given by $G(t) = -t(1 - t)$. By Proposition 1, the divergence function (3) induced by this regular strictly proper scoring rule is a variational lower bound of the following divergence function,

$$\Delta(P, Q) = \frac{1}{8} \int \frac{(dP - dQ)^2}{dP + dQ},$$

known as the triangular discrimination. The sample version (5) belongs to the family of least-squares GANs proposed by Mao et al. (2017).

4. *Boosting Score.* The boosting score was introduced by Buja et al. (2005) with $S(t, 1) = -\left(\frac{1-t}{t} \right)^{1/2}$ and $S(t, 0) = -\left(\frac{t}{1-t} \right)^{1/2}$ and has an connection to the AdaBoost algorithm. The corresponding convex function in the Savage representation is given by $G(t) = -2\sqrt{t(1 - t)}$. The induced divergence function $D_{\mathcal{T}}(P, Q)$ is thus a variational lower bound of the squared Hellinger distance

$$H^2(P, Q) = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ} \right)^2.$$

5. *Beta Score.* A general Beta family of proper scoring rules was introduced by Buja et al. (2005) with $S(t,1) = -\int_t^1 c^{\alpha-1}(1-c)^\beta dc$ and $S(t,0) = -\int_0^t c^\alpha (1-c)^{\beta-1} dc$ for any $\alpha, \beta > -1$. The log score, the quadratic score and the boosting score are special cases of the Beta score with $\alpha = \beta = 0$, $\alpha = \beta = 1$, $\alpha = \beta = -1/2$. The zero-one score is a limiting case of the Beta score by letting $\alpha = \beta \to \infty$. Moreover, it also leads to asymmetric scoring rules with $\alpha \neq \beta$.

2.6 Relation to The Matrix Depth Function

Consider the covariance matrix estimation problem with contaminated observations $X_1, ..., X_n \overset{iid}{\sim} (1-\epsilon)N(0,\Sigma) + \epsilon Q$. A robust estimator is defined by

$$\widehat{\Sigma} = \underset{\Gamma \in \mathcal{E}_p}{\operatorname{argmin}} \max_{\|u\|=1} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \leq \beta u^T \Gamma u\} \vee \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{|u^T X_i|^2 \geq \beta u^T \Gamma u\} \right], \quad (6)$$

where $\mathcal{E}_p$ is the set of all $p \times p$ covariance matrices, and β is a scalar determined by the equation $\mathbb{P}(N(0,1) \leq \sqrt{\beta}) = 3/4$. The estimator (6) is the minimizer of what is known as the matrix depth function (Zhang, 2002; Chen et al., 2018; Paindaveine and Van Bever, 2018). The intuition is that $\beta u^T \widehat{\Sigma} u$ is approximately the squared median absolute deviation of $u^T X_1, ..., u^T X_n$ for all directions u's, and the scalar β is defined so that $u^T \widehat{\Sigma} u$ estimates the variance on the direction of u. It was shown by Chen et al. (2018) that the estimator (6) achieves the error bound

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}}^2 \lesssim \frac{p}{n} \vee \epsilon^2,$$

with high probability as long as $\|\Sigma\|_{\text{op}} = O(1)$. Moreover, the rate $\frac{p}{n} \vee \epsilon^2$ was also proved to be minimax optimal by Chen et al. (2018).

In this section, we will discuss the relation between (6) and GAN, and we will show that (6) can be viewed as a variant of TV-GAN with a particular choice of the discriminator class. This fact motivates us to investigate more general GANs that are appropriate for robust covariance matrix estimation in Section 3.

Consider the zero-one score, and then (5) is specialized as

$$\widehat{P} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{T(X_i) \geq 1/2\} + \mathbb{E}_{X \sim Q} \mathbb{I}\{T(X) < 1/2\} \right]. \quad (7)$$

We also consider a variation of (7) defined by

$$\widehat{P} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \max_{T \in \mathcal{T}_Q} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}\{T(X_i) \geq 1/2\} + \mathbb{E}_{X \sim Q} \mathbb{I}\{T(X) < 1/2\} \right]. \quad (8)$$

The subtle difference of (8) compared with (7) is the dependence of the discriminator class on Q. In fact, both (8) and (7) can be regarded as the minimizers of variational lower bounds of the total variation distance. The connection between (8) and various depth functions in robust estimation was discussed in an f-Learning framework by Gao et al. (2019). For the

purpose of covariance matrix estimation, we now show that (8) can be specialized into (6). We first set

$$\mathcal{Q} = \{N(0, \Gamma) : \Gamma \in \mathcal{E}_p\},$$

and then choose

$$\mathcal{T}_{N(0,\Gamma)} = \left\{ T = \frac{dN(0, \beta\widetilde{\Gamma})}{dN(0, \beta\widetilde{\Gamma}) + dN(0, \beta\Gamma)} : \widetilde{\Gamma}^{-1} = \Gamma^{-1} + \widetilde{r}uu^T \in \mathcal{E}_p, |\widetilde{r}| \leq r, \|u\| = 1 \right\}.$$

The choice of the local discriminator class $\mathcal{T}_{N(0,\Gamma)}$ can be motivated by the conclusion of Proposition 1 that the optimal discriminator between P and Q is $\frac{dP}{dP+dQ}$. By direct calculation, we have

$$\mathbb{I}\left\{ \frac{dN(0, \beta\widetilde{\Gamma})}{dN(0, \beta\widetilde{\Gamma}) + dN(0, \beta\Gamma)}(X) \geq 1/2 \right\} = \mathbb{I}\left\{ \widetilde{r}|u^T X|^2 \leq \log(1 + \beta\widetilde{r}u^T\Gamma u) \right\}. \qquad (9)$$

Therefore, we can write (8) as

$$\begin{aligned}
\widehat{\Sigma} &= \underset{\Gamma \in \mathcal{E}_p}{\operatorname{argmin}} \max_{\substack{\|u\|=1 \\ |\widetilde{r}| \leq r}} \left[\frac{1}{n} \sum_{i=1}^n \mathbb{I}\left\{ \widetilde{r}|u^T X_i|^2 \leq \log(1 + \beta\widetilde{r}u^T\Gamma u) \right\} \right. \\
&\qquad\qquad\qquad\qquad \left. - \mathbb{P}_{X \sim N(0,\Gamma)}\left\{ \widetilde{r}|u^T X|^2 \leq \log(1 + \beta\widetilde{r}u^T\Gamma u) \right\} \right]. \qquad (10)
\end{aligned}$$

Since $\lim_{\widetilde{r}\to 0} \frac{\log(1+\beta\widetilde{r}u^T\Gamma u)}{\widetilde{r}u^T\Gamma u} = \beta$, the limiting event of (9) is either $\mathbb{I}\{|u^T X|^2 \leq \beta u^T\Gamma u\}$ or $\mathbb{I}\{|u^T X|^2 \geq \beta u^T\Gamma u\}$, depending on whether $\widetilde{r}$ tends to zero from left or from right. Moreover, $\mathbb{P}_{X \sim N(0,\Gamma)}\{|u^T X|^2 \leq \beta u^T\Gamma u\} = 1/2$ by the definition of β. Therefore, as $r \to 0$, the formula (10) becomes (6).

Despite the statistical optimality of the estimator (6), its optimization is computationally infeasible in practice whenever the dimension exceeds 10 (Chen et al., 2018). This is partly due to the fact that the zero-one loss is not smooth. However, even the smooth version of the TV-GAN was shown to be computationally intractable, which motivates Gao et al. (2019) to consider alternative proper scoring rules such as the log score in the setting of robust mean estimation.

The rest of the paper is devoted to the investigation of GANs that not only enjoy the statistical optimality (6), but also use discriminator classes that are easier to optimize. For this purpose, we are going to replace the zero-one score in (6)-(8) by smooth scoring rules. To leverage the computational strategies developed in the deep learning community (Goodfellow et al., 2014; Radford et al., 2015; Salimans et al., 2016), we will consider discriminator classes that are family of neural network classifiers. It turns out the choice of the neural network structure is a very subtle issue in robust covariance matrix estimation.

3. Network Structures for Robust Covariance Matrix Estimation

We study the property of the estimator (5) in the context of robust covariance matrix estimation. Define

$$\mathcal{E}_p(M) = \{\Gamma \in \mathcal{E}_p : \|\Gamma\|_{\mathrm{op}} \leq M\}.$$

For covariance matrix estimation, we write (5) as

$$\widehat{\Sigma} = \operatorname*{argmin}_{\Gamma \in \mathcal{E}_p(M)} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) + \mathbb{E}_{X \sim N(0,\Gamma)} S(T(X), 0) \right]. \tag{11}$$

We consider i.i.d. observations drawn from Huber's ϵ-contamination model (Huber, 1964, 1965). That is,

$$X_1, ..., X_n \overset{iid}{\sim} (1 - \epsilon) N(0, \Sigma) + \epsilon Q. \tag{12}$$

In other words, each observation has an ϵ probability to be drawn from an unknown contamination distribution Q. A more general data generating process is called the *strong contamination model*. In such a setting, we have

$$X_1, ..., X_n \overset{iid}{\sim} P \quad \text{for some } P \text{ satisfying } \mathsf{TV}(P, N(0, \Sigma)) \leq \epsilon, \tag{13}$$

which means that the observations are drawn from an unknown probability distribution in a total variation neighborhood of $N(0, \Sigma)$. It is easy to see that (12) implies (13) so that (13) is a more general notion of contamination. While the contamination is only allowed to be added into good samples in (12), the adversarial can now choose to remove some good samples after looking at the data in the setting of (13). See Diakonikolas et al. (2016) for a detailed discussion on various contamination models.

Under both (12) and (13), the minimax rate of covariance matrix estimation with respect to the loss function $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2$ is $\frac{p}{n} \vee \epsilon^2$, and can be achieved by (6). This fact was proved by Chen et al. (2018) under Huber's contamination model (12), and the same proof can be extended to the strong contamination model (13).

In this section, we study the statistical properties of (11) with general proper scoring rules. We will discuss appropriate choices of the discriminator class $\mathcal{T}$ for robust covariance matrix estimation. We consider $\mathcal{T}$ that is some family of neural network classifiers. Then, the structure of the neural nets is essential in determining the statistical properties of (11). We will present two network structures that are not appropriate for robust covariance matrix estimation, and then show simple modifications of the two structures lead to optimal estimation.

3.1 Inappropriate Network Structures

Consider the following two discriminator classes. The first class contains two-layer sigmoid neural nets,

$$\mathcal{T}_1 = \left\{ T(x) = \mathsf{sigmoid}\left(\sum_{j \geq 1} w_j \mathsf{sigmoid}(u_j^T x) \right) : \sum_{j \geq 1} |w_j| \leq \kappa, u_j \in \mathbb{R}^p \right\}. \tag{14}$$

The second class also contains two-layer neural nets, but uses ReLU activations in the hidden layer,

$$\mathcal{T}_2 = \left\{ T(x) = \mathsf{sigmoid} \left(\sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T x) \right) : \sum_{j \geq 1} |w_j| \leq \kappa, \|u_j\| \leq 1 \right\}. \tag{15}$$

The network structures of $\mathcal{T}_1$ and $\mathcal{T}_2$ are visualized in Figure 1. The reasons that they do

Figure 1: Two structures of neural nets that are not suitable for robust covariance matrix estimation.

not work are different for the two structures. To construct concrete counterexamples, we focus on the log score in this section. That is, we consider the estimator

$$\widehat{\Sigma} = \underset{\Gamma \in \mathcal{E}_p(M)}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]. \tag{16}$$

The first class (14) leads to optimal robust mean estimation, but fails to learn the covariance matrix even if there is no contamination in the data. The following result shows the capability of (14) in learning a mean vector.

Proposition 2 *Consider the estimator*

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \max_{T \in \mathcal{T}_1} \left[\frac{1}{n} \sum_{i=1}^{n} \log T(X_i) + \mathbb{E}_{X \sim N(\eta, I_p)} \log(1 - T(X)) \right],$$

with $\mathcal{T}_1$ specified by (14). Assume $\frac{p}{n} + \epsilon^2 \leq c$ for some sufficiently small constant $c > 0$, and set $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$. With i.i.d. observations $X_1, ..., X_n \sim P$, we have

$$\|\widehat{\theta} - \theta\|^2 \leq C\left(\frac{p}{n} \vee \epsilon^2\right),$$

with probability at least $1 - e^{-C'(p+n\epsilon^2)}$ uniformly over all $\|\theta\| \leq M = O(1)$ and all P such that $\mathsf{TV}(P, N(\theta, I_p)) \leq \epsilon$. The constants $C, C' > 0$ are universal.

The success of robust estimation via a two-layer neural network was first proved by Gao et al. (2019) under Huber's ϵ-contamination model. Proposition 2 extends the result to the strong contamination model. However, the same neural network structure cannot learn a covariance matrix, as is shown by the following result.

Proposition 3 *With $\mathcal{T}_1$ specified by (14), the function*

$$F(\Sigma, \Gamma) = \max_{T \in \mathcal{T}_1} \left[\mathbb{E}_{X \sim N(0,\Sigma)} \log T(X) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T(X)) \right]$$

is a constant for all $\Sigma, \Gamma \in \mathcal{E}_p$.

In an ideal situation where $\epsilon = 0$ and $n = \infty$, the estimator (16) becomes $\operatorname{argmin}_\Gamma F(\Sigma, \Gamma)$. However, Proposition 3 shows that the objective function $F(\Sigma, \Gamma)$ is completely flat, and thus every Γ is a global minimizer.

The second discriminator class (15) has a different problem. It actually leads to optimal covariance matrix estimation when $\epsilon = 0$, but it does not lead to robust estimation.

Proposition 4 *Consider the estimator (16), where $\mathcal{T} = \mathcal{T}_2$ is specified by (15) with at least two units in the hidden layer. Assume $\frac{p}{n} \leq c$ for some sufficiently small constant $c > 0$, and set $\kappa = O\left(\sqrt{\frac{p}{n}}\right)$. With i.i.d. observations $X_1, ..., X_n \sim N(0, \Sigma)$, we have*

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \leq C\frac{p}{n},$$

with probability at least $1 - e^{-C'p}$ uniformly over all $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$. The constants $C, C' > 0$ are universal.

The comparison between Proposition 3 and Proposition 4 shows that the subtle difference between the activation functions in the hidden layer directly affects the consistency of covariance matrix estimation. A simple change from sigmoid to ReLU leads to an optimal error rate in Proposition 4. Mathematically, this can be explained by the *feature matching effect* of GANs (Liu et al., 2017; Gao et al., 2019), which asserts that GAN cannot distinguish two generating processes with the same expectation of $g(\cdot)$. In other words, only features in the form of $\mathbb{E}g(X)$ can be learned from the data. Here $g(\cdot)$ is a feature map such that $\mathsf{sigmoid}(g(\cdot))$ is an element of the discriminator class and X is distributed according to the generative process. For the class (14), we have

$$\mathbb{E}_{X \sim N(0,\Sigma)}\mathsf{sigmoid}(u^T X) = 1/2, \tag{17}$$

for any u and any Σ, due to the symmetric property of the sigmoid activation function. The fact that the right hand side of (17) is a constant, independent of Σ, means that different choices of Σ's cannot be differentiated by the classifier $\{\mathsf{sigmoid}(u^T X) : u \in \mathbb{R}^p\}$. On the other hand, (15) uses the ReLU activation, and thus

$$\mathbb{E}_{X \sim N(0,\Sigma)}\mathsf{ReLU}(u^T X) = \sqrt{u^T \Sigma u}\sqrt{\frac{2}{\pi}}. \tag{18}$$

Since the right hand side of (18) depends on Σ through $u^T \Sigma u$, the factor $u^T \Sigma u$, which is the variance of $u^T X$, can be learned. Thus, the classifier $\{\mathsf{ReLU}(u^T X) : u \in \mathbb{R}^p\}$ is able to learn the whole covariance Σ by learning $u^T \Sigma u$ with all $u \in \mathbb{R}^p$. In summary, the difference

between Proposition 3 and Proposition 4 is a consequence of the symmetric property of the activation function.

However, as long as there is contamination in the data, the structure (15) does not lead to robust estimation. We show a one-dimensional counterexample in the following proposition.

Proposition 5 *With $\mathcal{T}_2$ specified by (15), we have*

$$[(1-\epsilon)\sigma+\epsilon\tau]^2 \in \underset{\gamma^2}{\text{argmin}} \max_{T\in\mathcal{T}_2} \left[\mathbb{E}_{X\sim(1-\epsilon)N(0,\sigma^2)+\epsilon N(0,\tau^2)} \log T(X) + \mathbb{E}_{X\sim N(0,\gamma^2)} \log(1 - T(X))\right],$$

for any $\sigma^2, \tau^2 > 0$.

The proposition considers a setting with an ϵ fraction of contaminated observations generated from $N(0, \tau^2)$. Then, in the ideal situation with $n = \infty$, $[(1 - \epsilon)\sigma + \epsilon\tau]^2$ is global minimizer. Since the value of τ is not restricted, this global minimizer can be arbitrarily far away from the variance σ^2 of the good samples. Intuitively, this is because the ReLU activation function is unbounded, and therefore, an ϵ change of the probability can change the right hand side of (18) arbitrarily.

3.2 Appropriate Network Structures

The network structures (14) and (15) can both be slightly modified to achieve optimal robust covariance matrix estimation. For the first discriminator class (14), we only need to add an intercept node in the input layer, which leads to the definition of the following discriminator class,

$$\mathcal{T}_3 = \left\{T(x) = \text{sigmoid}\left(\sum_{j\geq 1} w_j \text{sigmoid}(u_j^T x + b_j)\right) : \sum_{j\geq 1} |w_j| \leq \kappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R}\right\}. \quad (19)$$

For the second class (15), we need to add an extra sigmoid hidden layer. This gives

$$\mathcal{T}_4 = \left\{T(x) = \text{sigmoid}\left(\sum_{j\geq 1} w_j \text{sigmoid}\left(\sum_{l=1}^{H} v_{jl} \text{ReLU}(u_l^T x)\right)\right) : \right.$$

$$\left. \sum_{j\geq 1} |w_j| \leq \kappa_1, \sum_{l=1}^{H} |v_{jl}| \leq \kappa_2, \|u_l\| \leq 1\right\}. \quad (20)$$

These two modifications of (14) and (15) are illustrated in Figure 2.

We study the covariance matrix estimator (11) with a general regular proper scoring rule. Recall that a regular proper scoring rule admits the Savage representation (2) with a convex function $G(\cdot)$. We impose the following assumption on the convex function $G(\cdot)$.

Condition 6 (Smooth Scoring Rules) *We assume $G^{(2)}(1/2) > 0$ and $G^{(3)}(t)$ is continuous at $t = 1/2$. Moreover, there is a universal constant $c_0 > 0$, such that $2G^{(2)}(1/2) \geq G^{(3)}(1/2) + c_0$.*
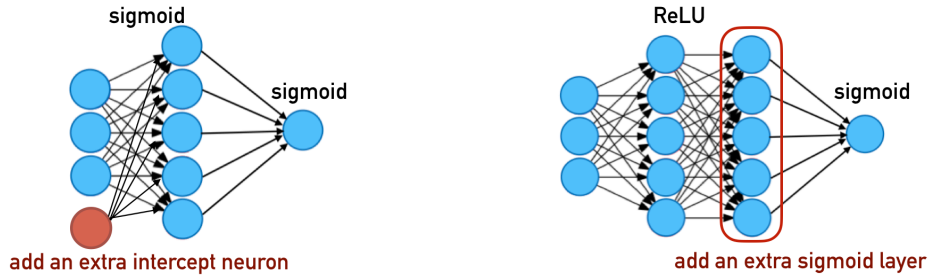
Figure 2: Simple fixes of the two network structures (14) and (15).

Condition 6 implies the scoring rule $\{S(\cdot, 1), S(\cdot, 0)\}$ is induced by two smooth functions, which excludes the zero-one loss. This is fine, because the zero-one loss was already studied as the matrix depth function in Chen et al. (2018). This paper only focuses on scoring rules that are feasible to optimize, and thus it is sufficient to restrict our results to smooth ones. The condition $2G^{(2)}(1/2) \geq G^{(3)}(1/2) + c_0$ is automatically satisfied by a symmetric scoring rule, because $S(t, 1) = S(1 - t, 0)$ immediately implies that $G^{(3)}(1/2) = 0$. For the Beta score with $S(t, 1) = -\int_t^1 c^{\alpha-1}(1-c)^{\beta} dc$ and $S(t, 0) = -\int_0^t c^{\alpha}(1-c)^{\beta-1} dc$ for any $\alpha, \beta > -1$, it is easy to check that such a c_0 (only depending on α, β) exists as long as $|\alpha - \beta| < 1$.

Theorem 7 *Consider the estimator (11) that is induced by a regular proper scoring rule that satisfies Condition 6, and $\mathcal{T} = \mathcal{T}_3$ is specified by (19). Assume $\frac{p}{n} + \epsilon^2 \leq c$ for some sufficiently small constant $c > 0$, and set $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$. Then, under the data generating process (13), we have*

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}}^2 \leq C\left(\frac{p}{n} \vee \epsilon^2\right),$$

with probability at least $1 - e^{-C'(p+n\epsilon^2)}$ uniformly over all $\|\Sigma\|_{\text{op}} \leq M = O(1)$. The constants $C, C' > 0$ are universal.

The theorem shows that the discriminator class (19) leads to optimal robust covariance matrix estimation, while the only difference between (14) and (19) is the inclusion of the intercept neuron in the bottom layer of the network in the class (19). In contrast to the common understanding that whether to include the intercept neuron in a neural network structure is only a matter of data normalization, here for the purpose of robust covariance matrix estimation using proper scoring rules, it is a fundamental issue.

The number κ explicitly controls the size of the weights of the networks, and thus the Rademacher complexity of the discriminator class. The constraint $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$ is critical for the robust estimator $\widehat{\Sigma}$ to achieve the optimal rate under the contamination model. In addition, a small κ also ensures that the closeness of the objective functions implies the closeness of the model parameters. On the other hand, if the population version is considered (with $n = \infty$ and $\epsilon = 0$), the true covariance matrix Σ can be exactly identified by the network structure (19) without any restriction on κ.

We also remark that the width of the network structure is not important. It can take any number from 1 to arbitrarily large. This is because the complexity of a neural network class can be characterized by the norm constraint instead of the explicit number of nodes (Bartlett, 1998; Bartlett and Mendelson, 2002). In practice, however, wide network structures will improve the behavior of gradient-based algorithms.

Theorem 8 *Consider the estimator (11) that is induced by a regular proper scoring rule that satisfies Condition 6, and $\mathcal{T} = \mathcal{T}_4$ is specified by (20). Assume $\frac{p}{n} + \epsilon^2 \leq c$ for some sufficiently small constant $c > 0$. Set $H \geq 2$, $\kappa_1 = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$, and $1 \leq \kappa_2 = O(1)$. Then, under the data generating process (13), we have*

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \leq C\left(\frac{p}{n} \vee \epsilon^2\right),$$

with probability at least $1 - e^{-C'(p+n\epsilon^2)}$ uniformly over all $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$. The constants $C, C' > 0$ are universal.

The ReLU activation function is widely used in training deep neural network models because of its superior optimization properties (Glorot et al., 2011). To estimate a covariance matrix, Theorem 8 shows that it can only be used after the two top layers. Otherwise, according to Proposition 5, the estimator would not be robust against arbitrary contamination.

4. Simultaneous Estimation of Mean and Covariance

In this section, we consider a more general setting where the data generating process is

$$X_1, ..., X_n \overset{iid}{\sim} P \quad \text{for some } P \text{ satisfying } \mathsf{TV}(P, N(\theta, \Sigma)) \leq \epsilon. \tag{21}$$

That is, both the mean vector θ and the covariance matrix Σ are unknown. We consider the estimation procedure

$$(\widehat{\theta}, \widehat{\Sigma}) = \underset{\eta \in \mathbb{R}^p, \Gamma \in \mathcal{E}_p(M)}{\mathrm{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n}\sum_{i=1}^{n} S(T(X_i), 1) + \mathbb{E}_{X \sim N(\eta, \Gamma)} S(T(X), 0)\right]. \tag{22}$$

Note that the generator class is $\{N(\eta, \Gamma) : \eta \in \mathbb{R}^p, \Gamma \in \mathcal{E}_p(M)\}$ compared with the centered class in (11).

We also introduce a more general discriminator class of deep neural nets. We first define a sigmoid bottom layer

$$\mathcal{G}_{\mathsf{sigmoid}} = \left\{g(x) = \mathsf{sigmoid}(u^T x + b) : u \in \mathbb{R}^p, b \in \mathbb{R}\right\}.$$

Then, with $\mathcal{G}^1(B) = \mathcal{G}_{\mathsf{sigmoid}}$, we inductively define

$$\mathcal{G}^{l+1}(B) = \left\{g(x) = \mathsf{ReLU}\left(\sum_{h \geq 1} v_h g_h(x)\right) : \sum_{h \geq 1} |v_h| \leq B, g_h \in \mathcal{G}^l(B)\right\}.$$

Note that the neighboring two layers are connected via ReLU activation functions. Finally, the network structure is defined by

$$\mathcal{T}^L(\kappa, B) = \left\{ T(x) = \mathsf{sigmoid}\left(\sum_{j \geq 1} w_j g_j(x) \right) : \sum_{j \geq 1} |w_j| \leq \kappa, g_j \in \mathcal{G}^L(B) \right\}. \tag{23}$$

This is a neural network class that consists of L hidden layers. When $L = 1$, (23) recovers the definition of the class (19).

Theorem 9 *Consider the estimator (22) that is induced by a regular proper scoring rule that satisfies Condition 6. The discriminator class $\mathcal{T} = \mathcal{T}^L(\kappa, B)$ is specified by (23). Assume $\frac{p}{n} + \epsilon^2 \leq c$ for some sufficiently small constant $c > 0$. Set $1 \leq L = O(1)$, $1 \leq B = O(1)$, and $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon \right)$. Then, under the data generating process (21), we have*

$$\|\widehat{\theta} - \theta\|^2 \leq C\left(\frac{p}{n} \vee \epsilon^2 \right),$$

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 \leq C\left(\frac{p}{n} \vee \epsilon^2 \right),$$

with probability at least $1 - e^{-C'(p+n\epsilon^2)}$ uniformly over all $\theta \in \mathbb{R}^p$ and all $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$. The constants $C, C' > 0$ are universal.

5. Elliptical Distributions

One of the most important statistical properties of the depth-based estimator (6) is its ability to adapt to general elliptical distributions (Chen et al., 2018). In this section, we show that the same property can also be achieved by robust estimators induced by proper scoring rules.

Definition 10 (Fang (2017)) *A random vector $X \in \mathbb{R}^p$ follows an elliptical distribution if and only if it has the representation $X = \theta + \xi AU$, where $\theta \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p \times r}$ are model parameters. The random variable U is distributed uniformly on the unit sphere $\{u \in \mathbb{R}^r : \|u\| = 1\}$ and $\xi \geq 0$ is a random variable in $\mathbb{R}$ independent of U. The vector θ and the matrix $\Sigma = AA^T$ are called the location and the scatter of the elliptical distribution.*

For any unit vector u, the distribution of $\xi u^T U$ does not depend on u because of the symmetry of U. Define $H(\cdot)$ to be the distribution function of $\xi u^T U$. Since there is a one-to-one relation between $H(\cdot)$ and the distribution of ξ, the distribution of $X = \theta + \xi AU$ is fully determined by the triplet (θ, Σ, H), and therefore we write the distribution as $E(\theta, \Sigma, H)$. See Figure 3 for an illustration.

Note that Σ and H are not identifiable, this is because $\xi AU = (a\xi)(a^{-1}A)U$ for any $a > 0$. To overcome this issue, we restrict H to the following class

$$\mathcal{H} = \left\{ H \text{ is a distribution function} : H(t) + H(-t) \equiv 1, \int R(|t|)dH(t) = \int R(|t|)d\Phi(t) \right\},$$
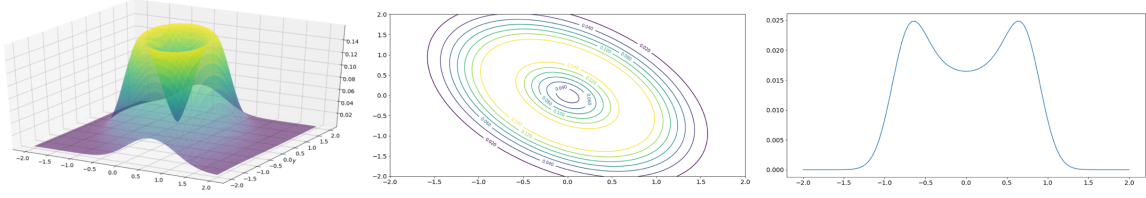
Figure 3: An illustration of a bivariate elliptical distribution (θ, Σ, H). From left to right: the density function, the contour whose shape is determined by Σ, and the marginal density of H.

where $\Phi(\cdot)$ is the distribution function of $N(0,1)$, and

$$R(|t|) = \begin{cases} |t|, & |t| \leq 1, \\ 1, & |t| > 1, \end{cases} \tag{24}$$

which is recognized as the clipped ℓ_1 function. The restriction $H \in \mathcal{H}$ is without loss of generality. This is because the function $F(a) = \mathbb{E}R(|a\xi u^T U|)$ is increasing for all $a > 0$, so that the equation $F(a) = \int R(|t|)d\Phi(t)$ must have a solution. Here, we do not use the simpler absolute function, because the first moment of $\xi u^T U$ may not exist.

Definition 11 *The elliptical distribution $X = \theta + \xi AU$ has a canonical parametrization (θ, Σ, H) with $\Sigma = AA^T$ and $H \in \mathcal{H}$. We use the notation $E(\theta, \Sigma, H)$ to denote the elliptical distribution in its canonical form.*

With the canonical representation, the parameters θ, Σ, H are all identifiable. The scatter matrix Σ is proportion to the covariance matrix whenever the covariance matrix exists. Moreover, for multivariate Gaussian $N(\theta, \Sigma)$, its canonical parametrization is (θ, Σ, Φ), and the scatter matrix and the covariance matrix are identical.

The goal of this section is to estimate both the location θ and the scatter Σ with observations

$$X_1, ..., X_n \overset{iid}{\sim} P \quad \text{for some } P \text{ satisfying } \mathsf{TV}(P, E(\theta, \Sigma, H)) \leq \epsilon. \tag{25}$$

To achieve this goal, we further require that H belongs to the following class

$$\mathcal{H}(M') = \left\{ H \in \mathcal{H} : \int_{1/4}^{1/3} dH(t) \geq \frac{1}{M'} \right\}.$$

The number $M' > 0$ is assumed to be some large constant. It is easy to see that $\mathcal{H} = \cup_{M'>0}\mathcal{H}(M')$. The regularity condition $H \in \mathcal{H}(M')$ will be easily satisfied as long as there is a constant probability mass of H contained in the interval $[1/4, 1/3]$. This condition prevents some of the probability mass from escaping to infinity.

Define the estimator

$$(\widehat{\theta}, \widehat{\Sigma}, \widehat{H}) = \underset{\eta \in \mathbb{R}^p, \Gamma \in \mathcal{E}_p(M), H \in \mathcal{H}(M')}{\operatorname{argmin}} \max_{T \in \mathcal{T}} \left[\frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) + \mathbb{E}_{X \sim E(\eta, \Gamma, H)} S(T(X), 0) \right]. \tag{26}$$

To accommodate for the more general generator class in (26), we consider the discriminator class $\bar{\mathcal{T}}^L(\kappa, B)$, which has the same definition (23), except that

$$\mathcal{G}^1(B) = \mathcal{G}_{\mathsf{ramp}} = \left\{ g(x) = \mathsf{ramp}(u^T x + b) : u \in \mathbb{R}^p, b \in \mathbb{R} \right\}.$$

In other words, $\bar{\mathcal{T}}^L(\kappa, B)$ and $\mathcal{T}^L(\kappa, B)$ only differs in the choice of the nonlinear activation function of the bottom layer. We remark that the discriminator class $\mathcal{T}^L(\kappa, B)$ also works for the elliptical distributions, but the theory would require a condition that is less transparent. The theoretical guarantee of the estimator (26) is given by the following theorem.

Theorem 12 *Consider the estimator (26) that is induced by a regular proper scoring rule that satisfies Condition 6. The discriminator class is specified by $\mathcal{T} = \bar{\mathcal{T}}^L(\kappa, B)$ with the dimension of (w_j) to be at least 2. Assume $\frac{p}{n} + \epsilon^2 \leq c$ for some sufficiently small constant $c > 0$. Set $2 \leq L = O(1)$, $1 \leq B = O(1)$, and $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$. Then, under the data generating process (25), we have*

$$
\begin{aligned}
\|\widehat{\theta} - \theta\|^2 &\leq C\left(\frac{p}{n} \vee \epsilon^2\right), \\
\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}^2 &\leq C\left(\frac{p}{n} \vee \epsilon^2\right),
\end{aligned}
$$

with probability at least $1 - e^{-C'(p+n\epsilon^2)}$ uniformly over all $\theta \in \mathbb{R}^p$, all $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$, and all $H \in \mathcal{H}(M')$ with $M' = O(1)$. The constants $C, C' > 0$ are universal.

6. Numerical Studies

We present our numerical results in this Section. We first give details of implementations in Section 6.1. We then compare our proposed methods with other methods in the literature in Section 6.2. Simultaneous location and scatter estimation and comparisons of different scoring rules are investigated in Section 6.3 and Section 6.4, respectively.

6.1 Implementations

In order to implement the proposed methods for scatter estimation, we need to specify the generator networks. Depending on whether the data is centered and whether the distribution family is known, we consider the following four types of generator networks:

- $G_1(Z; A) = AZ$. The random vector Z is sampled from $E(0, I_p, H)$. Then, according to Definition 10, we have $G_1(Z; A) \sim E(0, \Sigma, H)$ with $\Sigma = AA^T$. This is the simplest generator network suitable for centered observations with a known distribution family.

- $G_2(U, z; A, w_g) = g_{w_g}(z)AU$. When the distribution family is unknown, we can represent an $E(0, AA^T, H)$ random variable by ξAU according to Definition 10. This leads to the generator network $G_2(U, z; A, w_g)$, where we model ξ by $g_{w_g}(z)$, a neural work with parameter w_g and input vector $z \sim N(0, I_q)$. The input dimension q will be specified later.

- $G_3(Z; A, \theta) = \theta + AZ$. This is an extension of G_1 when the observations are not centered.
- $G_4(U, z; A, w_g, \theta) = \theta + g_{w_g}(z)AU$. This is an extension of G_2 when the observations are not centered.

The algorithm to implement proper scoring rule GAN with the generator network $G_1(Z; A)$ is given below.

Algorithm 1 $\operatorname{argmin}_A \max_w \left[\frac{1}{n} \sum_{i=1}^{n} S(D_w(X_i), 1) + \mathbb{E}_{Z \sim E(0, I_p, H)} S(D_w(G_1(Z; A)), 0) \right]$

Input:

1. Observations $O = \{X_1, \ldots, X_n\} \subset \mathbb{R}^p$,
2. Learning rates γ_d / γ_g for the discriminator/generator,
3. Batch size m,
4. Iterations for discriminator/generator steps in each epoch K_d / K_g,
5. Total epochs T,
6. Average epochs T_0.

Initialization:

1. Initialize $\widehat{\Sigma}_0$ by scaled Kendall's τ, apply singular value decomposition $\widehat{\Sigma}_0 = V_0 \Lambda_0 V_0^T$, and set $A_0 = V_0 \Lambda_0^{1/2}$.

2. Initialize the discriminator network by Xavier (Glorot and Bengio, 2010), where the first layer of $D_w(x)$ is initialized by some Gaussian distribution.

1: **for** $t = 1, \ldots, T$ **do**
2:     **for** $k = 1, \ldots, K_d$ **do**
3:         Sample mini-batch $\{X_1, \ldots, X_m\}$ from O. Sample $\{Z_1, \ldots, Z_m\}$ from $E(0, I_p, H)$;
4:         $g_w \leftarrow \nabla_w \left[\frac{1}{m} \sum_{i=1}^{m} S(D_w(X_i), 1) + \frac{1}{m} \sum_{i=1}^{m} S(D_w(G_1(Z_i; A)), 0) \right]$;
5:         $w \leftarrow w + \gamma_d g_w$;
6:     **end for**
7:     **for** $k = 1, \ldots, K_g$ **do**
8:         Sample $\{Z_1, \ldots, Z_m\}$ from $E(0, I_p, H)$;
9:         $g_A \leftarrow \nabla_A \left[\frac{1}{m} \sum_{i=1}^{m} S(D_w(G_1(Z_i; A)), 0) \right]$;
10:         $A \leftarrow A - \gamma_g g_A$;
11:     **end for**
12: **end for**

Return: The average over the last T_0 epochs $\widehat{\Sigma} = \frac{1}{T_0} \sum_{t=T-T_0+1}^{T} A_t A_t^T$.

Several remarks for Algorithm 1 are given below:

- *Some variations and extra details.* We provide general proper scoring rule GAN in Algorithm 1, and different choices of $\{S(\cdot, 1), S(\cdot, 0)\}$ in Section 2.5 leads to variant GANs. An extra implementation detail is that after every T_1 epochs, we update the learning rates γ_g and γ_d by $\alpha \gamma_g$ and $\alpha \gamma_d$ with some $\alpha \in (0, 1)$ for better convergence.

- *Discriminator network structures.* The structure of the discriminator network is set as p-$2p$-$\lfloor p/2 \rfloor$-1. We consistently observe that this wide structure outperforms a narrow ones such as p-$\lfloor p/2 \rfloor$-$\lfloor p/4 \rfloor$-1. The choice of nonlinearity follows that of (20), except we use $\mathsf{LeakyReLU}(x) = \max(0.2x, x)$ instead of $\mathsf{ReLU}(x)$ to avoid vanishing gradients. This slight change will not affect the theoretical results in the paper.

- *Identifiability and calibration.* Suppose $X \sim E(\theta, AA^T, H)$ so that it has the representation $X = \theta + \xi AU$. Then, the matrix $\widehat{A}\widehat{A}^T$ output by Algorithm 1 is an estimator for AA^T up to a multiplicative constant due to the identifiability issue discussed in Section 5. Therefore, we need to define the version of AA^T and calibrate the estimator $\widehat{A}\widehat{A}^T$ accordingly. In Section 5, the multiplicative constant is determined through the equation $\mathbb{E}R(|a\xi u^T U|) = \int R(|t|)d\Phi(t)$. Note that here the choice of Φ is arbitrary, and thus in our numerical studies, it is more convenient to replace the Gaussian cumulative distribution function (CDF) Φ by the CDF of the distribution that we are working with. For example, consider a multivariate t-distribution $T_v(\theta, \Sigma)$ with density proportional to $\left(1 + (x - \theta)^T \Sigma^{-1}(x - \theta)/v\right)^{-(v+p)/2}$, we will find a factor $a > 0$ such that $\mathbb{E}_{U,z}R(ag_{\widehat{w}_g}(z)u^T U) = \mathbb{E}_{X \sim T_v(0,1)}R(|X|)$ is satisfied, where U is uniformly distributed on the unit sphere and $z \sim N(0, I_q)$. Our final estimator is given by $a^{-2}\widehat{A}\widehat{A}^T$. With this scaling, the estimator is directly targeted at the Σ in the formula of the density of $T_v(\theta, \Sigma)$.

The following table summarizes the hyperparameters that can reproduce our numerical results for $p = 100$. Hyperparameters for other dimensions can be found in `https://github.com/zhuwzh/Robust-GAN-Scatter`.

generator	ξ network $g_{w_g}(z)$	discriminator	σ_1	γ_d/γ_g	K_d/K_g	T/T_0	α/T_1
G_1/G_3	-	100-200-50-1	0.0025	0.025/0.1	12/3	500/25	0.2/200
G_2/G_4	48-32-24-12-1	100-200-50-1	0.025	0.05/0.025	12/3	500/25	0.2/200

Table 1: Hyperparameters used in our numerical studies. Here, σ_1 is the standard deviation of the Gaussian initialization in the first layer (next to the input) of the discriminator network. Other parameters are introduced in Algorithm 1.

6.2 Comparisons with Other Methods

In this section, we compare the performance of JS-GAN against other methods for robust scatter estimation in the literature. We first introduce some other robust scatter matrix estimators that we will compare with. The definitions of these robust matrix estimator are all up to some scaling factor. Tyler's M-estimator (Tyler, 1987) is defined as a solution of $\sum_{i=1}^n \frac{X_i X_i^T}{X_i^T \Sigma^{-1} X_i} = c\Sigma$ for some $c > 0$. Note that it is a special case of Maronna's M-estimator (Maronna, 1976). Properties of Tyler's M-estimator were studied by Dümbgen (1998); Dümbgen and Tyler (2005); Zhang (2002); Zhang et al. (2016). The second robust estimator of scatter that we will compare with is the scaled Kendall's τ. The Kendall's τ

correlation coefficient (Kendall, 1938) between the jth and the kth variables is defined as

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{i<i'} \text{sign}\left((X_i - X_{i'})_j (X_i - X_{i'})_k\right).$$

Then, $\widehat{K} = (\widehat{K}_{jk})$ with $\widehat{K}_{jk} = \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right)$ is an estimator of the correlation matrix (Kruskal, 1958; Han and Liu, 2014). To obtain an estimator for the scatter matrix, define a diagonal matrix $\widehat{S}$ with diagonal entries $\widehat{S}_{jj} = \text{Median}(\{X_{ij}^2\}_{i=1}^n)$. Then, the scaled Kendall's τ estimator for the scatter matrix is $\widehat{S}^{1/2}\widehat{K}\widehat{S}^{1/2}$. Thirdly, we introduce the minimum volume ellipsoid estimator (MVE) by Leroy and Rousseeuw (1987). It finds the ellipsoid covering at least $n/2$ points of $\{X_i\}_{i=1}^n$ with the minimum volume and then use the shape of the ellipsoid as the scatter matrix estimator. Properties of MVE have been studied by Davies (1992). Finally, we consider the dimension halving method proposed by Lai et al. (2016) based on the idea of higher moment certification. We remark that among all the methods that we compare here, dimension halving is the only method that is designed for Huber's contamination model. For comparison of performances, we rescale all the estimators by some constant factors so that all of them are targeted at the same population scatter matrix.[1]

The comparisons cover the following scenarios from different perspectives.

Influence of Tail. We first study the influence of the tail behavior. We consider i.i.d. observations from $(1-\epsilon)T_v(0, \Sigma_{\text{ar}}) + \epsilon T_v(5\mathbb{1}_p, 5I_p)$, where $(\Sigma_{\text{ar}})_{jk} = (1/2)^{|j-k|}$ and $\mathbb{1}_p$ stands for the p-dimensional vector with all entries 1's. Note that the second moment of the multivariate t-distribution exists only when $v > 2$.

degrees of freedom v	$G_1(Z;A)$	$G_2(U,z;A,w_g)$	Dimension Halving	Tyler's M-estimator	Kendall's τ	MVE	Sample Covariance
1	0.2808 (0.0440)	0.3350 (0.0681)	-	372.9637 (582.3385)	52.5653 (0.6361)	50.2995 (0.6259)	-
2	0.3450 (0.0157)	0.4059 (0.0254)	-	55.5152 (1.1901)	64.7625 (0.4798)	20.1941 (1.8645)	-
4	0.2751 (0.0147)	0.2775 (0.0456)	1.2834 (0.0512)	38.7569 (0.2740)	72.8037 (0.3369)	0.1920 (0.0299)	64.5601 (0.3930)
8	0.2131 (0.0162)	0.2113 (0.0306)	0.8902 (0.0728)	39.0265 (0.2014)	77.2117 (0.3486)	0.1753 (0.0218)	72.7087 (0.3512)
16	0.1764 (0.0120)	0.2076 (0.0210)	0.8354 (0.0926)	39.1167 (0.3200)	79.2252 (0.2728)	0.1683 (0.0136)	75.4209 (0.2811)
32	0.1576 (0.0067)	0.2056 (0.0202)	0.8572 (0.0687)	39.1985 (0.2153)	80.2075 (0.1706)	0.1493 (0.0085)	76.5790 (0.4061)

Table 2: Simulation results with $n = 50,000, p = 100, \epsilon = 0.2$ and $v \in \{1, 2, 4, 8, 16, 32\}$. $G_1(Z;A) = AZ$, and $G_2(U, z; A, w_g) = g_{w_g}(z)AU$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

Table 2 summarizes the results with the degrees of freedom v varying from 1 to 32. We observe that JS-GANs (with generator networks G_1 and G_2) are overall the two best methods, especially when $v \in \{1, 2\}$. Dimension halving performs better than Tyler's M-estimator and Kendall's τ, because it is the only one that is designed for Huber's contamination model among all other methods. Among the remaining three methods, MVE is greatly influenced by the value of v, while Kendall's τ and Tyler's M-estimator does not seem to be robust in this setting.

---

1. Dimension halving is designed to estimate the covariance matrix when it exists. For the t-distribution with degrees of freedom v, the final estimator needs to be scaled by $\frac{v}{v-2}$ when $v > 2$. The results for $v \in \{1, 2\}$ are omitted for dimension halving because the covariance does not exist.

Distance of Contamination. We then study the effect of the distance of the contamination distribution. In this experiment, we sample i.i.d. observations from $(1-\epsilon)N(0, I_p) + \epsilon\delta_{s\mathbb{1}_p}$. That is, we model the contamination by a Dirac distribution with each coordinate being s, and we vary the value of s in this experiment.

factor s	$G_1(Z; A)$	$G_2(U, z; A, w_g)$	Dimension Halving	Tyler's M-estimator	Kendall's τ	MVE[2]	Sample Covariance
0.5	0.2000 (0.0251)	0.2057 (0.0104)	0.1949 (0.0085)	98.7102 (3.7054)	3.8098 (0.0158)	5.1546 (0.0424)	4.7303 (0.0307)
1	0.1699 (0.0168)	0.1607 (0.0084)	0.2163 (0.0077)	103.3438 (2.4459)	23.4627 (0.0732)	17.5985 (0.0964)	16.5368 (0.0667)
2	0.1705 (0.0145)	0.1576 (0.0097)	0.2762 (0.0159)	137.6124 (1.1288)	84.5634 (0.0750)	45.9484 (0.1704)	43.9837 (0.1812)
4	0.1530 (0.0059)	0.1557 (0.0142)	0.2982 (0.0110)	264.7385 (0.5537)	92.0462 (0.1206)	76.8558 (0.2466)	75.5743 (0.1890)
8	0.1600 (0.0094)	0.1555 (0.0116)	0.3008 (0.0184)	366.0933 (0.3620)	92.1079 (0.1124)	92.4466 (0.0811)	91.8662 (0.1037)

Table 3: Simulation results with $n = 50,000, p = 100, \epsilon = 0.2$ and $s \in \{0.5, 1, 2, 4, 8\}$. $G_1(Z; A) = AZ$, and $G_2(U, z; A, w_g) = g_{w_g}(z)AU$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

The results summarized in Table 3 show that JS-GANs and dimension halving are much better than the other three methods. The errors of Kendall's τ and MVE exhibit clear increasing patterns as s grows. Tyler's M-estimator does not seem to be robust under this setting.

It is also interesting to note that JS-GANs and dimension halving have different error behaviors. The error of dimension halving grows slowly as s increases, while the errors of JS-GANs decrease. This is because as the contamination distribution $\delta_{s\mathbb{1}_p}$ is moving away from $N(0, I_p)$, the discriminator network in GANs gets better at distinguishing contamination from the true model.

Dependence on (ϵ, n, p). Lastly, we show the comparison results in Tables 4-6 when the contamination proportion ϵ, the sample size n, and the dimension p vary. We consider i.i.d. observations from $(1-\epsilon)N(0, \Sigma_{\mathsf{ar}}) + \epsilon Q$ with $(\Sigma_{\mathsf{ar}})_{jk} = (1/2)^{|j-k|}$ and $Q \in \{N(5\mathbb{1}_p, 5I_p), \delta_{5\mathbb{1}_p}\}$.

ϵ	Q	$G_1(Z; A)$	$G_2(U, z; A, w_g)$	Dimension Halving	Tyler's M-estimator	Kendall's τ	MVE	Sample Covariance
0.02	$N(5\mathbb{1}_p, 5I_p)$	0.1438 (0.0109)	0.1829 (0.0221)	0.7709 (0.0720)	4.1114 (0.0410)	5.9289 (0.0362)	0.1286 (0.0115)	49.1036 (0.1110)
	$\delta_{5\mathbb{1}_p}$	0.1373 (0.0053)	0.1782 (0.0106)	0.7758 (0.0688)	105.4027 (2.1131)	6.4535 (0.0385)	20.0688 (17.1654)	48.9504 (0.0237)
0.05	$N(5\mathbb{1}_p, 5I_p)$	0.1434 (0.0096)	0.1834 (0.0123)	0.7633 (0.0639)	11.4392 (0.0739)	15.3559 (0.0430)	0.1313 (0.0117)	118.9069 (0.1821)
	$\delta_{5\mathbb{1}_p}$	0.1443 (0.0076)	0.1888 (0.0172)	0.7781 (0.0533)	130.9782 (0.3241)	16.8037 (0.0559)	49.7855 (17.4420)	118.6118 (0.0534)
0.1	$N(5\mathbb{1}_p, 5I_p)$	0.1500 (0.0111)	0.1881 (0.0143)	0.7548 (0.0435)	22.5296 (0.1511)	33.1473 (0.0867)	0.1420 (0.0169)	225.1574 (0.2899)
	$\delta_{5\mathbb{1}_p}$	0.1470 (0.0072)	0.1957 (0.0135)	0.7740 (0.0678)	198.0822 (0.6326)	36.5264 (0.0924)	63.8083 (22.3797)	224.6905 (0.0994)
0.2	$N(5\mathbb{1}_p, 5I_p)$	0.1620 (0.0102)	0.2000 (0.0157)	0.7285 (0.0862)	39.3082 (0.2273)	81.2740 (0.1081)	0.1420 (0.0107)	400.3554 (0.4956)
	$\delta_{5\mathbb{1}_p}$	0.1632 (0.0206)	0.1938 (0.0155)	0.9365 (0.0788)	235.5581 (0.8650)	92.1920 (0.1281)	81.8603 (0.1722)	399.3472 (0.1132)

Table 4: Simulation results with $n = 50,000, p = 100$ and $\epsilon \in \{0.02, 0.05, 0.1, 0.2\}$. $G_1(Z; A) = AZ$, and $G_2(U, z; A, w_g) = g_{w_g}(z)AU$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

---

2. We apply the function cov.rob in the R package MASS. However, the function cannot be applied when some observations are linearly dependent. We add a small random perturbation $N(0, 0.01)$ to the original data before applying cov.rob.

23

n	Q	$G_1(Z;A)$	$G_2(U,z;A,w_g)$	Dimension Halving	Tyler's M-estimator	Kendall's τ	MVE	Sample Covariance
5,000	$N(5\mathbb{1}_p,5I_p)$	1.0186 (0.0677)	0.6449 (0.0576)	1.0813 (0.0593)	38.6494 (0.4477)	81.0570 (0.4043)	0.4724 (0.0623)	400.5247 (0.8093)
	$\delta_{5\mathbb{1}_p}$	0.9142 (0.0717)	0.6152 (0.0751)	1.1796 (0.1257)	233.2116 (2.9582)	92.0669 (0.2453)	82.1385 (0.3717)	399.5117 (0.3461)
10,000	$N(5\mathbb{1}_p,5I_p)$	0.3465 (0.0722)	0.5852 (0.0816)	0.9489 (0.0563)	39.1785 (0.5876)	81.1670 (0.2038)	0.3225 (0.0350)	400.5247 (0.8093)
	$\delta_{5\mathbb{1}_p}$	0.3293 (0.0166)	0.4244 (0.0632)	1.0558 (0.0796)	235.6172 (1.8062)	92.1067 (0.3140)	82.1225 (0.4402)	399.2921 (0.3031)
20,000	$N(5\mathbb{1}_p,5I_p)$	0.2382 (0.0084)	0.2867 (0.0183)	0.7934 (0.0527)	39.1790 (0.2643)	81.1848 (0.1246)	0.2198 (0.0162)	400.5247 (0.8093)
	$\delta_{5\mathbb{1}_p}$	0.2459 (0.0151)	0.2835 (0.0287)	0.9463 (0.0793)	234.2984 (1.8534)	92.1426 (0.2296)	81.9923 (0.3232)	399.4243 (0.1186)
50,000	$N(5\mathbb{1}_p,5I_p)$	0.1547 (0.0082)	0.1866 (0.0211)	0.8185 (0.0847)	39.2105 (0.2365)	81.2640 (0.0727)	0.1447 (0.0127)	400.5247 (0.8093)
	$\delta_{5\mathbb{1}_p}$	0.1580 (0.0101)	0.1947 (0.0201)	0.8988 (0.0821)	235.0871 (0.9106)	92.1250 (0.1733)	82.0914 (0.2233)	399.5099 (0.1755)

Table 5: Simulation results with $p = 100, \epsilon = 0.2$ and $n \in \{5000, 10000, 20000, 50000\}$. $G_1(Z;A) = AZ$, and $G_2(U,z;A,w_g) = g_{w_g}(z)AU$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

p	Q	$G_1(Z;A)$	$G_2(U,z;A,w_g)$	Dimension Halving	Tyler's M-estimator	Kendall's τ	MVE	Sample Covariance
10	$N(5\mathbb{1}_p,5I_p)$	0.0498 (0.0116)	0.0541 (0.0139)	0.5800 (0.0268)	2.8370 (0.0345)	8.1053 (0.0540)	0.0475 (0.0133)	40.4843 (0.1282)
	$\delta_{5\mathbb{1}_p}$	0.0519 (0.0093)	0.1072 (0.0424)	0.5587 (0.0236)	10.8566 (0.0327)	9.2230 (0.0838)	6.3590 (0.0458)	39.4842 (0.0345)
25	$N(5\mathbb{1}_p,5I_p)$	0.0813 (0.0127)	0.1059 (0.0106)	0.6090 (0.0256)	8.7515 (0.0373)	20.2953 (0.0788)	0.0637 (0.0084)	100.4339 (0.2658)
	$\delta_{5\mathbb{1}_p}$	0.0813 (0.0089)	0.1188 (0.0269)	0.6010 (0.0261)	43.3639 (0.2047)	23.0388 (0.1104)	18.8767 (0.0685)	99.4246 (0.0432)
50	$N(5\mathbb{1}_p,5I_p)$	0.1171 (0.0111)	0.1296 (0.0089)	0.6238 (0.0197)	18.9115 (0.1781)	40.6553 (0.0731)	0.1034 (0.0081)	200.4463 (0.2247)
	$\delta_{5\mathbb{1}_p}$	0.1044 (0.0090)	0.1444 (0.0174)	0.6420 (0.0438)	125.5211 (0.5539)	46.0510 (0.0870)	40.0609 (0.2258)	199.4370 (0.0693)
100	$N(5\mathbb{1}_p,5I_p)$	0.1538 (0.0052)	0.1893 (0.0131)	0.6406 (0.0211)	39.2077 (0.1659)	81.1823 (0.0865)	0.1454 (0.0147)	400.4037 (0.2775)
	$\delta_{5\mathbb{1}_p}$	0.1552 (0.0087)	0.1957 (0.0150)	0.6385 (0.0279)	235.5848 (1.2379)	92.0734 (0.0988)	82.0031 (0.2379)	399.3505 (0.1138)
200	$N(5\mathbb{1}_p,5I_p)$	0.2154 (0.0066)	0.2630 (0.0125)	0.6824 (0.0274)	80.0833 (0.5008)	162.3562 (0.1491)	0.2115 (0.0100)	800.2610 (0.4365)
	$\delta_{5\mathbb{1}_p}$	0.2159 (0.0057)	0.2559 (0.0099)	0.6628 (0.0206)	731.8936 (1.5718)	184.2577 (0.1533)	166.0049 (0.3709)	799.4495 (0.1780)

Table 6: Simulation results with $n = 50,000, \epsilon = 0.2$ and $p \in \{10, 25, 50, 100, 200\}$. $G_1(Z;A) = AZ$, and $G_2(U,z;A,w_g) = g_{w_g}(z)AU$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

The results are summarized in Tables 4-6. As ϵ gets larger, Table 4 shows that the errors of Tyler's M-estimator, Kendall's τ and MVE all grow significantly, while the dependence on ϵ for the other three estimators are very mild. The same pattern also appears in Table 6 as the dimension p grows. For Table 5, we observe that the errors of the two GANs and dimension halving all decrease as n grows. In contrast, the errors of Tyler's M-estimator and Kendall's τ almost stay as constant when n varies because of the contamination. MVE exhibits different behaviors for the two contamination distributions. Its error decreases as n grows when the contamination distribution is $N(5\mathbb{1}_p, 5I_p)$ and stays as constant when the contamination distribution is $\delta_{5\mathbb{1}_p}$.

In summary, our numerical results show that Tyler's M-estimator and Kendall's τ do not work well under Huber's contamination model. MVE works well for certain contamination distributions but is certainly not robust against all contamination distributions. The GANs proposed in the paper and dimension halving all work very well because they are designed for robust estimation under Huber's contamination model. Among the three, the GANs constantly have smaller errors than dimension halving. This is not surprising given the minimax optimality of the GANs under Huber's contamination model established in this paper.

6.3 Simultaneous Estimation of Location and Scatter

In this section, we study robust simultaneous location and scatter estimation. Our goal is to compare the performances of the four generator networks G_1, G_2, G_3 and G_4. Table 7 summarizes the numerical results under four different settings of contamination models. We observe that the networks G_3 and G_4 have similar performances in terms of estimating the scatter as the networks G_1 and G_2 that only estimate the scatter. We can also compare G_2 with G_1 and G_4 with G_3, since G_1 and G_3 assume the knowledge of the distribution family, while G_2 and G_4 estimate the distribution via the additional $g_{w_g}(z)$. Even though the additional knowledge of the distribution family does help G_1 and G_3 to outperform G_2 and G_4, the advantage is not significant in terms of estimating the scatter and is negligible in terms of location estimation. These observations imply that the most complicated network G_4 works very well for adaptive estimation under the general elliptical distribution family.

(P, Q)	$G_1(z; A) = Az$	$G_3(z; A, \mu) = Az + \mu$		$G_2(u, z; A, w_g) = g_{w_g}(z)Au$	$G_4(u, z; A, w_g, \mu) = g_{w_g}(z)Au + \mu$	
	$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$	$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$	$\|\widehat{\theta} - \theta\|$	$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$	$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$	$\|\widehat{\theta} - \theta\|$
$(N(0, I_p), N(5, 5I_p))$	0.1615 (0.0134)	0.1537 (0.0155)	0.0508 (0.0054)	0.1624 (0.0141)	0.1694 (0.0105)	0.0519 (0.0048)
$(N(0, \Sigma_{\mathrm{ar}}), \delta_{4\mathbb{1}_p})$	0.1530 (0.0059)	0.1640 (0.0106)	0.0547 (0.0039)	0.1557 (0.0142)	0.1880 (0.0134)	0.0544 (0.0073)
$(T_1(0, \Sigma_{\mathrm{ar}}), T_1(5, 5I_p))$	0.2808 (0.0440)	0.2512 (0.0479)	0.0656 (0.0065)	0.3350 (0.0681)	0.4678 (0.0498)	0.0575 (0.0048)
$(T_2(0, \Sigma_{\mathrm{ar}}), T_2(5, 5I_p))$	0.3450 (0.0157)	0.3743 (0.0097)	0.0640 (0.0056)	0.4059 (0.0254)	0.4704 (0.0299)	0.0642 (0.0040)

Table 7: Simulation results with i.i.d. observations generated from $(1 - \epsilon)P + \epsilon Q$, where $n = 50,000, p = 100$ and $\epsilon = 0.2$. We show the average errors $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ and $\|\widehat{\theta} - \theta\|$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

6.4 Comparisons of Proper Scoring Rules

As we have shown in Section 2.5, JS-GAN can be understood as a special case of a more general class Beta-GAN(α, β) with $\alpha, \beta > -1$. Our theoretical results are valid for any Beta-GAN(α, β) as long as $|\alpha - \beta| < 1$ (Condition 6). In this section, we study the performance of this wide class of GANs with various choices of (α, β). Table 8 summarizes the numerical results under four different settings of contamination models. The comparison includes JS-GAN and LS-GAN, which correspond to $(\alpha, \beta) = (0, 0)$ and $(\alpha, \beta) = (1, 1)$. In addition, we also consider $(\alpha, \beta) \in \{(-0.5, -0.5), (0.5, 0.5), (0.5, 1), (1, 0.5), (2, 2), (4, 4)\}$. We observe that the GANs with $0 \leq \alpha, \beta \leq 1$ all have very similar performances. On the other hand, the errors increase as the values of α and β grow, which is shown in the last two columns of Table 8 for $(\alpha, \beta) \in \{(2, 2), (4, 4)\}$. In fact, our additional experiments show that the performance of Beta-GAN(α, β) is not acceptable anymore as soon as $\alpha, \beta \geq 8$. This may be caused by the bad landscape of the objective function for large α and β. In fact, as $\alpha = \beta \to \infty$, we recover TV-GAN, which is known to have a bad landscape for robust estimation (Gao et al., 2019). The boosting score, which corresponds to $(\alpha, \beta) = (-0.5, -0.5)$, leads to worse errors than JS-GAN and LS-GAN.

We also note that some asymmetric GANs (e.g. $(\alpha, \beta) = (1, 0.5)$) have better performance than JS-GAN and LS-GAN. It is interesting to further explore the properties of different scores from both theoretical and experimental perspectives in the future work.

(P, Q, p)	Beta(-0.5,-0.5)	JS-GAN	Beta(0.5, 0.5)	Beta(0.5,1)	Beta(1,0.5)	LS-GAN	Beta(2,2)	Beta(4,4)
$(N(0, I_p), N(5, 5I_p), 100)$	0.1557 (0.0093)	0.1188 (0.0046)	0.1228 (0.0045)	0.1201 (0.0033)	0.1040 (0.0017)	0.1283 (0.0095)	0.1402 (0.0063)	0.3478 (0.0035)
$(N(0, I_p), N(5, 5I_p), 200)$	0.3346 (0.0149)	0.1720 (0.0032)	0.1677 (0.0045)	0.1697 (0.0054)	0.1599 (0.0026)	0.1749 (0.0048)	0.1978 (0.0031)	0.3508 (0.0034)
$(T_2(0, \Sigma_{\mathrm{ar}}), T_2(5, 5I_p), 100)$	0.5653 (0.5065)	0.1848 (0.0106)	0.1941 (0.0087)	0.2016 (0.0190)	0.1925 (0.0149)	0.1882 (0.0152)	0.3371 (0.0378)	0.9689 (0.1124)
$(T_4(0, \Sigma_{\mathrm{ar}}), T_4(5, 5I_p), 100)$	0.2726 (0.0083)	0.2009 (0.0079)	0.1923 (0.0125)	0.2133 (0.0105)	0.1758 (0.0122)	0.1999 (0.0130)	0.3334 (0.0213)	0.7740 (0.0432)

Table 8: Simulation results with i.i.d. observations generated from $(1 - \epsilon)P + \epsilon Q$, where $n = 50,000, p \in \{100, 200\}$ and $\epsilon = 0.2$. We show the average error $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in each cell with standard deviation in parenthesis from 10 repeated experiments.

7. Proofs

7.1 Some Lemmas

Before proving the main results, we introduce some lemmas, whose proofs are given in Section 7.6.

Lemma 13 *Given i.i.d. observations $X_1, ..., X_n \sim \mathbb{P}$, and $\mathcal{T} \in \{\mathcal{T}_1, \mathcal{T}_3\}$ defined by either (14) or (19), we have for any $\delta > 0$,*

$$\sup_{T \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^{n} \log T(X_i) - \mathbb{E} \log T(X) \right| \leq C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right),$$

with probability at least $1 - \delta$ for some universal constant $C > 0$.

Lemma 14 *Given i.i.d. observations $X_1, ..., X_n \sim N(0, \Sigma)$ and the function class $\mathcal{T}_2$ defined by (15). Assume $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$. We have for any $\delta > 0$,*

$$\sup_{T \in \mathcal{T}_2} \left| \frac{1}{n} \sum_{i=1}^{n} \log T(X_i) - \mathbb{E} \log T(X) \right| \leq C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right), \tag{27}$$

with probability at least $1 - \delta$ for some universal constant $C > 0$.

Lemma 15 *Given i.i.d. observations $X_1, ..., X_n \sim \mathbb{P}$ and the function class $\mathcal{T}_3$ defined by (19). Assume $\{S(\cdot, 1), S(\cdot, 0)\}$ is a regular proper scoring rule that satisfies Condition 6 and $\kappa \leq c$ for some sufficiently small constant $c > 0$. We have for any $\delta > 0$,*

$$\sup_{T \in \mathcal{T}_3} \left| \frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) - \mathbb{E}S(T(X), 1) \right| \leq C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right),$$

with probability at least $1 - \delta$ for some universal constant $C > 0$.

Lemma 16 *Given i.i.d. observations $X_1, ..., X_n \sim N(0, \Sigma)$ and the function class $\mathcal{T}_4$ defined by (20). Assume $\{S(\cdot, 1), S(\cdot, 0)\}$ is a regular proper scoring rule that satisfies Condition 6, $\kappa_1 \leq c$ for some sufficiently small constant $c > 0$, and $\|\Sigma\|_{\mathrm{op}} \leq M = O(1)$. We have for any $\delta > 0$,*

$$\sup_{T \in \mathcal{T}_4} \left| \frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) - \mathbb{E}S(T(X), 1) \right| \leq C\kappa_1\kappa_2 \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right), \qquad (28)$$

with probability at least $1 - \delta$ for some universal constant $C > 0$.

Lemma 17 *Given i.i.d. observations $X_1, ..., X_n \sim \mathbb{P}$. Assume $\{S(\cdot, 1), S(\cdot, 0)\}$ is a regular proper scoring rule that satisfies Condition 6 and $\kappa \leq c$ for some sufficiently small constant $c > 0$. We have for any $\delta > 0$,*

$$\sup_{T \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) - \mathbb{E}S(T(X), 1) \right| \leq C\kappa(2B)^{L-1} \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right),$$

for both $\mathcal{T} = \mathcal{T}^L(\kappa, B)$ and $\mathcal{T} = \bar{\mathcal{T}}^L(\kappa, B)$ with probability at least $1 - \delta$ for some universal constant $C > 0$.

7.2 Proofs of Proposition 3 and Proposition 5

Proof [Proof of Proposition 3] Define

$$F(w, u; \Sigma, \Gamma) = \mathbb{E}_{X \sim N(0,\Sigma)} \log T_{w,u}(X) + \mathbb{E}_{X \sim N(0,\Gamma)} \log(1 - T_{w,u}(X)),$$

where $T_{w,u}(x) = \mathsf{sigmoid}\left(\sum_{j \geq 1} w_j \mathsf{sigmoid}(u_j^T x) \right)$. Then, we have

$$F(\Sigma, \Gamma) = \max_{\|w\|_1 \leq \kappa, u_j \in \mathbb{R}^p} F(w, u; \Sigma, \Gamma).$$

We calculate the gradient and Hessian of $F(w, u; \Sigma, \Gamma)$ with respect to w. To do this, we define $g_u(x)$ to be a vector with the same dimension as w and each of its coordinate takes $\mathsf{sigmoid}(u_j^T x)$. With this notation, we can write $T_{w,u}(x) = \mathsf{sigmoid}(w^T g_u(x))$. By standard calculation, we get

$$
\begin{aligned}
\nabla_w F(w, u; \Sigma, \Gamma) &= \mathbb{E}_{X \sim N(0,\Sigma)} \left(1 - T_{w,u}(X) \right) g_u(X) - \mathbb{E}_{X \sim N(0,\Gamma)} T_{w,u}(X) g_u(X), \\
\nabla_w^2 F(w, u; \Sigma, \Gamma) &= -\mathbb{E}_{X \sim N(0,\Sigma)} T_{w,u}(X) \left(1 - T_{w,u}(X) \right) g_u(X) g_u(X)^T \\
&\quad -\mathbb{E}_{X \sim N(0,\Gamma)} T_{w,u}(X) \left(1 - T_{w,u}(X) \right) g_u(X) g_u(X)^T.
\end{aligned}
$$

For any $u_j \in \mathbb{R}^p$,

$$
\begin{aligned}
\mathbb{E}_{X \sim N(0,\Sigma)} \mathsf{sigmoid}(u_j^T X) &= \mathbb{E}_{X \sim N(0,\Sigma)} \mathsf{sigmoid}(-u_j^T X) \\
&= \mathbb{E}_{X \sim N(0,\Sigma)} (1 - \mathsf{sigmoid}(u_j^T X)),
\end{aligned}
$$

27

which immediately implies $\mathbb{E}_{X \sim N(0,\Sigma)}\mathsf{sigmoid}(u_j^T X) = 1/2$. By the same argument, we also have $\mathbb{E}_{X \sim N(0,\Gamma)}\mathsf{sigmoid}(u_j^T X) = 1/2$. Therefore,

$$\nabla_w F(w, u; \Sigma, \Gamma)|_{w=0} = \frac{1}{2}\mathbb{E}_{X \sim N(0,\Sigma)}g_u(X) - \frac{1}{2}\mathbb{E}_{X \sim N(0,\Gamma)}g_u(X) = 0.$$

Moreover, since $-\nabla_w^2 F(w, u; \Sigma, \Gamma)$ is positive semi-definite, $F(w, u; \Sigma, \Gamma)$ is a concave function in w. This implies

$$\max_{\|w\|_1 \le \kappa} F(w, u; \Sigma, \Gamma) = F(0, u; \Sigma, \Gamma) = -\log 4.$$

Taking maximum over u, we have $F(\Sigma, \Gamma) = -\log 4$, regardless of the values of Σ and Γ. ∎

Proof [Proof of Proposition 5] We first introduce some notation. We define

$$J(\gamma^2) = \max_{\|w\|_1 \le \kappa, u_j \in \mathbb{R}} F(w, u; \gamma^2),$$

where

$$F(w, u; \gamma^2) = \mathbb{E}_{X \sim (1-\epsilon)N(0,\sigma^2)+\epsilon N(0,\tau^2)} \log T_{w,u}(X) + \mathbb{E}_{X \sim N(0,\gamma^2)} \log(1 - T_{w,u}(X)),$$

with $T_{w,u}(x)$ defined by $T_{w,u}(x) = \mathsf{sigmoid}\left(\sum_{j \ge 1} w_j \mathsf{ReLU}(u_j x)\right)$. With these notation, we need to prove $J([(1-\epsilon)\sigma + \epsilon\tau]^2) = \min_{\gamma^2} J(\gamma^2)$.

The gradient and Hessian of $F(w, u; \gamma^2)$ with respect to w are given by

$$
\begin{aligned}
\nabla_w F(w, u; \gamma^2) &= \mathbb{E}_{X \sim (1-\epsilon)N(0,\sigma^2)+\epsilon N(0,\tau^2)} \left(1 - T_{w,u}(X)\right) g_u(X) \\
&\quad -\mathbb{E}_{X \sim N(0,\gamma^2)} T_{w,u}(X) g_u(X), \\
\nabla_w^2 F(w, u; \gamma^2) &= -\mathbb{E}_{X \sim (1-\epsilon)N(0,\sigma^2)+\epsilon N(0,\tau^2)} T_{w,u}(X) \left(1 - T_{w,u}(X)\right) g_u(X) g_u(X)^T \\
&\quad -\mathbb{E}_{X \sim N(0,\gamma^2)} T_{w,u}(X) \left(1 - T_{w,u}(X)\right) g_u(X) g_u(X)^T,
\end{aligned}
$$

where we use $g_u(X)$ to denote the vector whose jth coordinate is $\mathsf{ReLU}(u_j X)$. It is not hard to see that

$$\mathbb{E}_{X \sim N(0,\gamma^2)}\mathsf{ReLU}(u_j X) = |u_j|\gamma\sqrt{\frac{2}{\pi}},$$

and

$$\mathbb{E}_{X \sim (1-\epsilon)N(0,\sigma^2)+\epsilon N(0,\tau^2)}\mathsf{ReLU}(u_j X) = |u_j|[(1 - \epsilon)\sigma + \epsilon\tau]\sqrt{\frac{2}{\pi}}.$$

Therefore, we have the identity

$$\mathbb{E}_{X \sim N(0,[(1-\epsilon)\sigma+\epsilon\tau]^2)}\mathsf{ReLU}(u_j X) = \mathbb{E}_{X \sim (1-\epsilon)N(0,\sigma^2)+\epsilon N(0,\tau^2)}\mathsf{ReLU}(u_j X),$$

which is equivalent to $\nabla_w F(w, u; [(1-\epsilon)\sigma+\epsilon\tau]^2)|_{w=0} = 0$. Moreover, since $-\nabla_w^2 F(w, u; [(1-\epsilon)\sigma+\epsilon\tau]^2)$ is positive semi-definite, $F(w, u; [(1-\epsilon)\sigma+\epsilon\tau]^2)$ is a concave function in w. This implies

$$\max_{\|w\|_1 \le \kappa} F(w, u; [(1-\epsilon)\sigma+\epsilon\tau]^2) = F(0, u; [(1-\epsilon)\sigma+\epsilon\tau]^2) = -\log 4.$$

Taking maximum over u, we have $J([(1 - \epsilon)\sigma + \epsilon\tau]^2) = -\log 4$. For any γ^2, $J(\gamma^2) \geq F(0, u; \gamma^2) = -\log 4$. Hence, we have the desired conclusion that $J([(1 - \epsilon)\sigma + \epsilon\tau]^2) = \min_{\gamma^2} J(\gamma^2)$. ∎

7.3 Proofs of Proposition 2 and Proposition 4

Proof [Proof of Proposition 2] We use the notation $F_{w,u}(P, Q) = \mathbb{E}_{X \sim P} \log T_{w,u}(X) + \mathbb{E}_{X \sim Q} \log(1 - T_{w,u}(X))$, with $T_{w,u}(x) = \mathsf{sigmoid}\left(\sum_{j \geq 1} w_j \mathsf{sigmoid}(u_j^T x)\right)$. Thus, we can write

$$\widehat{\theta} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \, F(\mathbb{P}_n, N(\eta, I_p)), \tag{29}$$

where $F(P, Q) = \max_{\|w\|_1 \leq \kappa, u_j \in \mathbb{R}^p} F_{w,u}(P, Q)$, and $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ is the empirical measure. Let P be the data generating process that satisfies $\mathsf{TV}(P, N(\theta, I_p)) \leq \epsilon$, and then there exist probability distributions Q_1 and Q_2, such that

$$P + \epsilon Q_1 = N(\theta, I_p) + \epsilon Q_2.$$

The explicit construction of Q_1, Q_2 is given in the proof of Theorem 5.1 of Chen et al. (2018). This implies that

$$
\begin{aligned}
&|F(P, N(\eta, I_p)) - F(N(\theta, I_p), N(\eta, I_p))| \\
\leq \quad & \sup_{\|w\|_1 \leq \kappa, u} |F_{w,u}(P, N(\eta, I_p)) - F_{w,u}(N(\theta, I_p), N(\eta, I_p))| \\
= \quad & \epsilon \sup_{\|w\|_1 \leq \kappa, u} |\mathbb{E}_{X \sim Q_2} \log(2 T_{w,u}(X)) - \mathbb{E}_{X \sim Q_1} \log(2 T_{w,u}(X))| \\
\leq \quad & 2\kappa\epsilon. \tag{30}
\end{aligned}
$$

Then,

$$
\begin{aligned}
F(N(\theta, I_p), N(\widehat{\theta}, I_p)) \quad &\leq \quad F(P, N(\widehat{\theta}, I_p)) + 2\kappa\epsilon && (31) \\
&\leq \quad F(\mathbb{P}_n, N(\widehat{\theta}, I_p)) + C\kappa\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 2\kappa\epsilon && (32) \\
&\leq \quad F(\mathbb{P}_n, N(\theta, I_p)) + C\kappa\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 2\kappa\epsilon && (33) \\
&\leq \quad F(P, N(\theta, I_p)) + 2C\kappa\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 2\kappa\epsilon && (34) \\
&\leq \quad F(N(\theta, I_p), N(\theta, I_p)) + 2C\kappa\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4\kappa\epsilon. && (35)
\end{aligned}
$$

The inequalities (31) and (35) are direct consequences of (30). We have used Lemma 13 for (32) and (34), and (33) is implied by the definition of the estimator (29). By the definition

of $F(N(\theta, I_p), N(\widehat{\theta}, I_p)) - F(N(\theta, I_p), N(\theta, I_p))$, we obtain the following inequality that

$$F_{w,u}(N(\theta, I_p), N(\widehat{\theta}, I_p)) + \log 4 \leq 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4\kappa\epsilon,$$

uniformly over all $\|w\|_1 \leq \kappa$ and all u with probability at least $1 - \delta$. Choose $w_1 = \kappa$, $w_j = 0$ for all $j \geq 2$ and $u_1 = v$ for some unit vector $\|v\| = 1$, and then we have

$$f(\kappa; v^T\theta, v^T\widehat{\theta}) \leq 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4\kappa\epsilon, \tag{36}$$

where

$$f(t; \delta_1, \delta_2) = \mathbb{E} \log \frac{2}{1 + e^{-t\mathsf{sigmoid}(Z+\delta_1)}} + \mathbb{E} \log \frac{2}{1 + e^{t\mathsf{sigmoid}(Z+\delta_2)}},$$

with $Z \sim N(0, 1)$. Direct calculations give

$$\begin{aligned}
\frac{\partial}{\partial t} f(t; \delta_1, \delta_2) &= \mathbb{E} \frac{1}{1 + e^{t\mathsf{sigmoid}(Z+\delta_1)}} \mathsf{sigmoid}(Z + \delta_1) \\
&\quad -\mathbb{E} \frac{1}{1 + e^{-t\mathsf{sigmoid}(Z+\delta_2)}} \mathsf{sigmoid}(Z + \delta_2), \\
\frac{\partial^2}{\partial t^2} f(t; \delta_1, \delta_2) &= -\mathbb{E} \frac{e^{t\mathsf{sigmoid}(Z+\delta_1)}}{(1 + e^{t\mathsf{sigmoid}(Z+\delta_1)})^2} |\mathsf{sigmoid}(Z + \delta_1)|^2 \\
&\quad -\mathbb{E} \frac{e^{t\mathsf{sigmoid}(Z+\delta_2)}}{(1 + e^{t\mathsf{sigmoid}(Z+\delta_2)})^2} |\mathsf{sigmoid}(Z + \delta_2)|^2.
\end{aligned}$$

Therefore, we have $f(0; \delta_1, \delta_2) = 0$, $\frac{\partial}{\partial t} f(t; \delta_1, \delta_2)|_{t=0} = \frac{1}{2} \left(\mathbb{E}\mathsf{sigmoid}(Z + \delta_1) - \mathbb{E}\mathsf{sigmoid}(Z + \delta_2) \right)$, and $\frac{\partial^2}{\partial t^2} f(t; \delta_1, \delta_2) \geq -\frac{1}{2}$, which then implies

$$f(\kappa; \delta_1, \delta_2) \geq f(0; \delta_1, \delta_2) + \kappa \frac{\partial}{\partial t} f(t; \delta_1, \delta_2)|_{t=0} - \frac{1}{4}\kappa^2.$$

or equivalently $\kappa \frac{\partial}{\partial t} f(t; \delta_1, \delta_2)|_{t=0} \leq f(\kappa; \delta_1, \delta_2) + \frac{1}{4}\kappa^2$. In view of the bound (36), we have

$$\begin{aligned}
& \frac{\kappa}{2} \int \left(\mathsf{sigmoid}(z + v^T\theta) - \mathsf{sigmoid}(z + v^T\widehat{\theta}) \right) \phi(z)dz \\
\leq \ & 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4\kappa\epsilon + \frac{1}{4}\kappa^2,
\end{aligned}$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. A symmetric argument with $u_1 = -v$ leads to the same bound for $\frac{\kappa}{2} \int \left(\mathsf{sigmoid}(z + v^T\widehat{\theta}) - \mathsf{sigmoid}(z + v^T\theta) \right) \phi(z)dz$, and thus

$$\begin{aligned}
& \left| \int \left(\mathsf{sigmoid}(z + v^T\widehat{\theta}) - \mathsf{sigmoid}(z + v^T\theta) \right) \phi(z)dz \right| \\
\leq \ & 2C \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4\epsilon + \frac{1}{4}\kappa.
\end{aligned}$$

Define the following function $h(t) = \int \mathsf{sigmoid}(z + v^T\theta + t)\phi(z)dz$, take $\kappa = O(\sqrt{p/n} + \epsilon)$ and $\delta = e^{C'(p + n\epsilon^2)}$, and then the above bound becomes $|h(v^T(\widehat{\theta} - \theta)) - h(0)| = O(\sqrt{p/n} + \epsilon)$. It is easy to see that $h'(0) \geq \int \mathsf{sigmoid}(z + M)(1 - \mathsf{sigmoid}(z + M))\phi(z)dz$, which is a constant, and the continuity of $h'(t)$ implies that there are small constants $c_1, c_2 > 0$, such that $\inf_{|t| \leq c_1} |h'(t)| \geq c_2$. Thus, as long as $|h(t) - h(0)|$ is sufficiently small, we have $|h(t) - h(0)| \geq c_2|t|$, which implies that $|v^T(\widehat{\theta} - \theta)| = O(\sqrt{p/n} + \epsilon)$. The proof is complete by taking supreme over all unit vector v. ∎

Proof [Proof of Proposition 4] We use the same notation $F_{w,u}(P, Q)$ defined in the proof of Proposition 2, but $T_{w,u}(x) = \mathsf{sigmoid}(\sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T x))$. Then, by the same argument used in (31)-(35) (with $\epsilon = 0$ and Lemma 13 replaced by Lemma 14), we have

$$F_{w,u}(N(0, \Sigma), N(0, \widehat{\Sigma})) + \log 4 \leq 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right),$$

uniformly over all $\|w\|_1 \leq \kappa$ and all $\|u_j\| \leq 1$ with probability at least $1 - \delta$. Choose $w_1 = w_2 = \kappa/2$, $w_j = 0$ for all $j \geq 3$, $u_1 = v$ and $u_2 = -v$ for some unit vector $\|v\| = 1$, and then we have

$$f\left(\kappa; \sqrt{v^T\Sigma v}, \sqrt{v^T\widehat{\Sigma}v}\right) \leq 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right), \tag{37}$$

where

$$f(t; \delta_1, \delta_2) = \mathbb{E}\log\frac{2}{1 + e^{-\frac{t}{2}\delta_1|Z|}} + \mathbb{E}\log\frac{2}{1 + e^{\frac{t}{2}\delta_2|Z|}},$$

with $Z \sim N(0, 1)$. Then,

$$\frac{\partial}{\partial t}f(t; \delta_1, \delta_2) = \frac{1}{2}\mathbb{E}\frac{1}{1 + e^{t\delta_1|Z|/2}}\delta_1|Z| - \frac{1}{2}\mathbb{E}\frac{1}{1 + e^{-t\delta_2|Z|/2}}\delta_2|Z|,$$

$$\frac{\partial^2}{\partial t^2}f(t; \delta_1, \delta_2) = -\frac{1}{4}\mathbb{E}\frac{e^{t\delta_1|Z|/2}}{(1 + e^{t\delta_1|Z|/2})^2}|\delta_1 Z|^2 - \frac{1}{4}\mathbb{E}\frac{e^{t\delta_2|Z|/2}}{(1 + e^{t\delta_2|Z|/2})^2}|\delta_2 Z|^2.$$

Therefore, we have $f(0; \delta_1, \delta_2) = 0$, $\frac{\partial}{\partial t}f(t; \delta_1, \delta_2)|_{t=0} = \frac{1}{4}\mathbb{E}|Z|(\delta_1 - \delta_2)$, and

$$\frac{\partial^2}{\partial t^2}f\left(t; \sqrt{v^T\Sigma v}, \sqrt{v^T\widehat{\Sigma}v}\right) \geq -\frac{M}{8},$$

which implies

$$f(\kappa; \delta_1, \delta_2) \geq f(0; \delta_1, \delta_2) + \kappa\frac{\partial}{\partial t}f(t; \delta_1, \delta_2)|_{t=0} - \frac{M}{16}\kappa^2.$$

Then, by the bound (37), we have $\sqrt{v^T\Sigma v} - \sqrt{v^T\widehat{\Sigma}v} \lesssim \sqrt{\frac{p}{n}}$ by taking $\kappa = O(\sqrt{p/n})$ and $\delta = e^{C'p}$. A symmetric argument also leads to the same bound for $\sqrt{v^T\widehat{\Sigma}v} - \sqrt{v^T\Sigma v}$, so that $\left|\sqrt{v^T\widehat{\Sigma}v} - \sqrt{v^T\Sigma v}\right| \lesssim \sqrt{\frac{p}{n}}$, which is equivalent to $|v^T(\widehat{\Sigma} - \Sigma)v| \lesssim \left(\sqrt{v^T\widehat{\Sigma}v} + \sqrt{v^T\Sigma v}\right)\sqrt{\frac{p}{n}} \leq 2\sqrt{M}\sqrt{\frac{p}{n}}$. The proof is complete by taking supreme over all unit vector v. ∎

7.4 Proofs of Theorem 7 and Theorem 8

Proof [Proof of Theorem 7] We use the notation

$$F_{w,u,b}(P,Q) = \mathbb{E}_{X \sim P} S(T_{w,u,b}(X), 1) + \mathbb{E}_{X \sim Q} S(T_{w,u,b}(X), 0),$$

with $T_{w,u,b}(x) = \mathsf{sigmoid}(\sum_{j \geq 1} w_j \mathsf{sigmoid}(u_j^T x + b_j))$. We also write

$$F(P,Q) = \max_{\|w\|_1 \leq \kappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R}} F_{w,u,b}(P,Q).$$

For P that satisfies $\mathsf{TV}(P, N(0, \Sigma)) \leq \epsilon$, there exist probability distributions Q_1 and Q_2, such that $P + \epsilon Q_1 = N(0, \Sigma) + \epsilon Q_2$. Therefore,

$$
\begin{aligned}
&|F(P, N(0, \Gamma)) - F(N(0, \Sigma), N(0, \Gamma))| \\
\leq\ & \max_{\|w\|_1 \leq \kappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R}} |F_{w,u,b}(P, N(0, \Gamma)) - F_{w,u,b}(N(0, \Sigma), N(0, \Gamma))| \\
=\ & \epsilon \max_{\|w\|_1 \leq \kappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R}} |\mathbb{E}_{X \sim Q_2} S(T_{w,u,b}(X), 1) - \mathbb{E}_{X \sim Q_1} S(T_{w,u,b}(X), 1)| \\
\leq\ & 2\epsilon \sup_{|t - 1/2| \leq \kappa} \left| S(t, 1) - G\left(\frac{1}{2}\right) - \frac{1}{2} G'\left(\frac{1}{2}\right) \right| \\
\leq\ & 2\epsilon\kappa \sup_{|t - \frac{1}{2}| \leq \kappa} \left| (1 - t) G''(t) \right| \leq 2C_1 \epsilon\kappa,
\end{aligned}
$$

where we have used $\sup_{|t - \frac{1}{2}| \leq \kappa} |(1-t) G''(t)| \leq C_1$ because of the smoothness of $G(t)$ at $t = 1/2$ by Condition 6. In the second last inequality above, we have used the fact that $\frac{\partial}{\partial t} S(t, 1) = (1 - t) G''(t)$ and $S(t, 1) = G(t) + (1 - t) G'(t)$.

Then, by the same argument used in (31)-(35) (with Lemma 13 replaced by Lemma 15), we have

$$F_{w,u,b}(N(0, \Sigma), N(0, \widehat{\Sigma})) - 2G(1/2) \leq 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4C_1 \epsilon\kappa, \qquad (38)$$

uniformly over all $\|w\|_1 \leq \kappa$, all u_j and all b_j with probability at least $1 - \delta$. We choose $w_1 = \kappa$, $w_j = 0$ for all $j \geq 2$, $u_1 = v/\sqrt{v^T \Sigma v}$ for some unit vector v, and $b_1 = 1$. Then, we have

$$f\left(\kappa; \sqrt{\frac{v^T \widehat{\Sigma} v}{v^T \Sigma v}} \right) \leq 2C\kappa \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4C_1 \epsilon\kappa, \qquad (39)$$

where

$$f(t; \Delta) = \mathbb{E} S\left(\frac{1}{1 + e^{-t\mathsf{sigmoid}(Z-1)}}, 1 \right) + \mathbb{E} S\left(\frac{1}{1 + e^{-t\mathsf{sigmoid}(\Delta Z - 1)}}, 0 \right) - 2G(1/2),$$

with $Z \sim N(0, 1)$. We introduce some polynomials,

$$p(t) = (1-t)^2 t, \quad \bar{p}(t) = 2(1-t)^2 t^2 - (1-t)^3 t, \quad \tilde{p}(t) = (1-t)^3 t^2,$$

$$q(t) = (1-t)t^2, \quad \bar{q}(t) = 2(1-t)^2t^2 - (1-t)t^3, \quad \tilde{q}(t) = (1-t)^2t^3.$$

Then, by standard calculations, we get

$$
\begin{aligned}
\frac{\partial}{\partial t} f(t; \Delta) &= \mathbb{E}p\left(L(t,1,Z)\right) G''\left(L(t,1,Z)\right) \mathsf{sigmoid}(Z-1) \\
&\quad -\mathbb{E}q\left(L(t,\Delta,Z)\right) G''\left(L(t,\Delta,Z)\right) \mathsf{sigmoid}(\Delta Z - 1),
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial^2}{\partial t^2} f(t; \Delta) \\
&= -\mathbb{E}\left[\bar{p}(L(t,1,Z))G''(L(t,1,Z)) - \tilde{p}(L(t,1,Z))G'''(L(t,1,Z)) \right] |\mathsf{sigmoid}(Z-1)|^2 \\
&\quad -\mathbb{E}\left[\bar{q}(L(t,\Delta,Z))G''(L(t,\Delta,Z)) - \tilde{q}(L(t,\Delta,Z))G'''(L(t,\Delta,Z)) \right] |\mathsf{sigmoid}(\Delta Z - 1)|^2,
\end{aligned}
$$

where we use the notation $L(t,\delta,Z) = \mathsf{sigmoid}(t\mathsf{sigmoid}(\delta Z - 1))$. Note that $f(0; \Delta) = 0$ and

$$\frac{\partial}{\partial t} f(t; \Delta)|_{t=0} = \frac{1}{8} G''(1/2)\mathbb{E}\left(\mathsf{sigmoid}(Z-1) - \mathsf{sigmoid}(\Delta Z - 1)\right).$$

Since

$$\bar{p}(1/2)G''(1/2) - \tilde{p}(1/2)G'''(1/2) = \frac{1}{16}G''(1/2) - \frac{1}{32}G'''(1/2) \geq \frac{1}{32}c_0$$

by Condition 6, for a sufficiently small κ, we have

$$\inf_{|t-1/2|\leq\kappa} \left[\bar{p}(t)G''(t) - \tilde{p}(t)G'''(t) \right] > 0.$$

Moreover, there is some constant $C_2 > 0$, such that

$$\sup_{|t-1/2|\leq\kappa} \left[\bar{p}(t)G''(t) - \tilde{p}(t)G'''(t) \right] \leq C_2.$$

For the same reason, we also have

$$0 < \inf_{|t-1/2|\leq\kappa} \left[\bar{q}(t)G''(t) - \tilde{q}(t)G'''(t) \right] \leq \sup_{|t-1/2|\leq\kappa} \left[\bar{q}(t)G''(t) - \tilde{q}(t)G'''(t) \right] \leq C_2.$$

Since $|t| \leq \kappa$ implies $|L(t,\delta,Z) - 1/2| \leq \kappa$, we have

$$\inf_{|t|\leq\kappa} \frac{\partial^2}{\partial t^2} f(t; \Delta) \geq -C_2\mathbb{E}|\mathsf{sigmoid}(Z-1)|^2 - C_2\mathbb{E}|\mathsf{sigmoid}(\Delta Z - 1)|^2 \geq -2C_2.$$

Therefore,

$$f(\kappa; \Delta) \geq f(0; \Delta) + \kappa\frac{\partial}{\partial t} f(t; \Delta)|_{t=0} - C_2\kappa^2.$$

Together with the bound (39), we have

$$
\begin{aligned}
&\frac{1}{8}G''(1/2)\mathbb{E}\left(\mathsf{sigmoid}(Z-1) - \mathsf{sigmoid}(\Delta Z - 1)\right) \\
&\leq 2C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1\kappa + C_2\kappa,
\end{aligned}
$$

with $\Delta = \sqrt{\frac{v^T \widehat{\Sigma} v}{v^T \Sigma v}}$. A symmetric argument with $w_1 = -\kappa$ leads to the same bound for

$$\frac{1}{8} G''(1/2) \mathbb{E} \left(\mathsf{sigmoid}(\Delta Z - 1) - \mathsf{sigmoid}(Z - 1) \right).$$

Thus, with the choice $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon \right)$ and $\delta = e^{-C'(p+n\epsilon^2)}$, we have

$$|h(\Delta) - h(1)| \leq C_3 \left(\sqrt{\frac{p}{n}} + \epsilon \right),$$

where $h(t) = \int \mathsf{sigmoid}(tz - 1)(2\pi)^{-1/2} e^{-z^2/2} dz$. Note that

$$|h'(1)| = \left| \int \mathsf{sigmoid}(z - 1)(1 - \mathsf{sigmoid}(z - 1)) z (2\pi)^{-1/2} e^{-z^2/2} dz \right|$$

is a constant. The continuity of $h'(t)$ implies that there are small constants $c_1, c_2 > 0$, such that $\inf_{|t-1| \leq c_1} |h'(t)| \geq c_2$. Thus, as long as $|h(t) - h(1)|$ is sufficiently small, we have $|h(t) - h(1)| \geq c_2 |t - 1|$, which implies that $\left| \sqrt{v^T \widehat{\Sigma} v} - \sqrt{v^T \Sigma v} \right| \lesssim \sqrt{v^T \Sigma v} \left(\sqrt{\frac{p}{n}} + \epsilon \right) \lesssim \sqrt{\frac{p}{n}} + \epsilon$. Following the last several lines of the proof of Proposition 4, we obtain the desired result. ∎

Proof [Proof of Theorem 8] We use the notation $T_{w,v,u}(x) = \mathsf{sigmoid}\left(\sum_{j \geq 1} w_j \mathsf{sigmoid}\left(\sum_{l=1}^{H} v_{jl} \mathsf{ReLU}(u_l^T x) \right) \right)$, and then define

$$F_{w,v,u}(P, Q) = \mathbb{E}_{X \sim P} S(T_{w,v,u}(X), 1) + \mathbb{E}_{X \sim Q} S(T_{w,v,u}(X), 0).$$

Using Lemma 16 and following the same argument that leads to (38), we have

$$F_{w,v,u}(N(0, \Sigma), N(0, \widehat{\Sigma})) - 2G(1/2) \leq 2C\kappa_1 \kappa_2 \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4C_1 \epsilon \kappa_1,$$

uniformly over all $\|w\|_1 \leq \kappa_1$, all $\|v_j\|_1 \leq \kappa_2$, and all $\|u_l\| \leq 1$. Choose $w_1 = \kappa_1$, $w_j = 0$ for all $j \geq 2$, $v_{11} = v_{12} = \frac{1}{2}$, $v_{1l} = 0$ for all $l \geq 3$, and $u_1 = -u_2 = v$ for a unit vector v. Then, we have

$$f\left(\kappa_1; \sqrt{v^T \Sigma v}, \sqrt{v^T \widehat{\Sigma} v} \right) \leq 2C\kappa_1 \kappa_2 \left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + 4C_1 \epsilon \kappa_1,$$

where

$$f(t; \delta_1, \delta_2) = \mathbb{E} S\left(\frac{1}{1 + e^{-t\mathsf{sigmoid}(\delta_1 |Z|/2)}}, 1 \right) + \mathbb{E} S\left(\frac{1}{1 + e^{-t\mathsf{sigmoid}(\delta_2 |Z|/2)}}, 0 \right) - 2G(1/2).$$

Then, using the same argument in the proof of Theorem 7, we have

$$|h(\delta_1) - h(\delta_2)| \lesssim \sqrt{\frac{p}{n}} + \epsilon,$$

with the choice $\kappa_1 = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$, $\kappa_2 \asymp 1$, and $\delta = e^{-C'(p+n\epsilon^2)}$, where $\delta_1 = \sqrt{v^T \Sigma v}$, $\delta_2 = \sqrt{v^T \widehat{\Sigma} v}$, and $h(t) = \int \mathsf{sigmoid}(t|z|/2)\phi(z)dz$. The notation $\phi(\cdot)$ is used for the density function of $N(0,1)$. Since

$$
\begin{aligned}
h'(t) &= \int \mathsf{sigmoid}(t|z|/2)(1 - \mathsf{sigmoid}(t|z|/2))|z|\phi(z)dz \\
&\geq \int \mathsf{sigmoid}(M^{1/2}|z|/2)(1 - \mathsf{sigmoid}(M^{1/2}|z|/2))|z|\phi(z)dz,
\end{aligned}
$$

which means $h'(t)$ is lower bounded by a constant uniformly over $|t| \leq M^{1/2}$, we have $|h(\delta_1) - h(\delta_2)| \geq c|\delta_1 - \delta_2|$. Following the last several lines of the proof of Proposition 4, we obtain the desired result. ∎

7.5 Proofs of Theorem 9 and Theorem 12

Proof [Proof of Theorem 9] We use the notation $T_{w,g}(x) = \mathsf{sigmoid}\left(\sum_{j\geq 1} w_j g_j(x)\right)$, and write

$$F_{w,g}(P,Q) = \mathbb{E}_{X\sim P} S(T_{w,g}(X), 1) + \mathbb{E}_{X\sim Q} S(T_{w,g}(X), 0).$$

By Lemma 17 and the same argument that leads to (38), we have

$$F_{w,g}(N(\theta,\Sigma), N(\widehat{\theta},\widehat{\Sigma})) - 2G(1/2) \leq 2C\kappa(2B)^{L-1}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1 B^{L-1}\kappa\epsilon, \quad (40)$$

uniformly over $\|w\|_1 \leq \kappa$ and $g_j \in \mathcal{G}^L(B)$. The second term in the above bound is $4C_1 B^{L-1}\kappa\epsilon$ instead of $4C_1\kappa\epsilon$ in (38) because $T \in \mathcal{T}^L(\kappa, B)$ implies that $\sup_x |T(x) - 1/2| \leq B^{L-1}\kappa$ according to the proof of Lemma 17.

We need to show that the function $\mathsf{sigmoid}(u^T x + b) \in \mathcal{G}^L(B)$ for any $u \in \mathbb{R}^p$ and any $b \in \mathbb{R}$. Note that $\mathsf{sigmoid}(u^T x + b) \in \mathcal{G}^1(B)$ is obvious. This is also true for $\mathcal{G}^2(B)$ by taking $\mathsf{sigmoid}(u^T x + b) \in \mathcal{G}_{\mathsf{sigmoid}}$, $v_1 = 1$ and $v_h = 0$ for all $h \geq 2$, because $\mathsf{ReLU}(\mathsf{sigmoid}(\cdot)) = \mathsf{sigmoid}(\cdot)$. Suppose $\mathsf{sigmoid}(u^T x + b) \in \mathcal{G}^l(B)$, we also have $\mathsf{sigmoid}(u^T x + b) \in \mathcal{G}^{l+1}(B)$, because $\mathsf{ReLU}(\mathsf{ReLU}(\cdot)) = \mathsf{ReLU}(\cdot)$, and the claim is proved by an induction argument.

Choose $w_1 = \kappa$, $w_j = 0$ for all $j \geq 2$, and $g_1(x) = \mathsf{sigmoid}(u^T x + b)$, and then we have

$$
\begin{aligned}
&\mathbb{E}_{X\sim N(\theta,\Sigma)} S\left(\frac{1}{1 + e^{-\kappa\mathsf{sigmoid}(u^T X + b)}}, 1\right) + \mathbb{E}_{X\sim N(\widehat{\theta},\widehat{\Sigma})} S\left(\frac{1}{1 + e^{-\kappa\mathsf{sigmoid}(u^T X + b)}}, 0\right) \\
&\leq 2G(1/2) + 2C\kappa(2B)^{L-1}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1 B^{L-1}\kappa\epsilon, \quad (41)
\end{aligned}
$$

uniformly over all $u \in \mathbb{R}^p$ and all $b \in \mathbb{R}$, with probability at least $1 - \delta$.

We further specify that $u = v$ for some unit vector v and $b = -v^T\theta$. Then, the inequality (41) becomes

$$f\left(\kappa; \sqrt{v^T\Sigma v}, \sqrt{v^T\widehat{\Sigma} v}, v^T(\widehat{\theta} - \theta)\right) \leq 2C\kappa(2B)^{L-1}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1 B^{L-1}\kappa\epsilon,$$

where

$$f(t; \delta_1, \delta_2, \Delta) = \mathbb{E} S\left(\frac{1}{1 + e^{-t\mathsf{sigmoid}(\delta_1 Z)}}, 1\right) + \mathbb{E} S\left(\frac{1}{1 + e^{-t\mathsf{sigmoid}(\delta_2 Z + \Delta)}}, 0\right) - 2G(1/2),$$

with $Z \sim N(0, 1)$. Then, using the same argument in the proof of Theorem 7, we have

$$\left| \int \mathsf{sigmoid}(\delta_1 z)\phi(z)dz - \int \mathsf{sigmoid}(\delta_2 z + \Delta)\phi(z)dz \right| \lesssim \sqrt{\frac{p}{n}} + \epsilon, \qquad (42)$$

with the choice $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$, $B \asymp 1$, $L \asymp 1$ and $\delta = e^{-C'(p+n\epsilon^2)}$, where $\delta_1 = \sqrt{v^T \Sigma v}$, $\delta_2 = \sqrt{v^T \widehat{\Sigma} v}$, $\Delta = v^T(\widehat{\theta} - \theta)$, and $\phi(\cdot)$ is the density function of $N(0, 1)$. Since $\int \mathsf{sigmoid}(\delta_1 z)\phi(z)dz = 1/2 = \int \mathsf{sigmoid}(\delta_2 z)\phi(z)dz$, this is equivalent to the bound $|h(0) - h(\Delta)| \lesssim \sqrt{\frac{p}{n}} + \epsilon$ with $h(t) = \int \mathsf{sigmoid}(\delta_2 z + t)\phi(z)dz$. It is easy to see that $|h'(0)| \geq \inf_{|\delta_2| \leq M^{1/2}} \int \mathsf{sigmoid}(\delta_2 z)(1 - \mathsf{sigmoid}(\delta_2 z))\phi(z)dz$, which is a constant, and the continuity of $h'(t)$ implies that there are small constants $c_1, c_2 > 0$, such that $\inf_{|t| \leq c_1} |h'(t)| \geq c_2$. Thus, as long as $|h(t) - h(0)|$ is sufficiently small, we have $|h(t) - h(0)| \geq c_2|t|$, which implies that $|v^T(\widehat{\theta} - \theta)| \lesssim \sqrt{p/n} + \epsilon$. Taking supreme over all unit vector, we have $\|\widehat{\theta} - \theta\| \lesssim \sqrt{p/n} + \epsilon$ with high probability.

To show the error bound for the covariance matrix estimator, we choose $u = v$ and $b = -v^T\theta - 1$ for some unit vector v in the inequality (41). Then, (42) becomes

$$\left| \int \mathsf{sigmoid}(\delta_1 z - 1)\phi(z)dz - \int \mathsf{sigmoid}(\delta_2 z + \Delta - 1)\phi(z)dz \right| \lesssim \sqrt{\frac{p}{n}} + \epsilon.$$

Since $|\mathsf{sigmoid}(\delta_2 z + \Delta - 1) - \mathsf{sigmoid}(\delta_2 z - 1)| \leq |\Delta| \lesssim \sqrt{\frac{p}{n}} + \epsilon$, we have

$$\left| \int \mathsf{sigmoid}(\delta_1 z - 1)\phi(z)dz - \int \mathsf{sigmoid}(\delta_2 z - 1)\phi(z)dz \right| \lesssim \sqrt{\frac{p}{n}} + \epsilon,$$

which can be written as $|\bar{h}(\delta_1) - \bar{h}(\delta_2)| \lesssim \sqrt{\frac{p}{n}} + \epsilon$, with $\bar{h}(t) = \int \mathsf{sigmoid}(tz - 1)\phi(z)dz$. Since $\inf_{0 < t \leq M^{1/2}} |\bar{h}'(t)| = \inf_{0 < t \leq M^{1/2}} \left| \int \mathsf{sigmoid}(tz - 1)(1 - \mathsf{sigmoid}(tz - 1))z\phi(z)dz \right|$ is lower bounded by some constant, we have $|\delta_1 - \delta_2| \lesssim \sqrt{\frac{p}{n}} + \epsilon$, and by following the last several lines of the proof of Proposition 4, we obtain the desired result. ∎

Proof [Proof of Theorem 12] We use the same notation $T_{w,g}(x)$ and $F_{w,g}(P, Q)$ defined in the proof of Theorem 9. The same argument that leads to (40) gives the following inequality,

$$F_{w,g}(E(\theta, \Sigma, H), E(\widehat{\theta}, \widehat{\Sigma}, \widehat{H})) - 2G(1/2)$$

$$\leq 2C\kappa(2B)^{L-1}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1 B^{L-1}\kappa\epsilon, \qquad (43)$$

uniformly over $\|w\|_1 \leq \kappa$ and $g_j \in \mathcal{G}^L(B)$.

Choose $w_1 = \kappa$, $w_j = 0$ for all $j \geq 2$, and $g_1(x) = \mathsf{ramp}\left(\frac{v^T x - v^T \widehat{\theta}}{\sqrt{v^T \Sigma v}}\right)$, for some unit vector v, and then we have

$$f\left(\kappa; \sqrt{v^T \Sigma v}, \sqrt{v^T \widehat{\Sigma} v}, v^T(\theta - \widehat{\theta})\right) \leq 2C\kappa(2B)^{L-1}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1 B^{L-1}\kappa\epsilon,$$

where

$$\begin{aligned}
f(t; \delta_1, \delta_2, \Delta) &= \int S\left(\frac{1}{1 + e^{-t\mathsf{ramp}(z + \Delta/\delta_1)}}, 1\right) dH(t) \\
&+ \int S\left(\frac{1}{1 + e^{-t\mathsf{ramp}((\delta_2/\delta_1)z)}}, 0\right) d\widehat{H}(t) - 2G(1/2).
\end{aligned}$$

Note that $\mathsf{sigmoid}(u^T x + b) \in \mathcal{G}^L(B)$ for any $u \in \mathbb{R}^p$ and any $b \in \mathbb{R}$ has already been proved in the proof of Theorem 9. The same argument also leads to the same conclusion for $\mathsf{ramp}(u^T x + b))$ for any $u \in \mathbb{R}^p$ and any $b \in \mathbb{R}$. Then, using the same argument in the proof of Theorem 7 ($\mathsf{ramp}(\cdot)$ is bounded between 0 and 1 just as $\mathsf{sigmoid}(\cdot)$), we have

$$\left|\int \mathsf{ramp}(z + \Delta/\delta_1) dH(z) - \int \mathsf{ramp}((\delta_2/\delta_1)z) d\widehat{H}(z)\right| \lesssim \sqrt{\frac{p}{n}} + \epsilon,$$

with the choice $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$, $B \asymp 1$, $L \asymp 1$ and $\delta = e^{-C'(p + n\epsilon^2)}$, where $\delta_1 = \sqrt{v^T \Sigma v}$, $\delta_2 = \sqrt{v^T \widehat{\Sigma} v}$, $\Delta = v^T(\theta - \widehat{\theta})$. Since $\int \mathsf{ramp}((\delta_2/\delta_1)z) d\widehat{H}(z) = 1/2 = \int \mathsf{ramp}(z) dH(z)$ by $H(t) + H(-t) \equiv 1$, the above inequality is equivalent to $|h(\Delta/\delta_1) - h(0)| \lesssim \sqrt{\frac{p}{n}} + \epsilon$, where $h(t) = \int \mathsf{ramp}(z + t) dH(z)$. Note that $h'(t) = \mathbb{P}_{Z \sim H}(|Z + t| \leq 1/2)$. By the condition $H \in \mathcal{H}(M')$, we have $\inf_{|t| \leq 1/6} |h'(t)| \geq 1/M'$. By the monotonicity of $h(t)$, as long as $|h(t) - h(0)|$ is sufficiently small, we have $|h(t) - h(0)| \geq (M')^{-1}|t|$, which implies

$$\frac{|v^T(\widehat{\theta} - \theta)|}{\sqrt{v^T \Sigma v}} \lesssim \sqrt{\frac{p}{n}} + \epsilon. \tag{44}$$

Since $\delta_1 \leq M^{1/2}$, we have $|v^T(\widehat{\theta} - \theta)| \lesssim \sqrt{\frac{p}{n}} + \epsilon$. Taking supreme over all unit vector, we have $\|\widehat{\theta} - \theta\| \lesssim \sqrt{p/n} + \epsilon$ with high probability.

To show the error bound for the scatter matrix estimator, we choose $w_1 = \kappa/2$, $w_2 = -\kappa/2$, $w_j = 0$ for all $j \geq 3$, $g_1(x) = \mathsf{ramp}\left(\frac{v^T(x - \theta)}{\sqrt{v^T \Sigma v}} - \frac{1}{2}\right)$, and $g_2(x) = \mathsf{ramp}\left(-\frac{v^T(x - \theta)}{\sqrt{v^T \Sigma v}} - \frac{1}{2}\right)$ for some unit vector v in (43). Since

$$\sum_{j \geq 1} w_j g_j(x) = \frac{\kappa}{2} R\left(\left|\frac{v^T(x - \theta)}{\sqrt{v^T \Sigma v}}\right|\right),$$

the inequality (43) becomes

$$\bar{f}\left(\kappa; \sqrt{v^T \Sigma v}, \sqrt{v^T \widehat{\Sigma} v}, v^T(\theta - \widehat{\theta})\right) \leq 2C\kappa(2B)^{L-1}\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) + 4C_1 B^{L-1}\kappa\epsilon,$$

37

where

$$\bar{f}(t; \delta_1, \delta_2, \Delta) = \int S\left(\frac{1}{1 + e^{-tR(|z|)/2}}, 1\right) dH(z)$$
$$+ \int S\left(\frac{1}{1 + e^{-tR(|(\delta_2/\delta_1)z - \Delta/\delta_1|)/2}}, 0\right) d\widehat{H}(z) - 2G(1/2).$$

Then, using the same argument in the proof of Theorem 7, we have

$$\left| \int R(|z|)dH(z) - \int R(|(\delta_2/\delta_1)z - \Delta/\delta_1|)d\widehat{H}(z) \right| \lesssim \sqrt{\frac{p}{n}} + \epsilon,$$

with the choice $\kappa = O\left(\sqrt{\frac{p}{n}} + \epsilon\right)$, $B \asymp 1$, $L \asymp 1$ and $\delta = e^{-C'(p + n\epsilon^2)}$, where $\delta_1 = \sqrt{v^T \Sigma v}$, $\delta_2 = \sqrt{v^T \widehat{\Sigma} v}$, $\Delta = v^T(\theta - \widehat{\theta})$. Since $|R(|(\delta_2/\delta_1)z - \Delta/\delta_1|) - R(|(\delta_2/\delta_1)z|)| \lesssim |\Delta/\delta_1|$, we have

$$\left| \int R(|(\delta_2/\delta_1)z|)d\widehat{H}(z) - \int R(|(\delta_2/\delta_1)z - \Delta/\delta_1|)d\widehat{H}(z) \right| \lesssim |\Delta/\delta_1| \lesssim \sqrt{\frac{p}{n}} + \epsilon$$

by (44). By triangle inequality, we get

$$\left| \int R(|z|)dH(z) - \int R(|(\delta_2/\delta_1)z|)d\widehat{H}(z) \right| \lesssim \sqrt{\frac{p}{n}} + \epsilon,$$

which can be written as $|\bar{h}(1) - \bar{h}(\delta_2/\delta_1)| \lesssim \sqrt{\frac{p}{n}} + \epsilon$, with $\bar{h}(t) = \int R(|tz|)d\widehat{H}(z)$, because $\int R(|z|)dH(z) = \int R(|z|)d\Phi(z) = \int R(|z|)d\widehat{H}(z)$ by the condition that $H, \widehat{H} \in \mathcal{H}$. Since

$$\bar{h}'(t) = 2\int_0^{1/t} z d\widehat{H}(z)$$
$$\geq 2\int_{1/(8t)}^{1/t} z d\widehat{H}(z)$$
$$\geq \frac{1}{4t}\mathbb{P}_{Z \sim \widehat{H}}\left(\frac{1}{8t} \leq Z \leq \frac{1}{t}\right),$$

we have $\inf_{|t-1| \leq 1/2} |\bar{h}'(t)| \geq (8M')^{-1}$. Since $\bar{h}(t)$ is increasing for all $t > 0$, as long as $|\bar{h}(t) - \bar{h}(1)|$ is sufficiently small, we have $|\bar{h}(t) - \bar{h}(1)| \geq (8M')^{-1}|t - 1|$, which implies $|\delta_2/\delta_1 - 1| \lesssim \sqrt{\frac{p}{n}} + \epsilon$. Following the last several lines of the proof of Theorem 7, we obtain the desired result. ∎

7.6 Proofs of Auxiliary Lemmas

Proof [Proof of Lemma 13] The bound for the class $\mathcal{T}_3$ was proved by Lemma 7.2 of Gao et al. (2019). The same bound also holds for the class $\mathcal{T}_1$ because $\mathcal{T}_1 \subset \mathcal{T}_3$. ∎

Proof [Proof of Lemma 14] Since $X_i \sim N(0, \Sigma)$, we can write $X_i = \Sigma^{1/2} Z_i$ with $Z_i \sim N(0, I_p)$. Define

$$f(Z_1, ..., Z_n) = \sup_{T \in \mathcal{T}_2} \left| \frac{1}{n} \sum_{i=1}^{n} \log T(\Sigma^{1/2} Z_i) - \mathbb{E} \log T(\Sigma^{1/2} Z) \right|.$$

We show $f(Z_1, ..., Z_n)$ is a Lipschitz function. We have

$$|f(Z_1, ..., Z_n) - f(Y_1, ..., Y_n)|$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\|w\|_1 \leq \kappa, \|u_j\| \leq 1} \left| \log \mathsf{sigmoid} \left(\sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T \Sigma^{1/2} Z_i) \right) \right.$$

$$\left. - \log \mathsf{sigmoid} \left(\sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T \Sigma^{1/2} Y_i) \right) \right|$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\|w\|_1 \leq \kappa, \|u_j\| \leq 1} \left| \sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T \Sigma^{1/2} Z_i) - \sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T \Sigma^{1/2} Y_i) \right| \quad (45)$$

$$\leq \frac{\kappa}{n} \sum_{i=1}^{n} \max_{j \geq 1} \sup_{\|u_j\| \leq 1} \left| \mathsf{ReLU}(u_j^T \Sigma^{1/2} Z_i) - \mathsf{ReLU}(u_j^T \Sigma^{1/2} Y_i) \right|$$

$$= \frac{\kappa}{n} \sum_{i=1}^{n} \sup_{\|u\| \leq 1} \left| \mathsf{ReLU}(u^T \Sigma^{1/2} Z_i) - \mathsf{ReLU}(u^T \Sigma^{1/2} Y_i) \right|$$

$$\leq \frac{\kappa}{n} \sum_{i=1}^{n} \sup_{\|u\| \leq 1} \left| u^T \left(\Sigma^{1/2} Z_i - \Sigma^{1/2} Y_i \right) \right| \quad (46)$$

$$\leq \frac{M^{1/2} \kappa}{n} \sum_{i=1}^{n} \|Z_i - Y_i\|$$

$$\leq \frac{M^{1/2} \kappa}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \|Z_i - Y_i\|^2}. \quad (47)$$

The inequalities (45) and (46) are implied by the fact that both the functions $\log \mathsf{sigmoid}(\cdot)$ and $\mathsf{ReLU}(\cdot)$ have Lipschitz constants bounded by 1. Therefore, $f(Z_1, ..., Z_n)$ is a Lipschitz function with Lipschitz constant $\frac{M^{1/2} \kappa}{\sqrt{n}} = O\left(\frac{\kappa}{\sqrt{n}}\right)$. By Talagrand's inequality (Talagrand, 1995), we have

$$f(Z_1, ..., Z_n) \leq \mathbb{E} f(Z_1, ..., Z_n) + C\kappa \sqrt{\frac{\log(2/\delta)}{n}},$$

with probability at least $1 - \delta$.

To bound $\mathbb{E} f(Z_1, ..., Z_n)$, we use a standard symmetrization argument (Pollard, 2012) and obtain the following bound that involves Rademacher complexity,

$$\mathbb{E} f(Z_1, ..., Z_n) \leq 2\mathbb{E} \sup_{T \in \mathcal{T}_2} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log T(\Sigma^{1/2} Z_i) \right|, \quad (48)$$

where $\epsilon_1, ..., \epsilon_n$ are i.i.d. uniform random variables on $\{-1, 1\}$. To bound the Rademacher complexity, we have

$$
\mathbb{E} \sup_{T \in \mathcal{T}_2} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log T(\Sigma^{1/2} Z_i) \right|
$$

$$
\leq \quad 2\mathbb{E} \sup_{\|w\|_1 \leq \kappa, \|u_j\| \leq 1} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \log \mathsf{sigmoid} \left(\sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T \Sigma^{1/2} Z_i) \right) \right)
$$

$$
\leq \quad 2\mathbb{E} \sup_{\|w\|_1 \leq \kappa, \|u_j\| \leq 1} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(\sum_{j \geq 1} w_j \mathsf{ReLU}(u_j^T \Sigma^{1/2} Z_i) \right) \right) \tag{49}
$$

$$
\leq \quad 2\kappa \mathbb{E} \sup_{\|u\| \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathsf{ReLU}(u^T \Sigma^{1/2} Z_i) \right|
$$

$$
\leq \quad 4\kappa \mathbb{E} \sup_{\|u\| \leq 1} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathsf{ReLU}(u^T \Sigma^{1/2} Z_i) \right)
$$

$$
\leq \quad 4\kappa \mathbb{E} \sup_{\|u\| \leq 1} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i u^T \Sigma^{1/2} Z_i \right) \tag{50}
$$

$$
\leq \quad 4\kappa M^{1/2} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i Z_i \right\|
$$

$$
\leq \quad 4\kappa M^{1/2} \sqrt{\frac{p}{n}}. \tag{51}
$$

The inequalities (49) and (50) are by Theorem 7 of Meir and Zhang (2003). The last inequality (51) is because $\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i Z_i \right\| \leq \sqrt{\mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i Z_i \right\|^2} = \sqrt{\frac{p}{n}}$. The proof is complete by combining the bounds above. $\blacksquare$

Proof [Proof of Lemma 15] Let $f(X_1, ..., X_n) = \sup_{T \in \mathcal{T}_3} \left| \frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) - \mathbb{E} S(T(X), 1) \right|$. Since

$$
\sup_{T \in \mathcal{T}_3} \sup_{x} \left| S(T(x), 1) - G\left(\frac{1}{2} \right) - \frac{1}{2} G'\left(\frac{1}{2} \right) \right|
$$

$$
\leq \quad \sup_{|t - \frac{1}{2}| \leq \kappa} \left| S(t, 1) - G\left(\frac{1}{2} \right) - \frac{1}{2} G'\left(\frac{1}{2} \right) \right|
$$

$$
\leq \quad \kappa \sup_{|t - \frac{1}{2}| \leq \kappa} \left| (1 - t) G''(t) \right| \leq C_1 \kappa,
$$

where we have used $\sup_{|t - \frac{1}{2}| \leq \kappa} |(1 - t) G''(t)| \leq C_1$ because of the smoothness of $G(t)$ at $t = 1/2$ by Condition 6. In the second last inequality above, we have used the fact that $\frac{\partial}{\partial t} S(t, 1) = (1 - t) G''(t)$ and $S(t, 1) = G(t) + (1 - t) G'(t)$. This implies that

$$
\sup_{x_1, ..., x_n, x_i'} \left| f(x_1, ..., x_n) - f(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n) \right| \leq \frac{2C_1 \kappa}{n}.
$$

Therefore, by McDiarmid's inequality (McDiarmid, 1989), we have

$$f(X_1, ..., X_n) \leq \mathbb{E}f(X_1, ..., X_n) + C_1 \kappa \sqrt{\frac{2\log(1/\delta)}{n}}, \tag{52}$$

with probability at least $1 - \delta$. By the same argument of (48), it is sufficient to bound the Rademacher complexity $\mathbb{E}\sup_{T \in \mathcal{T}_3} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i S(T(X_i), 1) \right|$. Since $T \in \mathcal{T}_3$ implies $|T(X) - 1/2| \leq \kappa$, the function $S(\mathsf{sigmoid}(\cdot), 1)$ has a Lipschitz constant bounded by C_1 on the domain of interest. By Theorem 7 of Meir and Zhang (2003), we have

$$\begin{aligned}
&\mathbb{E} \sup_{T \in \mathcal{T}_3} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i S(T(X_i), 1) \right| \\
&\leq \quad 2C_1 \mathbb{E} \sup_{\|w\|_1 \leq \kappa, u_j \in \mathbb{R}^p, b_j \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \sum_{j \geq 1} w_j \mathsf{sigmoid}(u_j^T X_i + b_j) \right|.
\end{aligned} \tag{53}$$

By Hölder's inequality, we can further bound the above term by

$$2C_1 \kappa \mathbb{E} \sup_{u \in \mathbb{R}^p, b \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathsf{sigmoid}(u^T X_i + b) \right|.$$

Define $\mathcal{D} = \left\{ D(x) = \mathsf{sigmoid}(u^T x + b) : u \in \mathbb{R}^p, b \in \mathbb{R} \right\}$, and then the Rademacher complexity can be bounded by Dudley's integral entropy, which gives

$$\mathbb{E} \sup_{D \in \mathcal{D}} \left| \frac{1}{n} \epsilon_i D(X_i) \right| \leq \mathbb{E} \frac{1}{\sqrt{n}} \int_0^2 \sqrt{\log \mathcal{N}(\delta, \mathcal{D}, \|\cdot\|_n)} d\delta, \tag{54}$$

where $\mathcal{N}(\delta, \mathcal{D}, \|\cdot\|_n)$ is the δ-covering number of $\mathcal{D}$ with respect to the empirical ℓ_2 distance $\|f - g\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2}$. Since the VC-dimension of $\mathcal{D}$ is $O(p)$, we have $\mathcal{N}(\delta, \mathcal{D}, \|\cdot\|_n) \lesssim p(16e/\delta)^{O(p)}$ (see Theorem 2.6.7 of van der Vaart and Wellner (1996)). This leads to the bound $\frac{1}{\sqrt{n}} \int_0^2 \sqrt{\log \mathcal{N}(\delta, \mathcal{D}, \|\cdot\|_n)} d\delta \lesssim \sqrt{\frac{p}{n}}$, which gives the desired result. $\blacksquare$

Proof [Proof of Lemma 16] Following the proof of Lemma 14, we write $X_i = \Sigma^{1/2} Z_i$ with $Z_i \sim N(0, I_p)$. Define

$$f(Z_1, ..., Z_n) = \sup_{T \in \mathcal{T}_4} \left| \frac{1}{n} \sum_{i=1}^n S(T(\Sigma^{1/2} Z_i), 1) - \mathbb{E}S(T(\Sigma^{1/2} Z), 1) \right|.$$

Since $T \in \mathcal{T}_4$ implies $|T(X) - 1/2| \leq \kappa_1$, the function $S(\text{sigmoid}(\cdot), 1)$ has a Lipschitz constant bounded by some constant C_1 on the domain of interest. Therefore,

$$
\begin{aligned}
& |f(Z_1, ..., Z_n) - f(Y_1, ..., Y_n)| \\
\leq \quad & C_1 \frac{1}{n} \sum_{i=1}^{n} \sup_{\|w\|_1 \leq \kappa_1, \|v_j\|_1 \leq \kappa_2, \|u_l\| \leq 1} \left| \sum_{j \geq 1} w_j \text{sigmoid} \left(\sum_{l \geq 1} v_{jl} \text{ReLU}(u_l^T \Sigma^{1/2} Z_i) \right) \right. \\
& \left. - \sum_{j \geq 1} w_j \text{sigmoid} \left(\sum_{l \geq 1} v_{jl} \text{ReLU}(u_l^T \Sigma^{1/2} Y_i) \right) \right| .
\end{aligned}
$$

Then, by a successive argument of Hölder's inequalities and Lipschitz continuity similar to (45)-(47), we have

$$
|f(Z_1, ..., Z_n) - f(Y_1, ..., Y_n)| \leq C_2 \frac{\kappa_1 \kappa_2}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \|Z_i - Y_i\|^2}.
$$

By Talagrand's inequality (Talagrand, 1995), we have

$$
f(Z_1, ..., Z_n) \leq \mathbb{E} f(Z_1, ..., Z_n) + C \kappa_1 \kappa_2 \sqrt{\frac{\log(2/\delta)}{n}},
$$

with probability at least $1 - \delta$.

To bound $\mathbb{E} f(Z_1, ..., Z_n)$, it is sufficient to analyze the Rademacher complexity according to (48). Again, this can be done by following the same argument that leads to (51), and thus we have $\mathbb{E} f(Z_1, ..., Z_n) \lesssim \kappa_1 \kappa_2 \sqrt{\frac{p}{n}}$, which completes the proof. ∎

Proof [Proof of Lemma 17] We first prove the result for $\mathcal{T} = \mathcal{T}^L(\kappa, B)$. Let $f(X_1, ..., X_n) = \sup_{T \in \mathcal{T}^L(\kappa, B)} \left| \frac{1}{n} \sum_{i=1}^{n} S(T(X_i), 1) - \mathbb{E} S(T(X), 1) \right|$. For any $g \in \mathcal{G}_{\text{sigmoid}}$, we have $\sup_x |g(x)| \leq 1$. Suppose for any $g \in \mathcal{G}^l(B)$, $\sup_x |g(x)| \leq \tau$, then we have $\sup_x |g(x)| \leq B\tau$ for any $g \in \mathcal{G}^{l+1}(B)$ by Hölder's inequality. A mathematical induction argument then gives $\sup_x |g(x)| \leq B^{L-1}$ for any $g \in \mathcal{G}^L(B)$. Therefore, $T \in \mathcal{T}^L(\kappa, B)$ implies that $\sup_x |T(x) - 1/2| \leq B^{L-1}\kappa$. By the same argument that derives (52), we then have

$$
f(X_1, ..., X_n) \leq \mathbb{E} f(X_1, ..., X_n) + C_1 B^{L-1} \kappa \sqrt{\frac{2 \log(1/\delta)}{n}},
$$

with probability at least $1 - \delta$.

It is sufficient to analyze the Rademacher complexity, and we have

$$\mathbb{E} \sup_{T \in \mathcal{T}^L(\kappa, B)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i S(T(X_i), 1) \right|$$

$$\leq 2C_1 \mathbb{E} \sup_{\|w\|_1 \leq \kappa, g_j \in \mathcal{G}^L(B)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \sum_{j \geq 1} w_j g_j(X_i) \right| \tag{55}$$

$$\leq 2C_1 \kappa \mathbb{E} \sup_{g \in \mathcal{G}^L(B)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| \tag{56}$$

$$\leq 4C_1 \kappa \mathbb{E} \sup_{\|v\|_1 \leq B, g_h \in \mathcal{G}^{L-1}(B)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \sum_{h=1}^{H} v_h g_h(X_i) \right| \tag{57}$$

$$\leq 4C_1 B \kappa \mathbb{E} \sup_{g \in \mathcal{G}^{L-1}(B)} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| \tag{58}$$

$$\leq 2C_1 \kappa (2B)^{L-1} \mathbb{E} \sup_{g \in \mathcal{G}_{\text{sigmoid}}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right|. \tag{59}$$

We explain each of the inequalities above. The first inequality (55) follows the same argument that derives (53). The inequalities (56) and (58) are implied by Hölder's inequality. We have used Theorem 7 of Meir and Zhang (2003) to derive (57). Finally, (59) is from a mathematical induction argument. Note that $\mathbb{E} \sup_{g \in \mathcal{G}_{\text{sigmoid}}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| \lesssim \sqrt{\frac{p}{n}}$ can be derived from Dudley's integral entropy (54), and then we obtain the desired result.

The class $\bar{\mathcal{T}}^L(\kappa, B)$ only differs from $\mathcal{T}^L(\kappa, B)$ in the bottom layer. That is, we use sigmoid$(\cdot)$ in the bottom layer of $\mathcal{T}^L(\kappa, B)$ and ramp$(\cdot)$ in the bottom layer of $\bar{\mathcal{T}}^L(\kappa, B)$. When we prove the result for $\mathcal{T}^L(\kappa, B)$, we use the following properties of sigmoid$(\cdot)$: it is a function bounded between 0 and 1; it is increasing; it has a bounded Lipschitz constant. All of the properties hold for ramp$(\cdot)$, and thus the desired conclusion holds for $\bar{\mathcal{T}}^L(\kappa, B)$ as well. ∎

Acknowledgments

References

Nina Amenta, Marshall Bern, David Eppstein, and S-H Teng. Regression depth and center points. *Discrete & Computational Geometry*, 23(3):305–323, 2000.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

Yannick Baraud and Lucien Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3, 2005.

Timothy M Chan. An optimal randomized algorithm for maximum tukey depth. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 430–436. Society for Industrial and Applied Mathematics, 2004.

Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Laurie Davies. The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, 20(4):1828–1843, 1992.

A Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 73–84. IEEE, 2017.

Peter J Diggle and Richard J Gratton. Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

Simon S Du, Sivaraman Balakrishnan, and Aarti Singh. Computationally efficient robust estimation of sparse functionals. *arXiv preprint arXiv:1702.07709*, 2017.

Lutz Dümbgen. On Tyler's M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics*, 50(3):471–491, 1998.

Lutz Dümbgen and David E Tyler. On the breakdown properties of some multivariate M-functionals. *Scandinavian Journal of Statistics*, 32(2):247–264, 2005.

Kai Wang Fang. *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC, 2017.

Soheil Feizi, Changho Suh, Fei Xia, and David Tse. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

Yoav Freund and Robert E. Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory (COLT)*, 1996.

Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv preprint arXiv:1810.02030.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

Fang Han and Han Liu. Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint arXiv:1305.6916*, 2013.

Fang Han and Han Liu. Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *Journal of the American Statistical Association*, 109(505):275–287, 2014.

Peter J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1):73–101, 1964.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Jiantao Jiao, Thomas A Courtade, Kartik Venkat, and Tsachy Weissman. Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, 61(10):5357–5365, 2015.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

William H Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

Annick M Leroy and Peter J Rousseeuw. *Robust Regression and Outlier Detection*. John Wiley & Sons, 1987.

Tengyuan Liang. How well can generative adversarial networks (gan) learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017.

Ricardo Antonio Maronna. Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67, 1976.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141 (1):148–188, 1989.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

Ritwik Mitra and Cun-Hui Zhang. Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint arXiv:1403.6195*, 2014.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

Davy Paindaveine and Germain Van Bever. Halfspace depths for scatter, concentration and shape matrices. *The Annals of Statistics*, 46(6B):3276–3307, 2018.

David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Peter J Rousseeuw and Mia Hubert. Regression depth. *Journal of the American Statistical Association*, 94(446):388–402, 1999.

Peter J Rousseeuw and Anja Struyf. Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8(3):193–203, 1998.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1): 73–205, 1995.

Zhiqiang Tan, Yunfu Song, and Zhijian Ou. Calibrated adversarial algorithms for generative modelling. *Stat.*, 8:e224, 2019.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

David E Tyler. A distribution-free m-estimator of multivariate scatter. *The Annals of Statistics*, 15(1):234–251, 1987.

Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes.* Springer, 1996.

Marten Wegkamp and Yue Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226, 2016.

Jian Zhang. Some extensions of tukey's depth function. *Journal of Multivariate Analysis*, 82(1):134–165, 2002.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

Teng Zhang, Xiuyuan Cheng, and Amit Singer. Marčenko–Pastur law for Tyler's M-estimator. *Journal of Multivariate Analysis*, 149:114–123, 2016.

Banghua Zhu, Jiantao Jiao, and David Tse. Deconstructing generative adversarial networks. *arXiv preprint arXiv:1901.09465*, 2019.

Yijun Zuo and Robert Serfling. Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions. *Journal of Multivariate Analysis*, 75(1):62–78, 2000.