A Survey of GAN and Robust Statistics

Sihan Liu

September 22, 2021

1 GAN Introduction

In Generative Adversarial Learning, there are three key components:

- P_θ, the Target distribution implicitly parametrized by an unknown parameter θ
- $P_{\hat{\theta}}$, the Generator distribution parametrized by $\hat{\theta}$.
- $D : \mathbb{R}^d \mapsto [0,1]$, the Discriminator which tries to discern samples from P_θ and $P_{\hat{\theta}}$. It is chosen among a family of functions $\mathcal{T}$ often represented by a neural network architecture.

In the standard GAN framework, a common objective function [1]

$$\min_{\hat{\theta}} \max_{D \in \mathcal{T}} \mathcal{L}\left(P_\theta, P_{\hat{\theta}}; D\right) = \mathop{\mathbf{E}}_{\mathbf{x} \sim P_\theta} [\log D(\mathbf{x})] + \mathop{\mathbf{E}}_{\mathbf{x} \sim P_{\hat{\theta}}} [\log(1 - D(x))]$$

is shared between the Generator and the Discriminator. The two entities are then trained against each other to find equilibrium of the minimax game. In practice, only sample access of the target distribution is given. The game is hence defined with the empirical distribution $\mathbb{P}_\theta^n$ that consists of n samples from the real data distribution P_θ. If we ignore the optimization process, the final output given by the algorithm will be exactly

$$\widetilde{\theta} = \arg\min_{\hat{\theta}} \max_{D \in \mathcal{T}} \mathcal{L}\left(\mathbb{P}_\theta^n, P_{\hat{\theta}}; D\right).$$

Ideally, the estimator $\widetilde{\theta}$ will be close to θ and P_θ, $P_{\widetilde{\theta}}$ will be close in some statistical distance when the number of samples n taken are sufficiently large.

2 Robust Statistics

In the non-robust setting, the samples $\mathbf{x}^{(1)} \cdots \mathbf{x}^{(n)}$ will be i.i.d samples coming from the distribution P_θ exactly. In many cases, however, the samples may be corrupted. Here, we will focus on the Huber Contamination model and the Strong Contamination model.

Under the Huber Contamination Model, the samples will instead come from $(1-\epsilon)P_\theta + \epsilon Q$, where Q is an adversarially chosen distribution; under the Strong Contamination Model, the samples may come from any distribution P satisfying that $\mathrm{TV}\left(P, P_\theta\right) \leq \epsilon$. It is easy to see Huber Contamination model is weaker as the mixed distribution $\epsilon P_\theta + (1-\epsilon)Q$ is indeed ϵ closed to the original distribution. Most of the results in [WDHY20] and [GLYZ18] are shown under Huber's Contamination Model. Yet, as demonstrated in [GYZ20], the techniques can easily be extended to the Strong Contamination Model.

[1] Other objective functions may be used, which could lead to different training effects in practice. See [NCT16]

3 Sample Complexity Framework

We will present the framework used frequently in the literature. The results crucially depend on the architecture of the Discriminator and, more specifically, properties of the activation function.

At a high level, we first show that, in the end of training, the objective function under the final estimator is upper bounded i.e. $\sup_{D\in\mathcal{T}} \mathcal{L}\left(P_\theta, P_{\widetilde{\theta}}; D\right) \leq U$ where U goes to $O(\epsilon)$ with high probability when a sufficiently large number of samples are used. Then, a lower bound is established to show that, under some Discriminator $D_{\theta,\widetilde{\theta}} \in \mathcal{T}$ that depends on the two distributions, the loss is descriptive enough to characterize the "distance" between θ and $\widetilde{\theta}$ (in case of parameter estimation) or the "statistical distance" between $P_\theta, P_{\widetilde{\theta}}$ (in case of distribution learning) i.e. $\mathcal{L}\left(P_\theta, P_{\widetilde{\theta}}; D_{\theta,\widetilde{\theta}}\right) \gtrsim \left\|\theta - \widetilde{\theta}\right\|$.

Though there are still various details to work out, it is not hard to see that once the two bounds are established, the convergence rate of the estimator $\widetilde{\theta}$ follows. We will proceed to discuss the techniques used in obtaining the two bounds.

3.1 Population Loss of Sample-Optimal Estimator

In the non-robust setting, the upper bound simply characterizes how well the loss generalizes from sample to the population. Intuitively, the more complicated the family of Discriminator is, the weaker the loss function's generalization ability is. Hence, the Discriminator is chosen from a specific neural net architecture. A common one is the family

$$\mathcal{T} = \{D(\mathbf{x})\sigma(\sum_{j\geq 1} w_j \sigma(\mathbf{u}_j^T\mathbf{x} + b_j))\text{s.t} \sum_{j\geq 11} |w_j| \leq \kappa, \mathbf{u}_j \in \mathbb{R}^d, b_j \in \mathbb{R}\}\,, \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

3.1.1 Generalization Gap

Its generalization ability is given below

Lemma 1 (Lemma 8.2 of [GLYZ18]). *Given i.i.d observations $\mathbf{x}^{(1)}, \mathbf{x}^{(n)} \sim P \in \delta(\mathbb{R}^d)$, and the function class defined in , it holds*

$$\sup_{D\in\mathcal{T}} \left| \frac{1}{n}\sum_{i=1}^{n} \log D(\mathbf{x}^{(i)}) - \mathop{\mathbf{E}}_{\mathbf{x}\sim P} \log D(\mathbf{x}) \right| \leq C\kappa \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right),$$

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

The proof uses standard concentration inequalities and techniques involving Rademacher complexity of neural nets. At a high level, we first show the quantity is highly concentrated as $\log D(\mathbf{x})$ satisfies the bounded difference condition of McDiarmid's inequality (which contributes to the $\kappa\sqrt{\frac{\log(1/\delta)}{n}}$ term). Then, using a symmetrization argument, we claim the expected value of the difference is bounded by the Rademacher complexity of $\mathcal{T}$ (which contributes to the $\kappa\sqrt{\frac{d}{n}}$ term).

3.1.2 Robustness

In the robust setting, we also need to bound how much the loss function could deviate due to the corrupting samples. Quite surprisingly, that's the only accommodation we need to make for robustness.

Claim 1. *Suppose $TV(P, P_\theta) \le \epsilon$. We have*

$$\sup_{D \in \mathcal{T}, \hat{\theta}} \left| \mathcal{L}(P, P_{\hat{\theta}}; D) - \mathcal{L}(P_\theta, P_{\hat{\theta}}; D) \right| \le 2\kappa\epsilon \,,$$

where $\mathcal{T}$ is defined as in Equation (1).

The key observation is that $|\log D(\mathbf{x})| \approx \left|\sum_j w_j \sigma(\mathbf{u}_j^T \mathbf{x} + b_j)\right|$. Hence, the claim is true as long as $|\sigma(\cdot)|$ is bounded. This implies that any bounded activation functions lead to robust Discriminator, which includes but not limits to $f(x) = 1/(1+\exp(-x))$ (sigmoid), $f(x) = \mathbf{1}\{x \ge 10\}$ (step-wise function) and $f(x) = \max(\min(x + 1/2, 1), 0$ (ramp function).

On the other hand, if the network does not have a hidden layer (equivalent as using the identity function $I(x) = x$ as the activation function) or the activation function's range is unbounded ($\mathrm{ReLU}(x) = x$ if $x > 0$ else 0), the resulting estimator is most likely non-robust. Conceptually, the reason is due to GAN's "moment-matching" effect.

Lemma 2 (Proposition 3.1 in [GLYZ18])**.** *Given an arbitrary feature mapping $g : \mathbb{R}^d \mapsto \mathbb{R}^h$, define the divergence measure*

$$F_g(P, Q) = \sup_{\mathbf{w} \in \mathcal{W}} \left\{ \mathop{\mathbf{E}}_{\mathbf{x} \sim P} \left[\log \sigma \left(\mathbf{w}^T g(\mathbf{x}) \right) \right] + \mathop{\mathbf{E}}_{\mathbf{x} \sim Q} \left[\log \left(1 - \sigma \left(\mathbf{w}^T g(\mathbf{x}) \right) \right) \right] \right\},$$

where $\mathcal{W}$ is a convex set. Then, $F_g(P, Q) = 0$ if and only if $\mathbf{E}_{\mathbf{x} \sim P}\, g(\mathbf{x}) = \mathbf{E}_{\mathbf{x} \sim Q}\, g(\mathbf{x})$.

Naive moment matching is perfectly fine under the non-robust setting. However, if "outliers" are added to the training samples and the range of g is unbounded, the true first moment of the feature vector, $\mathbf{E}_{\mathbf{x} \sim P_\theta}\, g(\mathbf{x})$, can be arbitrarily far away from the corrupted one $(1-\epsilon)\, \mathbf{E}_{\mathbf{x} \sim P_\theta}\, g(\mathbf{x}) + \epsilon\, \mathbf{E}_{\mathbf{x} \sim Q}\, g(\mathbf{x})$. If the Generator still blindly follows the first moment, it will end up putting way too much mass around the outliers. Fortunately, this can be easily addressed by using some bounded activation function in the second last layer.

3.1.3 Loss at Equilibrium

We will provide a succinct proof sketch for the upper bound with the tools developed so far. The quantity of interest is defined with the real data distribution P_θ. The algorithm optimizes on the empirical distribution $\mathbb{P}^n$ defined on n samples from the corrupted distribution P satisfying $\mathrm{TV}(P, P_\theta) < \epsilon$. Our goal is to relate the two loss, one defined with $\mathbb{P}^n$ and the other defined with P_θ.

We will present the argument for the Discriminator class $\mathcal{T}$ defined in Equation (1).

$$\begin{aligned}
\sup_{D\in\mathcal{T}} \mathcal{L}(P_\theta, P_{\widetilde{\theta}}; D) &\le \sup_{D\in\mathcal{T}} \mathcal{L}(P, P_{\widetilde{\theta}}; D) + 2\kappa\epsilon && \text{(Robustness of } \mathcal{T}\text{)}\\
&\le \sup_{D\in\mathcal{T}} \mathcal{L}(P^n, P_{\widetilde{\theta}}; D) + 2\kappa\epsilon + C\kappa\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) && \text{(Generalization Gap of } \mathcal{T}\text{)}\\
&\le \sup_{D\in\mathcal{T}} \mathcal{L}(P^n, P_{\theta}; D) + 2\kappa\epsilon + C\kappa\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) && \text{(Optimality of } \widetilde{\theta} \text{ with } P^n\text{)}\\
&\le \sup_{D\in\mathcal{T}} \mathcal{L}(P, P_{\theta}; D) + 2\kappa\epsilon + 2C\kappa\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) && \text{(Generalization Gap of } \mathcal{T}\text{)}\\
&\le \sup_{D\in\mathcal{T}} \mathcal{L}(P_\theta, P_{\theta}; D) + 4\kappa\epsilon + 2C\kappa\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) && \text{(Robustness Gap of } \mathcal{T}\text{)}\\
&\le 4\kappa\epsilon + 2C\kappa\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right) && (\sup_{D\in\mathcal{T}} \mathcal{L}(P_\theta, P_\theta; D) = 0)
\end{aligned}$$

Notice that this is a general argument that depends only on the choice of the Discriminator family regardless of the family of distribution learned. As long as the family has reasonable Rademacher complexity and $\log D(\mathbf{x})$ is bounded, the resulting loss will enjoy good generalization property and robustness. We will discuss what constitutes a "descriptive" enough loss function for the underlying learning task in the next section.

3.2 Lower Bound of the Loss

The innovative part of the work in [GLYZ18], [GYZ20] and [WDHY20] is to give variational interpretation of the loss function. In particular, [GYZ20] explicitly connects it to Tukey Median and Matrix Depth Functions, which are themselves important tools in traditional robust mean/covariance estimation algorithms.

3.2.1 Mean Estimation

To illustrate the idea, we will analyze the following Discriminator architecture

$$\mathcal{T}_1 = \{D(\mathbf{x})\sigma(w\sigma(\mathbf{u}^T\mathbf{x})) \text{ s.t } |w| \le \kappa, \mathbf{u}_j \in \mathbb{R}^d, b_j \in \mathbb{R}\}, \tag{2}$$

where $\sigma(x) = 1/(1 + exp(-x))$. While the architecture specified in Equation (1) are also good, it turns out this simpler architecture suffices (as one can see, it is a strict subset).

Lemma 3. *Assume that $\|\theta\|_2 < M$ and C is a constant that depends only on M. Then, for $\mathcal{T}_1$ specified in Equation (2), if $\sup_{D\in\mathcal{T}_1} \mathcal{L}(\mathcal{N}(\theta, \mathbf{I}), \mathcal{N}(\hat{\theta}, \mathbf{I}); D) \le C$, it holds*

$$\sup_{D\in\mathcal{T}_1} \mathcal{L}(\mathcal{N}(\theta, \mathbf{I}), \mathcal{N}(\hat{\theta}, \mathbf{I}); D) \gtrsim \kappa \left\|\theta - \hat{\theta}\right\|_2,$$

for any $\kappa \in (0, 1]$.

Proof. As we can see, after the first layer, the two multivariate Gaussians are immediately projected on the direction $\mathbf{u}$. If we force $\mathbf{u}$ to be a unit vector, we then have

$$\begin{aligned}\mathcal{L}(\theta,\hat{\theta};D) &= \mathop{\mathbf{E}}_{z\sim(\mathcal{N}(\theta,\mathbf{I})}\left[\log\left(\frac{2}{1+\exp(-w\cdot\sigma(\mathbf{u}^T\mathbf{x}))}\right)\right] + \mathop{\mathbf{E}}_{z\sim(\mathcal{N}(\hat{\theta},\mathbf{I})}\left[\log\left(\frac{2}{1+\exp(w\cdot\sigma(\mathbf{u}^T\mathbf{x}))}\right)\right] \\ &= \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\left[\log\left(\frac{2}{1+\exp(-w\cdot\sigma(z+\mathbf{u}^T\theta))}\right)\right] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\left[\log\left(\frac{2}{1+\exp(w\cdot\sigma(z+\mathbf{u}^T\hat{\theta}))}\right)\right]\end{aligned}$$

We can see the two expressions in the expectations are all of form $f(a) = \log\frac{1}{1+\exp(a)}$. Intuitively, they are all "about linear" with respect to a up to some error terms. Based on that, we take Taylor Expansion for the two expressions. Notice that $f(0) = \log(1) = 0$. $f'(0) = \frac{1}{2}$. The rest of the lower order terms vanish in an exponential rate as long as $|a| \leq 1$, which is true whenever $|w| \leq 1$ [2]. Hence, only the first order terms are left. By linearity of expectation, we can conclude that

$$\mathcal{L}(\theta,\hat{\theta};D) \gtrsim \left|\mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\left[w\sigma(z+\mathbf{u}^T\theta)\right] - \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\left[w\sigma(z+\mathbf{u}^T\hat{\theta})\right]\right|.$$

Denote $h(t) = \mathbf{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z+\mathbf{u}^T\theta+t)\right]$. We will take a closer look at the landscape of the function. It is not hard to see that, due to the concentration of Gaussian variables, the function's derivative is lower bounded in a constant neighborhood. Namely, $\inf_{|t|<c_2} h'(t) \geq c_1$ for some constant c_1, c_2 that depends only on the magnitude of $\mathbf{u}^T\theta$, which is at most a constant M by assumption. [3]. On the other hand, $h(t)$ is obviously a monotonically increasing function. Hence, as long as $|h(t) - h(0)| < c_1 \cdot c_2$ $h(t)$ has lower bounded derivative.

Overall, this gives that $\mathcal{L}(\theta,\hat{\theta};D) \gtrsim w\left(\mathbf{u}^T\theta - \mathbf{u}^T\hat{\theta}\right)$ as long as $\mathcal{L}(\theta,\hat{\theta};D)$ is bounded by some constant C. Finally, taking $w = \kappa$ and $\mathbf{u} = (\theta - \hat{\theta})$ gives the result. □

As we can see in the proof, at a conceptual level, the Discriminator is simply probing the projected distance of $\mathbf{u}^T(\theta - \hat{\theta})$ on many different directions. If we can make sure that the projected distances are small in all possible directions, it is guaranteed that $\left\|\theta - \hat{\theta}\right\|_2$ is small.

3.2.2 Covariance Estimation

Such variation argument can also be used to prove optimality in covariance estimation. For $\mathbf{x} \sim \mathcal{N}(\theta, \mathbf{I})$, projecting it along a unit vector gives $\mathbf{u}^T\mathbf{x} \sim \mathcal{N}(\mathbf{u}^T\theta, 1)$; for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$, we similarly have $\mathbf{u}^T\mathbf{x} \sim \mathcal{N}(0, \mathbf{u}^T\Sigma\mathbf{u})$. This allows us to write the loss for covariance estimation as

$$\mathcal{L}(\theta,\hat{\theta};D) = \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\left[\log\left(\frac{2}{1+\exp(-w\sigma(\delta_1 z)}\right)\right] + \mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\left[\log\left(\frac{2}{1+\exp(w\sigma(\delta_2 z))}\right)\right].$$

Unfortunately, for covariance estimation, we have extra requirements on the activation function. In particular, symmetric activation functions like sigmoid won't work. This is again due to the "moment matching" effect as shown in Lemma 2. To be more specific, for any function $f : \mathbb{R} \mapsto \mathbb{R}$ satisfying

[2] In [GLYZ18], w is set to be $\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{d}{n}}$. The author doubts that this is not needed as the experiments in practice does not really need the constrain. We verify it here.

[3] However, the constant may be exponentially small with respect to M as the mass of the "unfiltered" region shrinks exponentially with $\mathbf{u}^t\theta$. In the original work of [GLYZ18], the argument is skipped and the dependency on $\|\theta\|_2$ is not made explicitly. This is fixed in the work of [GYZ20].

$f(x)+f(-x)=C$ for some constant C, we always have $\mathbf{E}_{z\sim\mathcal{N}(0,1)}\left[w\cdot f(\delta z)\right]=\int_0^\infty \phi(z)\cdot w\cdot(f(\delta z)+f(-\delta z))dz = w\cdot\frac{C}{2}$. Hence, the loss function will be a constant for whatever Discriminator chosen.

The issue can be easily addressed by making the activation function "asymmetric". ReLU is a natural choice that indeed leads to optimal estimator in the *non-robust* setting as shown in Proposition 4 in [GYZ20]. However, robustness requires the activation function to have bounded range. The solution taken by [GYZ20] is to add a constant offset b [4], which leads to the following family of Discriminator

$$\mathcal{T}_2 = \{D(\mathbf{x}) = \sigma(w\sigma(\mathbf{u}^T\mathbf{x}+b)) \text{ s.t } |w|\le\kappa, \mathbf{u}_j\in\mathbb{R}^d, b\in\mathbb{R}\}. \tag{3}$$

We can derive a Lemma related to covariance estimation that is very similar to Lemma 3.

Lemma 4. *Assume that $\|\Sigma\|_2 < M$ and C is a constant that depends only on M. Then, for $\mathcal{T}_2$ specified in Equation (3), if $\sup_{D\in\mathcal{T}_2}\mathcal{L}(\mathcal{N}(\mathbf{0},\Sigma),\mathcal{N}(\mathbf{0},\hat{\Sigma});D)\le C$ it holds*

$$\sup_{D\in\mathcal{T}_1}\mathcal{L}(\mathcal{N}(\theta,\mathbf{I}),\mathcal{N}(\hat{\theta},\mathbf{I});D)\gtrsim\kappa\frac{1}{2\sqrt{M}}\left\|\Sigma-\hat{\Sigma}\right\|_2,$$

for any $\kappa\in(0,1]$.

Proof Sketch. Following a similar argument as in the proof of Lemma 3, we derive that, for some Discriminator $D\in T_2$,

$$\mathcal{L}(\Sigma,\hat{\Sigma},D)\gtrsim\left|\mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\kappa\sigma(\delta_1 z+b)-\mathop{\mathbf{E}}_{z\sim\mathcal{N}(0,1)}\kappa\sigma(\delta_2 z+b)\right|,$$

where $\delta_1=\sqrt{\mathbf{u}^T\Sigma\mathbf{u}}, \delta_2=\sqrt{\mathbf{u}^T\hat{\Sigma}\mathbf{u}}$. When $\|\Sigma\|_2\le M$, we can again exploit the continuity of $f(t)=\mathbf{E}_{z\sim\mathcal{N}(0,1)}\,\sigma(t\cdot\delta z+b)$ around $t=1$ to obtain that

$$\sup_{D\in\mathcal{T}_2}\mathcal{L}(\Sigma,\hat{\Sigma},D)\gtrsim\kappa\left|\delta_1-\delta_2\right|\gtrsim\sup_{\|\mathbf{u}\|_2=1}\kappa\left|\delta_1^2-\delta_2^2\right|\frac{1}{\delta_1+\delta_2}=\kappa\left\|\Sigma-\hat{\Sigma}\right\|_2\frac{1}{2\sqrt{M}},$$

where the last inequality is obtained by taking $\mathbf{u}$ as the eigenvector corresponding to the largest eigenvalue of $\Sigma-\hat{\Sigma}$. □

4 Connection to Jensen-Shannon Divergence and Optimization

The typical interpretation for the Loss $\mathcal{L}(P_\theta,P_{\hat{\theta}};D)$ is as a divergence measure. Suppose the Discriminator is allowed to be any function of the form $\mathbb{R}^d\mapsto[0,1]$. It is easy to see that the cross-entropy loss $P(\mathbf{x})\log(D(\mathbf{x}))+Q(\mathbf{x})\log(1-D\mathbf{x})$ is point-wisely maximized when $D(\mathbf{x})=\frac{P(\mathbf{x})}{P(\mathbf{x})+Q(\mathbf{x})}$. Then, the Loss function becomes the Jensen-Shannon Divergence plus a constant offset.

$$\begin{aligned}\sup_D\mathcal{L}(P_\theta,P_{\hat{\theta}};D)&=\int_{\mathbf{x}}P(\mathbf{x})\log\frac{P(\mathbf{x})}{P(\mathbf{x})+Q(\mathbf{x})}+Q(\mathbf{x})\log\frac{Q(\mathbf{x})}{P(\mathbf{x})+Q(\mathbf{x})}d\mathbf{x}+2\log 2.\\&=\int_{\mathbf{x}}P(\mathbf{x})\log\left(\sigma(\log\frac{P(\mathbf{x})}{Q(\mathbf{x})})\right)+Q(\mathbf{x})\log\left(\sigma(\log\frac{Q(\mathbf{x})}{P(\mathbf{x})})\right)d\mathbf{x}+2\log 2.\end{aligned}$$

If so, the Loss function is for sure "descriptive" enough as it is itself a divergence measure.

[4] The offset is treated as a general parameter that can take any real value. However, from the proof part, we can tell all we need is just a constant offset.

Suppose P and Q are both Gaussian distributions with the form $P = \mathcal{N}(\mu_p, \mathbf{\Sigma}_p)$ and $Q = \mathcal{N}(\mu_q, \mathbf{\Sigma}_q)$, the term $\log \frac{P(\mathbf{x})}{Q(\mathbf{x})}$ has a closed form. Consequently, the optimal Discriminator is exactly

$$D(\mathbf{x}) = \sigma\left(\frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}_q^{-1} - \mathbf{\Sigma}_p^{-1})\mathbf{x} + \frac{1}{2}(\mu_p - \mu_q)^T\mathbf{x} + \frac{1}{2}\left(\|\mu_q\|_2^2 - \|\mu_p\|_2^2\right)\right). \tag{4}$$

Hence, for both mean and covariance estimation task, it is possible to come up with a Discriminator form that could reduce the loss exactly to the JS Divergence (though for covariance estimation, the Discriminator will have a weird quadratic layer).

Unfortunately, the closed form alone is unbounded, hence leading to non-robust estimators. As we have seen in previous section, one possible remedy is to wrap a "filtering" layer on top of the closed form. For example, if we add a sigmoid layer, we then have a Discriminator family very similar to $\mathcal{T}_1$ as specified in Equation (1) for mean estimation. Intuitively, the filtering layer enables Discriminator to focus on a smaller region where the influence of outliers become bounded. The interesting part is that it does not simply compute the divergence on truncated normal distributions where the truncation set has bounded radius. Instead, the truncation is performed solely on directions used by the Discriminator μ to approximate the divergence so that the approximation is robust to outliers and only a constant fraction of samples are lost.

The variational interpretation of the loss function is convenient in establishing sample complexity of the estimator. However, it does not distinguish GAN from other variational algorithm such as the Tukey Median, whose running time is shown to be exponential in dimension. Intuitively, GAN should have better running time as the "probing direction" is not arbitrarily chosen but a representative one in which the statistical distance of original distributions after projection is preserved.

From an optimization point of view, the non-robust learning task corresponds to a nonconvex-concave minimax game and stochastic gradient descent ascent can be used to find its equilibrium in polynomial number of iterations (See [LJJ20]). While the polluted samples can modify the landscape of the objective function, the filtering layer ensures the influence is limited (See Claim 1).

It is hence an interesting follow-up project to investigate how the landscape of Discriminator's objective function changes due to the presence of the filtering layer and what exactly the Loss function is reduced to with the new architecture. On top of that, one may connect it to literature studying the optimization of minimax game and obtain running time result for GAN based estimator.

References

[GLYZ18] Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets. *arXiv preprint arXiv:1810.02030*, 2018.

[GYZ20] Chao Gao, Yuan Yao, and Weizhi Zhu. Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective. *Journal of Machine Learning Research*, 21(160):1–48, 2020.

[LJJ20] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[NCT16] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

[WDHY20] Kaiwen Wu, Gavin Weiguang Ding, Ruitong Huang, and Yaoliang Yu. On minimax optimality of gans for robust mean estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4541–4551. PMLR, 2020.