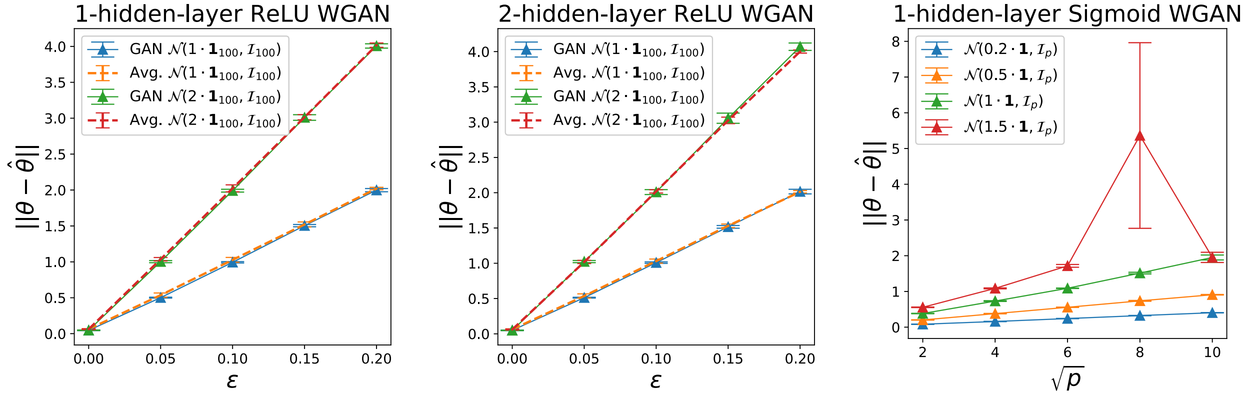
A Additional Experiments

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



(a) error *vs.* ϵ

(b) error *vs.* ϵ

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

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

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

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

B Technical Lemmas

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

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

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

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

In addition,

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

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

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

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

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

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

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

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

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

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

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

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

Hence,

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

which finishes the proof. □

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

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

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

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

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

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

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

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

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

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

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

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

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

In addition, we have

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

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

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

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

with probability at least $1 - \delta$.

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

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

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

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

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

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

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

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

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

which finishes the proof. □

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

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

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

Proof. On the one hand,

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

On the other hand,

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

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

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

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

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

we get

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

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

One can verify that

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

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

□

C f-GAN

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

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

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

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

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

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

then

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

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

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

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

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

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

We have

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

where

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

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

Thus

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

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

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

Thus

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

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

D MMD GAN

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\|\mu_{\mathbb{P}_\theta} - \mu_{\mathbb{P}_{\hat{\theta}}}\|_{\mathcal{H}}^2 &= \mathbf{E}_{x,x' \sim \mathbb{P}_\theta}[k(x,x')] - 2\mathbf{E}_{x\sim \mathbb{P}_\theta, x' \sim \mathbb{P}_{\hat{\theta}}}[k(x,x')] + \mathbf{E}_{x,x' \sim \mathbb{P}_{\hat{\theta}}}[k(x,x')] \\
&= 2\mathbf{E}_{x \sim N(0, 2I_p)}\left[\exp\left(-\frac{x^T x}{2\sigma^2}\right)\right] - 2\mathbf{E}_{x\sim N(\theta - \hat{\theta}, 2I_p)}\left[\exp\left(-\frac{x^T x}{2\sigma^2}\right)\right] \\
&= \sqrt{\left(\frac{\sigma^2}{2+\sigma^2}\right)^p}\left(1 - \exp\left(-\frac{1}{2(2+\sigma^2)}\|\hat{\theta} - \theta\|^2\right)\right).
\end{aligned}$$

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

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

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

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

Combining all of the above, we have proven that

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

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

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

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

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

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

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

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

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

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

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

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

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

We first prove the following claim.

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

If the claim holds, then

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

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

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

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

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

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

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

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

Further,

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

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

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

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

E Wasserstein GAN

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

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

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

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

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

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

Decompose (40) into two terms:

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

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

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

where ϕ is the density of the standard Gaussian distribution.

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

Decompose (40) into three terms:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

F Adaptation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

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

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

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

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

Define

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

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

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

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

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

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

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

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

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

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

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

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