# A. Proofs of properties of the SQ hard distribution

We start with the following lemma on Hermite polynomials:

**Lemma A.1.** *For every $k > 1$, the distance between any roots of $H_{k-1}(t)$ and $H_k(t)$ is at least $\Omega(1/\sqrt{k})$.*

*Proof.* It is known that extrema of $H_k$ are exactly zeros of $H_{k-1}$, which follows from $H_k' = 2kH_{k-1}$ and a lack of double roots. Thus, it is enough to show that extrema and zeros of $H_k$ are $\Omega(1/\sqrt{k})$-separated.

Consider the case where $0 \le u < v < w$ are such that $H_k(u) = H_k(w) = 0$, $H_k$ is positive between $u$ and $w$, and $H_k'(v) = 0$. Let us show how to lower bound $v - u$. Denote $F_k(t) = e^{-t^2/2}H_k(t)$. Clearly, $F_k(u) = F_k(w) = 0$ and $F_k$ is positive between $u$ and $w$ with a unique local maximum on $[u, w]$, which we denote by $v'$. It is not hard to check that $v' \le v$. Thus, it is enough to lower bound $v' - u$. It is known (see, e.g., (Szego, 1939)) that $F_k$ satisfies the ODE $Z'' + (2k + 1 - t^2)Z = 0$. By comparing with $Z'' + (2k + 1)Z = 0$, we can get that lower bound $v - u \ge v' - u \ge \frac{\pi}{2\sqrt{2k+1}} = \Omega(1/\sqrt{k})$.

Now let us lower bound $w - v$. It is known (Szego, 1939) that $H_k$ satisfies the ODE $Z'' - 2tZ' + 2kZ = 0$. By comparing this ODE with $Z'' - 2wZ' + 2kZ = 0$, we get that $w - v \ge \frac{\arctan\left(\frac{\sqrt{2k-w^2}}{w}\right)}{\sqrt{2k-w^2}} \ge \Omega(1/\sqrt{k})$. The latter step is due to $w \le \sqrt{2k}$ and that the lower bound on $w - v$ is nonincreasing in $w$.

Other cases can be treated similarly. $\square$

**Lemma 4.2.** *There exist two distributions $D_A$ and $D_B$ over $\mathbb{R}$ with everywhere positive p.d.f.'s $A(t)$ and $B(t)$ respectively such that:*

- *$D_A$ and $D_B$ match $N(0, 1)$ in the first $m$ moments;*
- *There exist two subsets $S_A, S_B \subset \mathbb{R}$ such that the distance between $S_A$ and $S_B$ is at least $\Omega(1/\sqrt{m})$, $\mathbb{P}_{x\sim D_A}[x \in S_A] \ge 1 - e^{-\Omega(m)}$, and $\mathbb{P}_{x\sim D_B}[x \in S_B] \ge 1 - e^{-\Omega(m)}$;*
- *$A, B \in C^\infty$, and for every $0 \le l \le m + 1$ and $t$, one has: $|\frac{d^l}{dt^l}\frac{A(t)}{G(t)}|, |\frac{d^l}{dt^l}\frac{B(t)}{G(t)}| \le m^{O(l+1)}$.*

*(See Figure 1 for the illustration.)*

*Proof.* Let $H_m(t)$ and $H_{m+1}(t)$ be two consecutive (physicist's) Hermite's polynomials. It is a classic result in Gaussian quadrature (see, e.g., (Szego, 1939)) that for every $k$, there exists a discrete distribution supported on the zeros of $H_k(t/\sqrt{2})$, which matches $N(0, 1)$ in the first $2k - 1$ moments. Let $\widetilde{D}_A$ denote such a distribution for $H_m$ and $\widetilde{D}_B$ the same for $H_{m+1}$. By Lemma A.1, the distance between the supports of $\widetilde{D}_A$ and $\widetilde{D}_B$ is at least $\Omega(1/\sqrt{m})$ and they both match $N(0, 1)$ in the first $2m - 1 \ge m$ moments.

Now, we obtain the desired distributions $D_A$ and $D_B$ as follows. Fix a small $\delta > 0$. The distribution $D_A$ is defined as $\sqrt{1-\delta} \cdot x + \sqrt{\delta} \cdot y$, where $x \sim \widetilde{D}_A$, $y \sim N(0, 1)$, and $x$ and $y$ are independent. The distribution $D_B$ is defined similarly, but instead of $\widetilde{D}_A$ we use $\widetilde{D}_B$. It is easy to check that $D_A$ and $D_B$ match the first $m$ moments of $N(0, 1)$. Now suppose that $\delta = 1/m^2$. The second property follows from the supports of $\widetilde{D}_A$ and $\widetilde{D}_B$ being $\Omega(1/\sqrt{m})$ separated and the standard concentration inequalities; specifically, we take $S_A$ to be the Minkowski sum of the support of scaled down $\widetilde{D}_A$ and the ball of radius $\Theta(1/\sqrt{m})$, and $S_B$ to be similar with $\widetilde{D}_B$ instead of $\widetilde{D}_A$. Then the chance $x \sim D_A$ is not in $S_A$ is at most the chance $y \sim N(0, 1)$ has $|\sqrt{\delta}y| > \Omega(1/\sqrt{m})$, which is $e^{-\Omega(m)}$.

Now let us prove the bounds on $\frac{d^l}{dt^l}\frac{A(t)}{G(t)}$, for the $B(\cdot)$ similar bounds follows exactly the same way.

Denote $x_1 < x_2 < \ldots < x_m$ the roots of $H_m(t)$.

One has:

$$A(t) = \frac{1}{\sqrt{2\pi\delta}} \sum_{i=1}^{m} p_i e^{-\frac{\left(t-\sqrt{2(1-\delta)}x_i\right)^2}{2\delta}},$$

where $p_i = \mathbb{P}_{x \sim \widetilde{D}_A}[x = \sqrt{2}x_i]$. Hence,

$$\begin{aligned}\frac{A(t)}{G(t)} &= \frac{1}{\sqrt{\delta}} \sum_{i=1}^{m} p_i e^{-\frac{\left(t-\sqrt{2(1-\delta)}x_i\right)^2}{2\delta}+\frac{t^2}{2}} \\ &= \frac{1}{\sqrt{\delta}} \sum_{i=1}^{m} p_i e^{-\frac{1}{2}\cdot\left(\left(t\cdot\sqrt{\frac{1}{\delta}-1}-\sqrt{\frac{2}{\delta}}\cdot x_i\right)^2-2x_i^2\right)}. \\ &= \frac{1}{\sqrt{\delta}} \sum_{i=1}^{m} p_i e^{-\frac{1}{2}\cdot\frac{(t\cdot\sqrt{1-\delta}-\sqrt{2}x_i)^2}{\delta}+x_i^2}.\end{aligned}$$

We have for every $i$ the bound $p_i e^{x_i^2} = O(1)$ (Gil et al., 2018). Therefore, if $Q(t)$ denotes the p.d.f. of $N(0, \delta/(1-\delta))$ we have

$$\sup_t \left|\frac{d^l}{dt^l}\frac{A(t)}{G(t)}\right| \le \frac{1}{\sqrt{\delta}} \cdot m \cdot O(1) \cdot \left(\sup_t \left|\frac{d^l}{dt^l}Q(t)\right|\right) = m(l/\delta)^{O(l+1)} = m^{O(l+1)}.$$

□

**Lemma 4.3.** *For every $k \le d^{\Omega(1)}$, there exists such a family $\mathcal{U}$ with $\varepsilon \le d^{-0.49}$ and $|\mathcal{U}| = 2^{d^{\Theta(1)}}$.*

*Proof.* Let $U$ and $V$ be uniformly random $k$-dimensional subspaces of $\mathbb{R}^d$. W.l.o.g. we can assume that $U$ is spanned by the first $k$ standard basis vectors. Let $V$ be spanned by an orthonormal basis $v_1, v_2, \ldots, v_k$ such that each $v_i$ is distributed uniformly on the unit sphere of $\mathbb{R}^d$. Consider an $\varepsilon'$-net $\mathcal{N}$ of the unit sphere of $U$ of size $(1/\varepsilon')^{O(k)}$. For every $u \in \mathcal{N}$ with probability at least $1 - e^{-\Omega(\varepsilon'^2 d)}$ the absolute value of the dot product of $u$ with a given $v_i$ is at most $\varepsilon'$. As a result, with probability at least $1 - (1/\varepsilon')^{O(k)} e^{-\Omega(\varepsilon'^2 d)}$, dot products between all elements of $\mathcal{N}$ and all $v_i$ are at most $\varepsilon'$ in the absolute value. But this implies that the dot products between all the unit vectors of $U$ and $V$ are at most $\varepsilon'\sqrt{k}$. So, by setting $\varepsilon' = \varepsilon/\sqrt{k}$ and by using the union bound, we get that we can set:

$$\log|\mathcal{U}| \le \Omega(\varepsilon^2 d/k) - O(k(\log k + \log(1/\varepsilon))).$$

Thus, we can set $\varepsilon = d^{-0.49}$, and $k \le d^{\sigma}$ for a sufficiently small positive $\sigma$, which yields $|\mathcal{U}| = 2^{d^{\Theta(1)}}$. □

**Lemma 4.4.** *There exist two sets $S_{U,A}, S_{U,B} \subset \mathbb{R}^d$ such that the distance between $S_{U,A}$ and $S_{U,B}$ is $\Omega(\sqrt{k/m})$, and for which $\mathbb{P}_{x \sim D_{U,A}}[x \in S_{U,A}] \ge 1 - e^{-\Omega(km)}$ and $\mathbb{P}_{x \sim D_{U,B}}[x \in S_{U,B}] \ge 1 - e^{-\Omega(km)}$.*

*Proof.* The sets are defined as follows:

$$S_{U,A} = \{x \in \mathbb{R}^d \mid \text{for at least } 0.9\text{-fraction of } 1 \le i \le k, \text{ one has } \langle x, u_i\rangle \in S_A\}$$

and

$$S_{U,B} = \{x \in \mathbb{R}^d \mid \text{for at least } 0.9\text{-fraction of } 1 \le i \le k, \text{ one has } \langle x, u_i\rangle \in S_B\}.$$

The points $x \in S_{U,A}$ and $y \in S_{U,B}$ are well-separated, since in at least a 0.8-fraction of $1 \le i \le k$, both $\langle x, u_i\rangle \in S_A$ and $\langle y, u_i\rangle \in S_B$. Since $S_A$ and $S_B$ are $\Omega(1/\sqrt{m})$-separated, we obtain the result.

The bounds on the probabilities follow from the respective bounds in Lemma 4.2 and standard Chernoff bounds. □

# B. SQ lower bound

## B.1. SQ lower bound

Now let us show that if we set all the parameters appropriately, it is hard in the SQ model to learn a good classifier (robust or otherwise) for distributions $D_{U,A}$ and $D_{U,B}$ defined above, where $U \in \mathcal{U}$ is an unknown subspace. The main idea is to show that if the subspace $U \in \mathcal{U}$ is chosen uniformly at random, unless we perform more than $2^{d^{\Omega(1)}}$ queries, we can not tell apart $D_{U,A}$ or $D_{U,B}$ from the standard Gaussian $N(0, I_d)$ (and as a result, from each other). Intuitively, any since query can only reliably distinguish $D_{U,A}$ from $N(0, I_d)$ for a tiny fraction of subspaces $U \in \mathcal{U}$. The result then follows by a simple counting argument. To formalize the above intuition, we use an argument similar at a high-level to the one used in (Diakonikolas et al., 2017).

Let $D, D_1, D_2$ be distributions over $\mathbb{R}^d$ with everywhere positive p.d.f.'s $P(x)$, $P_1(x)$, and $P_2(x)$, respectively. Then, the pairwise correlation of $D_1$ and $D_2$ w.r.t. $D$, denoted by $\chi_D(D_1, D_2)$, is defined as follows:

$$\chi_D(D_1, D_2) = \int_{\mathbb{R}^d} \frac{P_1(x)P_2(x)}{P(x)}\, dx - 1.$$

In Section B.2, we show that for an appropriate setting of parameters (namely, when $\varepsilon m^{\Theta(1)} k \le d^{-\Omega(1)}$), for every $U_1, U_2 \in \mathcal{U}$, one has:

$$\chi_{N(0,I_d)}(D_{U_1,A}, D_{U_2,A}) \le \begin{cases} m^{O(k)} & \text{if } U_1 = U_2 \\ m^{O(k)} \cdot d^{-\Omega(m)} & \text{otherwise} \end{cases}$$

and

$$\chi_{N(0,I_d)}(D_{U_1,B}, D_{U_2,B}) \le \begin{cases} m^{O(k)} & \text{if } U_1 = U_2 \\ m^{O(k)} \cdot d^{-\Omega(m)} & \text{otherwise.} \end{cases}$$

Then by repeating the proof of Lemma 3.3 from (Feldman et al., 2013), we get that if the number of queries is significantly smaller than:

$$\frac{|\mathcal{U}| \cdot (\tau^2 - m^{O(k)} d^{-\Omega(m)})}{m^{O(k)}},$$

then with high probability over a random subspace $U \in \mathcal{U}$, all the queries asked can be answered as if both $D_{U,A}$ and $D_{U,B}$ were $N(0, I_d)$. As a result, we cannot distinguish them from $N(0, I_d)$ and, as a result, between each other.

Suppose that $m \log d > Ck \log m$ for a sufficiently large constant $C$, so that the $m^{O(k)} d^{-\Omega(m)}$ term is less than $d^{-\Omega(m)} < m^{-\Omega(k)}$. Then we can set the precision $\tau$ to $m^{-\Theta(k)}$ and still be unable to distinguish $D_{U,A}$ from $D_{U,B}$ from $|\mathcal{U}| m^{-O(k)} = 2^{d^{\Omega(1)}} m^{-O(k)}$ queries. If $m^{O(k)} \le 2^{d^\sigma}$ for a sufficiently small positive $\sigma > 0$, this gives the desired lower bound of $2^{d^{\Omega(1)}}$ on the number of SQ queries the algorithm must ask.

### B.2. Upper bounding pairwise correlations

In this section, we show how to upper bound $\chi_{N(0,I_d)}(D_{U_1,A}, D_{U_2,A})$; upper bounding $\chi_{N(0,I_d)}(D_{U_1,B}, D_{U_2,B})$ is exactly the same. Denote $a(t) = \frac{A(t)}{G(t)} - 1$, where $G(t)$ is the p.d.f. of a standard Gaussian. By Lemma 4.2, one has $\mathbb{E}_{t \sim N(0,1)}[t^l \cdot a(t)] = 0$ for all $l \in \{1, 2, \ldots, m\}$.

We assume that $m^C \varepsilon k \le d^{-\Omega(1)}$ for a sufficiently large constant $C$ to be determined later. Since by Lemma 4.3 we can take $\varepsilon = d^{-0.49}$, the required inequality holds as long as $m$ and $k$ are at most small powers of $d$.

First, suppose that $U_1 = U_2 = U$. Suppose that $u_1, u_2, \ldots, u_d$ is an orthonormal basis of $\mathbb{R}^d$ such that $u_1, u_2, \ldots, u_k$ is a fixed basis of $U$. Then,

$$\begin{aligned}
\chi_{N(0,1)^d}(D_{U,A}, D_{U,A}) &= \int_{\mathbb{R}^d} \frac{A_U(x)^2}{\prod_{i=1}^d G(\langle x, u_i \rangle)}\, dx - 1 \\
&= \int_{\mathbb{R}^d} \prod_{i=1}^k (1 + a(\langle x, u_i \rangle))^2 \cdot \prod_{i=1}^d G(\langle x, u_i \rangle)\, dx - 1 \\
&= \mathbb{E}_{x \sim N(0,I_d)} \left[ \prod_{i=1}^k (1 + a(\langle x, u_i \rangle))^2 \right] - 1 \\
&= \prod_{i=1}^k \mathbb{E}_{x \sim N(0,I_d)} \left[ (1 + a(\langle x, u_i \rangle))^2 \right] - 1 \\
&\le m^{O(k)},
\end{aligned}$$

where the fourth step is due to the independence of $\langle x, u_i \rangle$ (which is implied by orthogonality of $u_i$), and the fifth step follows from Lemma 4.2.

Now suppose that $U_1 \neq U_2$. Suppose that $u_1, u_2, \ldots, u_d$ is an orthonormal basis of $\mathbb{R}^d$ such that $u_1, u_2, \ldots, u_k$ is a fixed basis of $U_1$, and, similarly, $v_1, v_2, \ldots, v_d$ is an orthonormal basis of $\mathbb{R}^d$ such that $v_1, v_2, \ldots, v_k$ is a fixed basis of $U_2$. Now,

$$\begin{aligned}
\chi_{N(0,I_d)}(D_{U_1,A}, D_{U_2,A}) &= \int \frac{A_{U_1}(x)A_{U_2}(x)}{\prod_{i=1}^d G(x_i)}\,dx - 1 \\
&= \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i=1}^k\Big(1 + a(\langle x, u_i\rangle)\Big)\cdot\prod_{i=1}^k\Big(1 + a(\langle x, v_i\rangle)\Big)\right] - 1 \\
&= \sum_{S,T\subseteq[k]} \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T} a(\langle x, v_i\rangle)\right] - 1 \\
&= \sum_{\substack{S,T\subseteq[k]:\\ S,T\neq\emptyset}} \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T} a(\langle x, v_i\rangle)\right], \qquad (2)
\end{aligned}$$

where the last step follows from the fact that if $S = \emptyset$ and $T \neq \emptyset$, then the expression factorizes due to the independence of $\langle x, v_i\rangle$, and we also use that $\mathbb{E}_{t\sim N(0,1)}[a(t)] = 0$. The case $S \neq \emptyset$ and $T = \emptyset$ is similar.

Now let us fix non-empty $S, T \subseteq [k]$. W.l.o.g., suppose that $|S| \geq |T|$. Denote $\widetilde{v}_i = v_i - \mathrm{proj}_{U_1} v_i$. Since $U_1, U_2 \in \mathcal{U}$ and $U_1 \neq U_2$, we have $\|\widetilde{v}_i - v_i\|_2 \leq \varepsilon$. One has for every $1 \leq i \leq k$ by a Taylor expansion that

$$a(\langle x, v_i\rangle) = \sum_{l=0}^m a^{(l)}(\langle x, \widetilde{v}_i\rangle)\cdot\frac{\langle x, v_i - \widetilde{v}_i\rangle^l}{l!} + a^{(m+1)}(\theta_i)\cdot\frac{\langle x, v_i - \widetilde{v}_i\rangle^{m+1}}{(m+1)!}, \qquad (3)$$

for some $\theta_i = \theta_i(x)$ that lies between $\langle x, \widetilde{v}_i\rangle$ and $\langle x, v_i\rangle$.

**Lemma B.1.** *Suppose $|S| \geq |T|$. For every $l\colon T \to \{0, 1, \ldots, m\}$, one has:*

$$\mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T}\left(a^{(l(i))}(\langle x, \widetilde{v}_i\rangle)\cdot\frac{\langle x, v_i - \widetilde{v}_i\rangle^{l(i)}}{l(i)!}\right)\right] = 0.$$

*Proof.* Since $v_i - \widetilde{v}_i \in U_1$, we can write $v_i - \widetilde{v}_i = \sum_{j=1}^k \alpha_{ij} u_j$. One has:

$$\begin{aligned}
&\mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T}\left(a^{(l(i))}(\langle x, \widetilde{v}_i\rangle)\cdot\frac{\langle x, v_i - \widetilde{v}_i\rangle^{l(i)}}{l(i)!}\right)\right] \\
&= \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T}\left(a^{(l(i))}(\langle x, \widetilde{v}_i\rangle)\cdot\frac{\left(\sum_{j=1}^k \alpha_{ij}\langle x, u_j\rangle\right)^{l(i)}}{l(i)!}\right)\right] \\
&= \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T}\left(a^{(l(i))}(\langle x, \widetilde{v}_i\rangle)\cdot\frac{1}{l(i)!}\cdot\sum_{\beta_{ij}:\sum_{j=1}^k \beta_{ij} = l(i)}\binom{l(i)}{\beta_{i1}\ldots\beta_{ik}}\prod_{j=1}^k(\alpha_{ij}\langle x, u_j\rangle)^{\beta_{ij}}\right)\right] \\
&= \sum_{\substack{\beta_{ij}\\ \forall i:\sum_{j=1}^k \beta_{ij} = l(i)}}\left(\prod_{i\in T}\frac{\binom{l(i)}{\beta_{i1}\ldots\beta_{ik}}}{l(i)!}\right)\cdot\mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T}\left(a^{(l(i))}(\langle x, \widetilde{v}_i\rangle)\cdot\prod_{j=1}^k(\alpha_{ij}\langle x, u_j\rangle)^{\beta_{ij}}\right)\right].
\end{aligned}$$

Now let us fix partitions $\beta_{ij}$ and show that:

$$\mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle)\cdot\prod_{i\in T}\left(a^{(l(i))}(\langle x, \widetilde{v}_i\rangle)\cdot\prod_{j=1}^k(\alpha_{ij}\langle x, u_j\rangle)^{\beta_{ij}}\right)\right] = 0. \qquad (4)$$

Since $\sum_{ij}\beta_{ij} = \sum_i l(i) \le |T| \cdot m$, there exists $j^* \in S$ such that: $\sum_i \beta_{ij^*} \le \frac{|T|\cdot m}{|S|} \le m$. Since $\langle x, u_{j^*}\rangle$ is independent from the remaining dot products, we can factor from (4) the expression

$$\mathbb{E}_{x\sim N(0,I_d)}[a(\langle x, u_{j^*}\rangle)\langle x, u_{j^*}\rangle^l] \tag{5}$$

with $l \le m$. But since $\langle x, u_{j^*}\rangle$ is distributed as $N(0,1)$, one has that (5) is equal to zero due to Lemma 4.2. $\square$

Let us continue upper bounding (2). For $i \in T$ and $0 \le j \le m+1$, denote:

$$\gamma_{ij} = \begin{cases} v_i - \widetilde{v}_i, \text{if } j \le m, \\ \theta_i, \text{if } j = m+1. \end{cases}$$

One has:

$$\begin{aligned}
&\mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle) \cdot \prod_{i\in T} a(\langle x, v_i\rangle)\right] \\
&= \sum_{l\colon T\to\{0,1,\ldots,m+1\}} \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle) \cdot \prod_{i\in T} a^{(l(i))}(\gamma_{i,l(i)}) \cdot \frac{\langle x, v_i - \widetilde{v}_i\rangle^{l(i)}}{l(i)!}\right] \\
&= \sum_{\substack{l\colon T\to\{0,1,\ldots,m+1\} \\ l^{-1}(m+1)\neq\emptyset}} \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in S} a(\langle x, u_i\rangle) \cdot \prod_{i\in T} a^{(l(i))}(\gamma_{i,l(i)}) \cdot \frac{\langle x, v_i - \widetilde{v}_i\rangle^{l(i)}}{l(i)!}\right] \\
&\le \sum_{\substack{l\colon T\to\{0,1,\ldots,m+1\} \\ l^{-1}(m+1)\neq\emptyset}} (\sup_t |a(t)|)^{|S|} \cdot \left(\prod_{i\in T} \frac{\sup_t |a^{l(i)}(t)|}{l(i)!}\right) \cdot \mathbb{E}_{x\sim N(0,I_d)}\left[\prod_{i\in T} \|\mathrm{proj}_{U_1} x\|_2^{l(i)} \cdot \|v_i - \widetilde{v}_i\|_2^{l(i)}\right] \\
&\le \sum_{\substack{l\colon T\to\{0,1,\ldots,m+1\} \\ l^{-1}(m+1)\neq\emptyset}} m^{O(|S|)} \cdot \left(\prod_{i\in T} \frac{m^{O(l(i))} \cdot \varepsilon^{l(i)}}{l(i)!}\right) \cdot \mathbb{E}_{y\sim N(0,I_k)}\left[\|y\|_2^{\sum_{i\in T} l(i)}\right] \\
&\le \sum_{\substack{l\colon T\to\{0,1,\ldots,m+1\} \\ l^{-1}(m+1)\neq\emptyset}} \frac{m^{O\left(|S| + \sum_{i\in T} l(i)\right)} \cdot \varepsilon^{\sum_{i\in T} l(i)} \cdot \left(k + \sum_{i\in T} l(i)\right)^{\sum_{i\in T} l(i)}}{\prod_{i\in T} l(i)!} \\
&\le \sum_{\substack{l\colon T\to\{0,1,\ldots,m+1\} \\ l^{-1}(m+1)\neq\emptyset}} \frac{m^{O(k)} d^{-\Omega\left(\sum_{i\in T} l(i)\right)}}{\prod_{i\in T} l(i)!} \\
&\le m^{O(k)} \cdot d^{-\Omega(m)},
\end{aligned} \tag{6}$$

where the first step follows from (3), the second step follows from Lemma B.1, the third step follows from Cauchy–Schwartz, the fourth step follows from Lemma 4.2 and from the bound $\|v_i - \widetilde{v}_i\|_2 \le \varepsilon$, the fifth step follows from the inequality $\mathbb{E}_{y\sim N(0,I_k)}[\|y\|_2^s] \le (k+s)^s$, the sixth step follows from $(\varepsilon m^{\Theta(1)} k) = d^{-\Omega(1)}$ and from $\sum_{i\in T} l(i) \le O(mk)$, and the last step follows from dropping the denominators, the sum having at most $(m+2)^{|T|} = m^{O(k)}$ terms, and that $\sum_i l(i) \ge m+1$.

Plugging (6) into (2), we get the result.

### B.3. Setting parameters

We obtain a $\Omega(\sqrt{k/m})$-robust classifier, and the precision of statistical queries can be as high as $m^{O(k)} \cdot d^{-\Omega(m)}$. Thus, for $0 < \gamma < 1/10$, we can set $m = d^{\Theta(\gamma)}$ and $k \ll \frac{m \log d}{\log m}$. As a result we get robustness $\Omega(\sqrt{k/m}) = \Omega(\sqrt{\log d / \log m}) = \Omega(\sqrt{1/\gamma})$, and the precision of statistical queries can be as good as $2^{-d^{\Omega(\gamma)}}$.

## C. Bound on covering number of generative models

**Lemma C.1.** *Let $g_w$ be a $\ell$-layer neural network architecture with at most $d$ activations in each layer and Lipschitz nonlinearities such as ReLUs. Then*

$$\|g_w(x) - g_{w'}(x)\|_2 \leq \|w - w'\|_1 \cdot \|x\|_2 \cdot (dB)^\ell$$

*Proof.* By the triangle inequality, it suffices to consider $w$ and $w'$ that differ in a single coordinate. Suppose this coordinate is in layer $i$. Since each layer's weight matrix $w_i$ has $\|w_i\| \leq \|w_i\|_F \leq dB$, and the initial layer has activation $\|x\|_2$, the $\ell_2$ norm of the activations in the $i$th layer is at most $\|x\|_2(dB)^i$. Therefore the change in activation in layer $i+1$ is at most $\|w - w'\|_1 \cdot \|x\|_2(dB)^i$, and the change in the last layer is at most $\|w - w'\|_1 \cdot \|x\|_2(dB)^{\ell-1}$. □

**Lemma 3.7.** *Let $g_w$ be an $\ell$-layer neural network architecture with at most $d$ activations in each layer and Lipschitz nonlinearities such as ReLUs. Consider any family of distribution pairs $\mathcal{D}$ such that for each $D \in \mathcal{D}$, and each $i \in \{0, 1\}$, there exists some $w \in [-B, B]^m$ with $W_\infty(D_i, D(g_w)) \leq \varepsilon$. Then*

$$\log\left(\mathcal{N}_{W_\infty,\mathrm{TV}}(\mathcal{D}, \varepsilon + \delta, \delta)\right) \leq O(m\ell \log(dB/\delta)).$$

*Proof.* First, consider any $w \in [-B, B]^m$ and $x \in \mathbb{R}^k$, and let $w'$ differ from $w$ in a single weight.

For some parameter $\alpha > 0$, we consider the net $\widetilde{\mathcal{N}} = \{D(g_w) \mid w \in [-B, B]^m \cap \alpha\mathbb{Z}^m\}$. Our cover of $\mathcal{D}$ will be $\mathcal{N} \times \mathcal{N}$. This has size $(1 + \frac{2B}{\alpha})^{2m}$, which is sufficiently small as long as $\alpha = (dB/\delta)^{-O(\ell)}$.

It suffices to show for each $D \in \mathcal{D}$ and $i \in \{0, 1\}$ that $D_i \in U_{\varepsilon+\delta,\delta}(\widetilde{D})$ for some $\widetilde{D} \in \widetilde{\mathcal{N}}$. Let $w^*$ be the $w$ for which $W_1(D_i, D(g_w)) \leq \varepsilon$ and $\widehat{w}$ be the nearest $w$ in our cover, so $\|\widehat{w} - w^*\|_\infty \leq \alpha$. Then for any $x \in \mathbb{R}^k$ with $\|x\|_2 \leq \sqrt{k}/\delta$,

$$\|g_{\widehat{w}}(x) - g_{w^*}(x)\|_2 \leq \delta$$

by Lemma C.1 and our chosen $\alpha$. Since $\|x\|_2 \leq \sqrt{k}/\delta$ with probability much higher than $1 - \delta$, this implies $D(g_{w^*}) \in U_{\delta,\delta}(D(g_{\widehat{w}}))$. The triangle inequality then gives $D_i \in U_{\varepsilon+\delta,\delta}(D(g_{\widehat{w}}))$ as desired. □