# A Spherical harmonic decomposition and kernel spectrum

Any function defined on the unit sphere has a spherical harmonic decomposition

$$g(x,y) = \sum_u \gamma_u \phi_u(x)\phi_u(y), \tag{29}$$

where $\phi_u(x) : \mathbb{S}^{d-1} \mapsto \mathbb{R}$ is a spherical harmonic basis. Note that $u = (t, j)$ is a multi-index: the first denotes the order of the basis and the latter denotes the index of bases with the same order.

For each order $t$, there are $N(d,t) = \frac{2t+d-2}{t}\begin{pmatrix} t+d-3 \\ t-1 \end{pmatrix}$ bases with the same coefficient. As a result, the spectrum $\gamma_u$ sorted by magnitude has the step like shape where each step is of length $N(d,t)$.

To compute the coefficients, we use the Legendre harmonics [Müller, 2012] with the following property

$$P_{t,d}(\langle x, y\rangle) = \frac{1}{N(d,t)} \sum_{j=1}^{N(d,t)} \phi_{t,j}(x)\phi_{t,j}(y). \tag{30}$$

The spherical harmonics also form an orthonormal basis on the unit sphere:

$$\mathbb{E}\left[\phi_{l,i}(x)\phi_{m,j}(x)\right] = \frac{1}{\left|\mathbb{S}^{d-1}\right|}\int_{\mathbb{S}^{d-1}} \phi_{l,i}(x)\phi_{m,j}(x)dx = \delta_{lm}\delta_{ij}, \tag{31}$$

where $\left|\mathbb{S}^{d-1}\right|$ denotes the surface area of the unit sphere.

Combining these properties, we can calculate the spectrum using

$$\gamma_{(t,j)} = \frac{\left|\mathbb{S}^{d-2}\right|}{\left|\mathbb{S}^{d-1}\right|}\int_{-1}^{1} g(\xi)P_{t,d}(\xi)(1-\xi^2)^{(d-3)/2}d\xi, \quad \text{for all } j \in [N(d,t)]\,. \tag{32}$$

# B Bounding $\lambda_m(G)$ using matrix concentration bound: Proof of Lemma 6

Recall that

$$g(x,y) = \sum_{u=1}^{\infty} \gamma_u \phi_u(x)\phi_u(y). \tag{33}$$

For an integer $r > 0$, define the truncated version of $g$ and the corresponding residue as

$$g^{[r]}(x,y) = \sum_{u=1}^{r} \gamma_u \phi_u(x)\phi_u(y) \tag{34}$$

$$e^r(x,y) = g(x,y) - g^{[r]}(x,y) \tag{35}$$

and define the matrices

$$\left[G^{[r]}\right]_{i,j} = g^{[r]}(x_i, x_j)$$
$$E^r = G - G^{[r]}. \tag{36}$$

**Lemma 10** *Let $c_g = \max_x g(x,x)$ then with probability at least $1 - m\exp\left(-\frac{m\gamma_m}{8c_g}\right)$,*

$$\lambda_m(G^{[m]}) \geq m\gamma_m/2.$$

**Proof** Define a matrix $A$ whose rows are

$$A^i := [\sqrt{\gamma_1}\phi_1(x_i), \ldots, \sqrt{\gamma_m}\phi_m(x_i)]$$

for $1 \leq i \leq m$. Define matrices

$$X_i = (A^i)^\top A^i.$$

Denote $Y = \sum_{i=1}^m X_i$. Then $\lambda_m(\mathbb{E}Y) = m\gamma_m$ using the fact that $\mathbb{E}\left[\phi_i(x)\phi_j(x)\right] = \delta_{ij}$. Furthermore, $X_i \succeq 0$ and

$$\|X_i\| \leq \mathrm{tr}(X_i) = \sum_{u=1}^{m} \gamma_u \phi_u^2(x_i) \leq g(x_i, x_i) = c_g.$$

Therefore, matrix Chernoff bound (e.g., [Tropp, 2012]) gives

$$\Pr\left[\lambda_m(Y) \leq (1-\epsilon)\lambda_m(\mathbb{E}Y)\right] \leq m \exp\left(-(1-\epsilon)^2 \lambda_m(\mathbb{E}Y)/(2c_g)\right).$$

Choose $\epsilon = 1/2$ and use the facts that $G^{[m]} = AA^\top$, $Y = A^\top A$ and $\lambda_m(G^{[m]}) = \lambda_m(Y)$, we finish the proof. ■

**Proof** [Proof of Lemma 6] By Weyl's theorem and the fact that $E^m$ is PSD,

$$\lambda_m(G) \geq \lambda_m(G^{[m]}) + \lambda_m(E^m) \geq \lambda_m(G^{[m]}).$$

Lemma 6 then follows from Lemma 10. ■

## C Bounding the difference between $\lambda_m(G)$ and $\lambda_m(G_n)$: Proof of Lemma 7

Using Weyl's theorem we have that

$$|\lambda_m(G_n) - \lambda_m(G)| \leq \|G_n - G\|. \tag{37}$$

We are going to give an upper bound on $\|G_n - G\|$:

$$\|G_n - G\| = \sup_{\|\alpha\|=1} \sum_{i,j=1}^{m} \alpha_i \alpha_j (x_i^\top x_j) E_{ij}, \tag{38}$$

$$\text{where } E_{i,j} = \frac{1}{n}\sum_{k=1}^{n} \sigma'(w_k^\top x_i)\sigma'(w_k^\top x_j) - \mathbb{E}_w[\sigma'(w^\top x_i)\sigma'(w^\top x_j)], \tag{39}$$

and the first expectation is taken over $w$ uniformly on the sphere $\mathbb{S}^{d-1}$.

Our bound heavily relies on the inner products $|\langle x_i, x_j\rangle|$ for all $i \neq j$ being small enough. In the next lemma, we provide such a result for uniformly distributed data.

**Lemma 11 (Tail bound for spherical distribution)** *If $a$ and $b$ are independent vectors uniformly distributed over the unit sphere $\mathbb{S}^{d-1}$, then there exists a constant $c > 0$, such that for any $u > 0$,*

$$\Pr\left[|\langle a, b\rangle| \geq \frac{cu}{\sqrt{d}}\right] \leq 2e^{-u^2}.$$

**Proof** Note that both $a$ and $b$ are sub-gaussian random variables with sub-gaussian norm $c/\sqrt{d}$ where $c$ is some constant [Vershynin, 2010].

Denote $\mathbb{E}_b[\cdot]$ the expectation over $b$. We can rewrite the probability as

$$\begin{aligned}\Pr\left[|\langle a, b\rangle| \geq \frac{cu}{\sqrt{d}}\right] &\leq \mathbb{E}_b \Pr\left[|\langle a, b\rangle| \geq \frac{cu}{\sqrt{d}} \mid b\right] \\ &\leq \mathbb{E}_b\left\{2\exp\left(-u^2\right)\right\} = 2\exp\left(-u^2\right).\end{aligned} \tag{40}$$

The last inequality uses the independence of $a$ and $b$ and $\|\langle a, b\rangle\|_{\psi_2} \leq \|b\|_2 \|a\|_{\psi_2}$ for a fixed $b$. ■

Decomposing the sum into diagonal and off-diagonal terms gives us

$$\|G_n - G\| \leq \sup_{\|\alpha\|=1} \sum_{i\neq j}^{m} \alpha_i \alpha_j \langle x_i, x_j \rangle E_{ij} + \sum_{i=1}^{m} \alpha_i^2 E_{ii} \tag{41}$$

$$\leq \sup_{\|\alpha\|=1} \sqrt{\sum_{i\neq j} \alpha_i^2 \alpha_j^2} \sqrt{\sum_{i\neq j} \langle x_i, x_j \rangle^2 E_{ij}^2} + \max_i |E_{ii}| . \tag{42}$$

Let $\mathcal{G}$ denote the event that for all $i \neq j \in [m]$, $|\langle x_i, x_j \rangle| \leq O\left(\frac{\log d}{\sqrt{d}}\right)$, then by Lemma 11 and the union bound

$$\Pr[\neg\mathcal{G}] \leq 2m^2 e^{-\log^2 d}. \tag{43}$$

Therefore, with probability at least $1 - 2m^2 e^{-\log^2 d}$, we have

$$\|G_n - G\| \leq c\frac{\log d}{\sqrt{d}} \sqrt{\sum_{i\neq j} E_{ij}^2} + \max_i |E_{ii}| . \tag{44}$$

Note that

$$U(\{x_1, \cdots, x_m\}) = \frac{1}{m(m-1)} \sum_{i\neq j} E_{ij}^2 \tag{45}$$

is a U-statistics.

Suppose $|E_{ij}| \leq B$, according to the concentration inequality (Theorem 2 in [Peel et al., 2010]), we have with probability at least $1 - \delta$

$$\sum_{i\neq j} E_{ij}^2 \leq m(m-1)\mathbb{E}_{\{x_1,x_2\}} E_{12}^2 + m(m-1)\left(\sqrt{\frac{4\Sigma^2}{m}\log\frac{1}{\delta}} + \frac{4B^2}{3m}\log\frac{1}{\delta}\right), \tag{46}$$

where $\Sigma^2 = \mathbb{E}\left[E_{1,2}^4\right] - \mathbb{E}\left[E_{1,2}^2\right]^2$ is the variance for the kernel in U-statistics.

Putting everything together, we have with probability at least $1 - \delta - 2m^2 e^{-\log^2 d}$

$$\|G_n - G\| \leq c\frac{\log d}{\sqrt{d}}\left(m\sqrt{\mathbb{E}_{\{x_1,x_2\}} E_{12}^2} + m\left(\frac{4\Sigma^2}{m}\log\frac{1}{\delta}\right)^{1/4} + mB\sqrt{\frac{4}{3m}\log\frac{1}{\delta}}\right) + B \tag{47}$$

## D Discrepancy of the weights

In this section, we relate the quantities $\mathbb{E}_{\{x_1,x_2\}} E_{12}^2$ and $B$ to the discrepancies of the weights. Note that for ReLU, $\sigma'(w^\top x)$ does not depend on the norm of $w$, so we can focus on $w$ on the unit sphere.

Given a set of $n$ points $W = \{w_i\}_{i=1}^n$ on the unit sphere $\mathbb{S}^{d-1}$, the discrepancy of $W$ with respect to a measurable subset $S \subseteq \mathbb{S}^{d-1}$ is defined as

$$\text{dsp}(W, S) = \frac{1}{n}|W \cap S| - A(S) \tag{48}$$

where $A(S)$ is the normalized area of $S$ (*e.g.*, the area of the whole sphere $A(\mathbb{S}^{d-1})$ is 1). Let $\mathcal{S}$ denote the family of slices in $\mathbb{S}^{d-1}$

$$\mathcal{S} = \left\{S_{xy} : x, y \in \mathbb{S}^{d-1}\right\}, \text{ where } S_{xy} = \left\{w \in \mathbb{S}^{d-1} : w^\top x \geq 0, w^\top y \geq 0\right\}. \tag{49}$$

The $L_\infty$ discrepancy of $W$ with respect to $\mathcal{S}$ is

$$L_\infty(W, \mathcal{S}) = \sup_{S\in\mathcal{S}} \text{dsp}(W, S), \tag{50}$$

and the $L_2$ discrepancy is

$$L_2(W, \mathcal{S}) = \sqrt{\mathbb{E}_{x,y}\text{dsp}(W, S_{xy})^2} \tag{51}$$

where the expectation is taken over $x, y$ uniformly on the sphere. We use $L_\infty(W)$ and $L_2(W)$ as their shorthands.

For ReLU, by definition, we have

$$\mathbb{E}E_{ij}^2 = \left(L_2(W)\right)^2, \tag{52}$$

$$B \leq L_\infty(W), \tag{53}$$

$$\Sigma^2 \leq \mathbb{E}\left[E_{1,2}^4\right] \leq \mathbb{E}\left[E_{1,2}^2\right] \max_{x_1,x_2} \left|E_{1,2}^2\right| \leq \left(L_\infty(W)L_2(W)\right)^2, \tag{54}$$

using the fact that $E_{ij} = \text{dsp}(W, S_{x_i x_j})$.

Therefore, the bound becomes

$$\|G_n - G\| \leq c\frac{\log d}{\sqrt{d}}\left(mL_2(W) + \sqrt{L_\infty(W)L_2(W)}m\left(\frac{4}{m}\log\frac{1}{\delta}\right)^{1/4} + mL_\infty(W)\sqrt{\frac{4}{3m}\log\frac{1}{\delta}}\right) + L_\infty(W) \tag{55}$$

In the following subsections, we will discuss the discrepancies.

### D.1 Computing $L_2$ discrepancy for ReLU

Note that the derivative of ReLU $\sigma'(w^\top x) = \mathbb{I}\left[w^\top x\right]$ does not depend on the norm of $w$. Without loss of generality, we can assume $\|w\| = 1$ throughout this subsection.

**Theorem 8** *Suppose* $W = \{w_i\}_{i=1}^n \subseteq \mathbb{S}^{d-1}$.

$$\left(L_2(W)\right)^2 = \frac{1}{n^2}\sum_{i,j=1}^n k(w_i, w_j)^2 - \mathbb{E}_{u,v}\left[k(u,v)^2\right]$$

*where* $\mathbb{E}_{u,v}$ *is over* $u$ *and* $v$ *uniformly distributed on* $\mathbb{S}^{d-1}$ *and the kernel* $k(\cdot,\cdot)$ *is*

$$k(u,v) = \frac{\pi - \arccos\langle u, v\rangle}{2\pi}.$$

**Proof** Let $d(u,v) = \frac{\arccos\langle u,v\rangle}{\pi}$. Let $S_{xy} = \left\{w \in \mathbb{S}^{d-1} : w^\top x \geq 0,\ w^\top y \geq 0\right\}$. It is clear that (up to sets of measure zero)

$$A(S_{xy}) = k(x,y) = \frac{1 - d(x,y)}{2}, \tag{56}$$

$$\mathbb{I}\left[z \in S_{xy}\right] = \frac{1}{4}\left(\text{sign}(x^\top z) + 1\right)\left(\text{sign}(y^\top z) + 1\right), \tag{57}$$

where $\mathbb{I}\left[\cdot\right]$ is the indicator function. Then

$$\text{dsp}(W, S_{xy}) = \frac{1}{n}\sum_{k=1}^n \mathbb{I}\left[w_k \in S_{xy}\right] - A(S_{xy}) \tag{58}$$

$$= \frac{1}{n}\sum_{k=1}^n \frac{1}{4}\left(\text{sign}(x^\top w_k) + 1\right)\left(\text{sign}(y^\top w_k) + 1\right) - \frac{1 - d(x,y)}{2}. \tag{59}$$

Let $s_{xi}$ be a shorthand for $\text{sign}(x^\top w_i)$. Then we have

$$\left(L_2(W)\right)^2 = \mathbb{E}_{x,y}\left(\text{dsp}(W, S_{xy})\right)^2 \tag{60}$$

$$= \int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\left(\frac{1}{n}\sum_{k=1}^n \frac{1}{4}\left(\text{sign}(x^\top w_k) + 1\right)\left(\text{sign}(y^\top w_k) + 1\right) - \frac{1 - d(x,y)}{2}\right)^2 dA(x)dA(y) \tag{61}$$

$$= \frac{1}{n^2}\sum_{i,j=1}^n \int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} \frac{(s_{xi}+1)(s_{yi}+1)}{4}\frac{(s_{xi}+1)(s_{yi}+1)}{4}dA(x)dA(y) \tag{62}$$

$$- \frac{2}{n}\sum_{i=1}^n \int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} \frac{1 - d(x,y)}{2}\frac{(s_{xi}+1)(s_{yi}+1)}{4}dA(x)dA(y) \tag{63}$$

$$+ \int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\left(\frac{1 - d(x,y)}{2}\right)^2 dA(x)dA(y). \tag{64}$$

Consider the first term, which is equal to

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\left(\int_{\mathbb{S}^{d-1}}\frac{(s_{xi}+1)(s_{xj}+1)}{4}dA(x)\right)\left(\int_{\mathbb{S}^{d-1}}\frac{(s_{yi}+1)(s_{yj}+1)}{4}dA(y)\right). \tag{65}$$

By Lemma 13,

$$\int_{\mathbb{S}^{d-1}}\frac{(s_{xi}+1)(s_{xj}+1)}{4}dA(x)=\frac{2-2d(w_i,w_j)}{4}, \tag{66}$$

so the first term is equal to

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\left(\int_{\mathbb{S}^{d-1}}\frac{(s_{xi}+1)(s_{xj}+1)}{4}dA(x)\right)^2=\frac{1}{n^2}\sum_{i,j=1}^{n}k(w_i,w_j)^2. \tag{67}$$

Now consider the second term. Note that the summand is invariant to $w_i$, so it can be replaced by an arbitrary $p\in\mathbb{S}^{d-1}$. The second term is then equal to

$$-2\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\frac{1-d(x,y)}{2}\frac{(\mathrm{sign}(x^\top p)+1)(\mathrm{sign}(y^\top p)+1)}{4}dA(x)dA(y) \tag{68}$$

$$=-2\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\frac{1-d(x,y)}{2}\mathbb{I}[p\in S_{xy}]dA(x)dA(y) \tag{69}$$

$$=-2\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\frac{1-d(x,y)}{2}\mathbb{I}[p\in S_{xy}]dA(x)dA(y)dA(p) \tag{70}$$

$$=-2\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\frac{1-d(x,y)}{2}\left[\int_{\mathbb{S}^{d-1}}\mathbb{I}[p\in S_{xy}]dA(p)\right]dA(x)dA(y) \tag{71}$$

$$=-2\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\frac{1-d(x,y)}{2}\frac{2-2d(x,y)}{4}dA(x)dA(y) \tag{72}$$

$$=-2\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\left(\frac{1-d(x,y)}{2}\right)^2dA(x)dA(y), \tag{73}$$

where the third step is by invariance to $p$ and the fourth step is by Lemma 13. The theorem then follows. ■

Theorem 8 lets us compute $L_2(W)$ for a fixed $W$. The next lemma gives a concrete bound for a special case where $W$ is uniformly distributed on the unit sphere.

**Lemma 12** *There exists a constant $c_g$, such that for any $0<\delta<1$, with probability at least $1-\delta$ over $W=\{w_i\}_{i=1}^n$ that are sampled from the unit sphere uniformly at random,*

$$(L_2(W))^2\le c_g\left(\sqrt{\frac{\log d}{nd}\log\frac{1}{\delta}}+\frac{1}{n}\log\frac{1}{\delta}\right).$$

**Proof** By Theorem 8, we have

$$(L_2(W))^2=\frac{1}{4n^2}\sum_{i,j=1}^{n}\left(\frac{1}{2}-d(w_i,w_j)\right)^2-\frac{1}{4}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\left(\frac{1}{2}-d(u,v)\right)^2dA(u)dA(v)+\frac{1}{4n^2}\sum_{i,j=1}^{n}\left(\frac{1}{2}-d(w_i,w_j)\right). \tag{74}$$

First consider $T_1=\frac{1}{n^2}\sum_{i,j=1}^n\left(\frac{1}{2}-d(w_i,w_j)\right)^2-\mu$ where $\mu=\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}}\left(\frac{1}{2}-d(u,v)\right)^2dA(u)dA(v)$. Rewrite $T_1=\frac{1}{4n}+\frac{n-1}{n}U(W)-\mu$ where $U(W)=\frac{1}{n(n-1)}\sum_{i\neq j}\left(\frac{1}{2}-d(w_i,w_j)\right)^2$ is a U-statistics. We upper bound $U(W)$ by using Bernstein's inequality when $W$ is uniform over the sphere.

By Taylor expansion, we have

$$\frac{1}{2}-d(u,v)=x/\pi+x^3/6\pi+O(x^5),\ \text{where } x=u^\top v.$$

Then let $\mathcal{G}$ denote the event that $|x|=|u^\top v|\le c\sqrt{\log d/d}$ for a sufficient large constant $c>0$, so that by Lemma 11,

$\Pr[\neg\mathcal{G}] \leq O(1/d^4)$. Then

$$\mathbb{E}\left[U(W)\right] = \mu = \mathbb{E}\left[\left(x/\pi + x^3/6\pi + O(x^5)\right)^2\right] \tag{75}$$

$$= \mathbb{E}[x^2/\pi^2 + x^4/6\pi^2 + O(x^6)] \tag{76}$$

$$\leq \mathbb{E}\left[[x^2/\pi^2 + x^4/6\pi^2 + O(x^6)] \mid \mathcal{G}\right] + \Pr[\neg\mathcal{G}] \max_{u,v} \left[\frac{1}{2} - d(u,v)\right]^2 \tag{77}$$

$$= O\left(\frac{\log d}{d}\right), \tag{78}$$

and thus

$$\mathrm{Var}\left[U(W)\right] = \mathbb{E}\left\{\left[\left(x/\pi + x^3/6\pi + O(x^5)\right)^2 - \mu\right]^2\right\} \tag{79}$$

$$= \mathbb{E}\left\{\left[x^2/\pi^2 + x^4/6\pi^2 + O(x^6) - \mu\right]^2\right\} \tag{80}$$

$$\leq \mathbb{E}\left\{\left[x^2/\pi^2 + x^4/6\pi^2 + O(x^6) - \mu\right]^2 \mid \mathcal{G}\right\} + \Pr[\neg\mathcal{G}] \max_{u,v} \left[\left(\frac{1}{2} - d(u,v)\right)^2 - u\right]^2 \tag{81}$$

$$= O\left(\frac{\log^2 d}{d^2}\right). \tag{82}$$

Then by Berstein's inequality, we have with probability at least $1-\delta$ over the $W$ uniformly on the sphere,

$$|T_1| \leq O\left(\frac{\log d}{d}\sqrt{\frac{1}{n}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{1}{\delta}\right). \tag{83}$$

A similar argument holds for $T_2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\left(\frac{1}{2} - d(w_i, w_j)\right)$. Note that

$$\mathrm{Var}\left\{\left(\frac{1}{2} - d(u,v)\right)\right\} = \mu = O\left(\frac{\log d}{d}\right). \tag{84}$$

We have that with probability at least $1-\delta$ over the $W$ uniform from the sphere,

$$|T_2| \leq O\left(\sqrt{\frac{\log d}{nd}\log\frac{1}{\delta}} + \frac{1}{n}\log\frac{1}{\delta}\right). \tag{85}$$

This completes the proof. ■

Below are some technical lemmas that are used in the analysis.

**Lemma 13**

$$\int_{\mathbb{S}^{d-1}} d(x,y)dA(x) = \frac{1}{2}, \forall y \in \mathbb{S}^{d-1}, \tag{86}$$

$$\int_{\mathbb{S}^{d-1}} \mathrm{sign}(x^\top y)dA(x) = 0, \forall y \in \mathbb{S}^{d-1}, \tag{87}$$

$$\int_{\mathbb{S}^{d-1}} \mathrm{sign}(x^\top z)\,\mathrm{sign}(y^\top z)dA(z) = 1 - 2d(x,y), \forall x, y \in \mathbb{S}^{d-1}. \tag{88}$$

**Proof** The first two are straightforward. The third is implicit in the proof of Theorem 1.21 in [Bilyk and Lacey, 2015]. ■

## E Rademacher complexity and final error bounds: Proof of Theorem 3 and Corollary 4

We apply the argument in [Bartlett and Mendelson, 2002] to our setting to get Lemma 14. Combining it with Theorem 1 leads to Theorem 3. Furthering combining with Lemma 9 leads to Corollary 4 follows from

**Lemma 14** *Suppose the data are bounded: $|y| \le Y$ and $\|x\|_2 \le 1$. Let*

$$\mathcal{F} = \left\{ f(x) = \sum_{k=1}^{n} v_k \sigma(w_k^\top x) : v_k \in \{-1, +1\}, \ \sum_k \|w_k\|_2 \le C_W \right\}.$$

*Then with probability $\ge 1-\delta$, for any $f \in \mathcal{F}$,*

$$\frac{1}{2}\mathbb{E}(y - f(x))^2 \le \frac{1}{2m}\sum_{l=1}^{m}(y_l - f(x_l))^2 + \frac{2(Y + C_W)C_W}{\sqrt{m}} + (Y^2 + C_W^2)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{89}$$

**Proof** For a sample $S = ((x_1, y_1), \cdots, (x_m, y_m))$, and a loss function $l(y, x) = \frac{1}{2}(y - f(x))^2$, we denote $\hat{\mathbb{E}}_S[l]$ as the empirical average $\hat{\mathbb{E}}_S[l] = \frac{1}{m}\sum_{l=1}^{m} l(y_l, x_l)$.

Define

$$\Phi(S) = \sup_{l \in \mathcal{L}} \mathbb{E}[l] - \hat{\mathbb{E}}_S[l] \tag{90}$$

where $\mathcal{L}$ is the set of loss functions

$$\mathcal{L} = \left\{ l(y, x) = \frac{1}{2}(y - f(x))^2 : f \in \mathcal{F} \right\}.$$

Let $S$ and $S'$ be two datasets that differ by exactly one data point $(x_i, y_i)$ and $(x_i', y_i')$. Then we have a bound on the difference of loss functions. Since $\|x\|_2 \le 1$ and $\sum_k \|w_k\|_2 \le C_W$, we have $|f| \le C_W$. Thus

$$|l(y, f(x)) - l(y', f(x'))| \le \frac{1}{2}\max\left\{(y - f(x))^2, (y' - f(x'))^2\right\} \le Y^2 + C_W^2. \tag{91}$$

This leads to an upper bound

$$\begin{aligned}\Phi(S) - \Phi(S') &\le \sup_{l \in \mathcal{L}} \hat{\mathbb{E}}_S[l] - \hat{\mathbb{E}}_{S'}[l] \\ &= \sup_{l \in \mathcal{L}} \frac{l(y_i, f(x_i)) - l(y_i', f(x_i'))}{m} \le \frac{Y^2 + C_W^2}{m}.\end{aligned} \tag{92}$$

Similarly, we can get the other side of the inequality and have $|\Phi(S) - \Phi(S')| \le \frac{Y^2 + C_W^2}{m}$.

From McDiamids' inequality, with probability at least $1 - \delta$ we get

$$\Phi(S) \le \mathbb{E}_S \Phi(S) + (Y^2 + C_W^2)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \tag{93}$$

The first term on the right-hand side can be bounded by Rademacher complexity as shown in the book Foundations of Machine Learning (3.13). In the end, we have the bound

$$\frac{1}{2}\mathbb{E}(y - f(x))^2 \le \frac{1}{2m}\sum_{l=1}^{m}(y_l - f(x_l))^2 + 2\mathcal{R}_m(\mathcal{L}) + (Y^2 + C_W^2)\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{94}$$

where $\mathcal{R}_m(\mathcal{L})$ is the Rademacher complexity of the function class $\mathcal{L}$.

We can find the Rademacher complexity by using composition rules. The Rademacher complexity of linear functions $\left\{w^\top x : \|w\|_2 \le b_W, \|x\|_2 \le 1\right\}$ is $b_W/\sqrt{m}$, where $m$ is the number of data points. If a function $\phi$ is $L$-Lipschitz, then for any function class $\mathcal{H}$, we have $\mathcal{R}(\phi \circ \mathcal{H}) \le L\mathcal{R}(\mathcal{H})$. In addition, we also have $\mathcal{R}(cH) = |c|\,\mathcal{R}(H)$ and $\mathcal{R}(\sum_k F_k) \le \sum_k \mathcal{R}(F_k)$.

So for the function class $\mathcal{F}$ that describes a neural network, we have

$$\mathcal{R}_m(\mathcal{F}) \le \frac{C_W}{\sqrt{m}}. \tag{95}$$

It is derived by using the fact that $\sigma'(\cdot)$ is 1-Lipschitz and $\sum_k \|w_k\|_2 \le C_W$.

Finally composing on the loss function we get

$$\mathcal{R}_m(\mathcal{L}) \leq \frac{(Y + C_W)C_W}{\sqrt{m}}, \tag{96}$$

using the fact that the ground truth in the loss should be bounded by $Y$ and the function bounded by $C_W$, thus the Lipschitz constant of the loss function is bounded by $Y + C_W$. ■

# F Discussions

In this section, we discuss and remark on further considerations and possible extensions of our current analysis.

## F.1 Other loss functions

Currently, our analysis is tied to the least squares loss $\ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$. It is fairly straightforward to generalize it to any strongly convex loss function, such as logistic loss. Note that the final objective function is *not* convex due to the non-convexity in neural networks, but most loss functions used in practice are strongly convex w.r.t. $f(x)$. Under the new setting, the residual is then

$$r = \frac{1}{m}\left(\ell'(y_1, f(x_1)), \cdots, \ell'(y_m, f(x_m))\right)^\top.$$

According to our analysis, the norm of the residual $\|r\|$ will be bounded. This in turn implies each individual $\ell'(y_l, f(x_l))$ will be small. Since the loss function $\ell(y, f(x))$ is strongly convex, the loss itself will be small.

## F.2 Other activation functions

We can consider a family of activation functions of the form $\sigma(u) = \max\{u, 0\}^t$, *i.e.*, rectified polynomials [Cho and Saul, 2009, Krotov and Hopfield, 2016]. This requires two modifications to the analysis.

One is the corresponding kernel $k(x, y) = \mathbb{E}_w\left[\sigma'(w^\top x)\sigma'(w^\top y)\right]$ and $g(x, y) = k(x, y)\langle x, y\rangle$. When the input distribution is uniform, we can also compute the kernels in closed form as shown in [Cho and Saul, 2009]:

$$k_t(x, y) = \frac{J_{t-1}(\theta)}{2\pi} \tag{97}$$

where

$$J_t(\theta) = (-1)^t(\sin\theta)^{2t+1}\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\right)^t\left(\frac{\pi - \theta}{\sin\theta}\right), \tag{98}$$

and $\theta = \arccos\langle x, y\rangle$. Note that the subscript is $t - 1$ in (97) because we are computing on the derivative $\sigma'(u)$.

Examples for the first few $t$ are listed as follows.

$$J_0(\theta) = \pi - \theta \tag{99}$$

$$J_1(\theta) = \sin\theta + (\pi - \theta)\cos\theta \tag{100}$$

$$J_2(\theta) = 3\sin\theta\cos\theta + (\pi - \theta)(1 + 2\cos^2\theta) \tag{101}$$

Larger $t$ corresponds to more nonlinear activation functions and leads to slower decaying spectrum since there are more high frequency components.

We also need to change the definition of the discrepancy to accommodate the new kernels. Let

$$\begin{aligned}(L_2(W))^2 &= \mathbb{E}_{x_i, x_j}\left[\mathbb{E}_w[\sigma'(w^\top x_i)\sigma'(w^\top x_j)] - \frac{1}{n}\sum_{k=1}^n \sigma'(w_k^\top x_i)\sigma'(w_k^\top x_j)\right]^2 \\ &= \frac{1}{n^2}\sum_{i,j=1}^n k(w_i, w_j)^2 - \mathbb{E}_{u,v}\left[k(u, v)^2\right].\end{aligned} \tag{102}$$

Therefore, the discrepancy is affected by how the kernels change due to change in activation functions.

The other modification is on the Rademacher complexity. Since the derivative $\sigma'(u) = t\max\{u, 0\}^{t-1}$, there is an additional factor of $t$ in front of the complexity. That is, larger $t$ leads to higher Rademacher complexity.

Table 2: Comparison of minimum eigenvalues with uniform and "matching" distributions. Note that the "matching" distribution corresponds to larger minimum eigenvalue for different dimensions. However, the difference becomes negligible when the dimension increases.

| $d$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| uniform | $3.96 \times 10^{-4}$ | 0.0015 | 0.0032 | 0.0072 |
| matching | $\mathbf{5.43 \times 10^{-4}}$ | **0.0017** | 0.0032 | 0.0072 |

In summary, the best parameter $t$ depends on the balance between the two conflicting effects. On one hand, larger $t$ corresponds to slower decaying spectrum and makes the minimum singular value more likely to be larger. On the other hand, smaller $t$ leads to better generalization since the Rademacher complexity is smaller.

### F.3 (Sub)gradient of the activation function

Throughout this paper, we have used one particular subgradient for the ReLU activation function: $\mathbb{I}[u > 0]$. At the point $u = 0$, there are many other valid subgradients as long as its value is between 0 and 1. However, our analysis is not restricted to this particular subgradient. First of all, all the subgradients only differ at one point $u = 0$, which is of probability zero. Second, our analysis is probabilistic in nature. The first term in Lemma 5 is the expectation over $W$, which is insensitive to the probability zero event $u = 0$. The second term in Lemma 5 is related to $L_2(W)$, which is again expectation over all possible data, thus insensitive to the difference.

In summary, though for some $W \in \mathcal{G}_W$ the loss is not differentiable, one can define $\partial L/\partial W$ by using subgradients of ReLU $\sigma$ as follows:

$$\sigma'(x) = \begin{cases} 0, & x < 0 \\ c, & x = 0 \\ 1, & x > 0 \end{cases} \tag{103}$$

for any $c \in [0, 1]$. Then under the conditions in our theorems, with high probability, for any $W \in \mathcal{G}_W$ and any definition of $\sigma'$ in (103), the guarantees hold.

Other activation functions such as rectified polynomials are differentiable and thus they do not have such issue.

### F.4 Other input distribution

When the input distribution is not uniform, the spectrum of the kernel function defined in (**??**) will be different because the spherical harmonic bases are defined with respect to the input distribution. To ensure the spectrum decays slowly, we need to find a corresponding distribution of $W$ that "matches" the input distribution.

We provide some intuitions in finding such "matching" distribution. Suppose the input distribution is uniform on the set $E \in \mathbb{S}^{d-1}$, if a hyperplane whose normal is $w$ does not "cut through" the set, then for all data points, they have the same sign $\mathbb{I}[w^\top x > 0]$. This will likely lead to rank deficiency in the extended feature matrix.

Therefore, we prefer $W$ to split the data points as much as possible. One such distribution of $W$ is uniform on the set $F_E = \left\{w \in \mathbb{S}^{d-1} : \text{there exists } \ u \in E, \ \langle u, w\rangle = 0\right\}$. For example, if $E$ is the intersection of the positive orthant and the unit sphere, $E = \left\{u \in \mathbb{S}^{d-1} : u_i \geq 0, \text{ for all } \ i \in [d]\right\}$, then the corresponding set $F_E$ is the whole sphere excluding $E$ and $-E$.

We have verified the phenomenon empirically. We have generated 3000 input data points uniform on the positive orthant $E$. We then compute the $3000 \times 3000$ Gram matrix, where the $(i, j)$-th entry is $\mathbb{E}_w\left[\sigma'(w^\top x_i)\sigma'(w^\top x_i)\langle x_i, x_j\rangle\right]$. The expectation is approximated by sampling 100,000 independent $w$'s and then averaging. We compare two distributions of $W$: 1) uniform on the whole unit sphere; 2) uniform on $F_E$.

In Table 2, we compare the minimum eigenvalues with the two distributions. The uniform distribution on $F_E$ always leads to larger or the same minimum eigenvalues. However, as dimension increases, the difference becomes negligible. Note that the difference between the uniform distribution on the whole sphere and uniform on $F_E$ becomes exponentially small when the dimension $d$ increases, because the proportion of $E$ and $-E$ shrinks exponentially. This suggests that in high dimensions, uniform on the whole unit sphere is a reasonable distribution for $W$.

Table 3: Performance comparison with/without regularization on MNIST dataset. Errors are all in $\%$.

| | $n = 200$ | | $n = 400$ | |
|---|---|---|---|---|
| | train | test | train | test |
| no-reg | 0.94 | 3.39 | **0.32** | 3.08 |
| reg | **0.56** | **3.22** | 0.33 | **2.90** |
| | $n = 600$ | | $n = 800$ | |
| | train | test | train | test |
| no-reg | 0.00065 | 2.67 | 0.11 | 2.90 |
| reg | **0.00057** | **2.62** | **0.0003** | **2.45** |

For a general input distribution $P(x)$, we can decompose it into small sets $dx$ and on every set, the distribution is uniform with measure $P(x)dx$. Then every small sets corresponds to a distribution of $W$. The final distribution of $W$ is the superposition of all such distributions, weighted by $P(x)dx$.

## G Further experiment on MNIST

We also compare the regularization effects on the MNIST dataset. The dataset contains 60,000 training and 10,000 test handwritten digits. To demonstrate the regularization effect, we train one hidden layer fully connected neural networks with $k = 200, 400, 600, 800$ units. The results are summarized in Table 3. Note that state-of-the-arts performance on MNIST are mostly obtained by convolutional neural networks. This experiment is not intended to achieve the state-of-the-arts but it tries to showcase the advantage of regularization on a real-world dataset.

From Table 3, we see regularization consistently leads to slightly better test error for all cases.