Table 3: Summary of Notation

Notation	Definition
$\mathbb{I}$	Input domain, subset of $\mathbb{R}^d$
$\bar{\mathbb{I}}$	$\bigcup_{N=1}^{\infty} \mathbb{I}^N$
$\mathcal{A}_i$	Classes of test functions $\mathbb{I} \to \mathbb{R}$
γ_i	Test function norm
$\mathcal{S}_i$	Class of functions mapping $\mathcal{P}(\mathbb{I}) \to \mathbb{R}$
$\|\cdot\|_{\mathcal{S}_i}$	Measure network norm
D	Map from vectors to empirical measures s.t. $D(x_1 \dots x_n) = \sum_{i=1}^n \delta_{x_i}$
$\hat{\mathcal{P}}(\mathbb{I})$	$\bigcup_{N=1}^{\infty} D(\mathbb{I}^N)$
κ	Fixed probability measure over $\mathbb{S}^d$ in first layer
ν	Signed measure over $\mathbb{S}^d$ in first layer
τ	Fixed probability measure over $\mathcal{A}_i$ in second layer
χ	Signed measure over $\mathcal{A}_i$ in second layer
$Y_{k,j}$	Orthogonal basis polynomial of degree k and index j on $\mathbb{S}^d$
P_k	Legendre polynomial of degree k
g_k	the kth spherical harmonic of a function $g : \mathbb{S}^d \to \mathbb{R}$

A. Omitted Proofs

A.1. Proof of Proposition 3.1

Proof:

We remind our notation. Given $f : \mathbb{I} \to \mathbb{R}$, the empirical extension $\hat{f} : \hat{\mathcal{P}}(\mathbb{I}) \to \mathbb{R}$ is defined as $\hat{f}(\mu) := f(x_\mu)$ where $x_\mu \in D^{-1}(\mu)$ and $\|x_\mu\|_0 = \min_{x \in D^{-1}(\mu)} \|x\|_0$. And for $\bar{f} : \mathcal{P}(\bar{\mathbb{I}}) \to \mathbb{R}$, we say this is a continuous extension of f if $\bar{f}$ is continuous in under the Wasserstein metric, and $f(x) = \bar{f}(D(x))$ for every real, finite-dimensional vector x.

For the forward implication, if $\bar{f}$ is a continuous extension, then clearly $\bar{f} = \hat{f}$ restricted to $\hat{\mathcal{P}}(\mathbb{I})$.

Furthermore, continuity of $\bar{f}$ and compactness of $\mathcal{P}(\mathbb{I})$ implies $\bar{f}$ is uniformly continuous, and therefore $\hat{f}$ is as well.

For the backward implication, we introduce $\hat{f}_\epsilon(\mu) = \sup_{\nu \in B_\epsilon(\mu) \cap \hat{\mathcal{P}}(\mathbb{I})} \hat{f}(\nu)$ where the ball $B_\epsilon(\mu)$ is defined with the Wasserstein metric. Note that $\hat{f}_\epsilon$ is defined over arbitrary probability measures, not just discrete measures. Now, we introduce $\bar{f}(\mu) = \inf_{\epsilon > 0} \hat{f}_\epsilon(\mu)$, where density of the discrete measures and uniform continuity of $\hat{f}$ guarantees that $\bar{f}$ is well-defined and finite.

Uniform continuity implies if $\mu \in \hat{\mathcal{P}}(\mathbb{I})$ then $\bar{f}(\mu) = \hat{f}(\mu)$. Consider any $y \in \mathbb{I}^M$ such that $\mu = D(y)$, and define a sequence of vectors $y^i = (z_i, y_2, \dots, y_M)$ where $z_i \to y_1$ and all z_i are distinct from elements of y. Every point $y^i \in \mathbb{I}^M$ has a unique coordinate and therefore $\hat{f}(D(y^i)) = f_M(y^i)$. Because $D(y^i) \rightharpoonup D(y)$, continuity implies $\hat{f}(D(y)) = f_M(y)$. Thus, for any $y \in \mathbb{I}^M$, $\bar{f}(D(y)) = f_M(y)$, which implies $\bar{f}$ is an extension.

Now, suppose we have an arbitrary convergent sequence of probability measures $\mu_n \rightharpoonup \mu$. By the density of discrete measures, we can define sequences $\mu_n^m \rightharpoonup \mu_n$ where $\mu_n^m \in \hat{\mathcal{P}}(\mathbb{I})$. In particular, we may choose these sequences such that for all n, $W_1(\mu_n^m, \mu_n) \leq \frac{1}{m}$. Then for any $\epsilon > 0$,

$$|\bar{f}(\mu) - \bar{f}(\mu_n)| \leq |\bar{f}(\mu) - \hat{f}_\epsilon(\mu)| + |\hat{f}_\epsilon(\mu) - \hat{f}(\mu_n^n)| + |\hat{f}(\mu_n^n) - \hat{f}_\epsilon(\mu_n)| + |\hat{f}_\epsilon(\mu_n) - \bar{f}(\mu_n)| \,.$$

Consider the simultaneous limit as $n \to \infty$ and $\epsilon \to 0$. On the RHS, the first term vanishes by definition, and the fourth by uniform continuity. For any $\nu \in B_\epsilon(\mu) \cap \hat{\mathcal{P}}(\mathbb{I})$, $W_1(\nu, \mu_n^n) \leq W_1(\nu, \mu) + W_1(\mu, \mu_n) + W_1(\mu_n, \mu_n^n) \to 0$ in the limit. So the second term vanishes as well by uniform continuity of $\hat{f}$. Similarly, for any $\nu \in B_\epsilon(\mu_n) \cap \hat{\mathcal{P}}(\mathbb{I})$, $W_1(\nu, \mu_n^n) \leq W_1(\nu, \mu_n) + W_1(\mu_n, \mu_n^n) \to 0$, and the third term vanishes by uniform continuity. This proves continuity of $\bar{f}$.

□

A.2. Proof of Proposition 5.1

Proof: We can decompose the generalization error:

$$\begin{aligned}
&\mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \left| \mathbb{E}_{\mu \sim \mathcal{D}} \ell(f^*(\mu), f(\mu)) - \frac{1}{n} \sum_{i=1}^n \ell(f^*(\mu_i), f(\mu_i)) \right| \\
&\leq 2\mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell(f^*(\mu_i), f(\mu_i)) \right| \\
&\leq 2\mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell(f^*(\mu_i), 0) \right| + 2\mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (\ell(f^*(\mu_i), 0) - \ell(f^*(\mu_i), f(\mu_i))) \right| \\
&\leq \frac{2RG\delta}{\sqrt{n}} + 4R^2 G \mathbb{E} \sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\mu_i) \right| ,
\end{aligned}$$

where the second step uses symmetrization through the Rademacher random variable ϵ, and the fourth is by assumption on the loss function ℓ, from the fact that $\|f\|_{\mathcal{S}_1} \leq \delta$ implies $\|f\|_\infty \leq 2R^2\delta$. We decompose the Rademacher complexity (removing the absolute value by symmetry):

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\mu_i) \right] &= \mathbb{E}\left[\sup_{\substack{\chi \in \mathcal{M}(\mathcal{A}) \\ \|\chi\|_{TV} \leq \delta}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \int \sigma(\langle \phi, \mu_i \rangle) \chi(d\phi) \right] \\
&= \delta \mathbb{E}\left[\sup_{\gamma_1(\phi) \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \sigma(\langle \phi, \mu_i \rangle) \right] \\
&\leq \delta \mathbb{E}\left[\sup_{\gamma_1(\phi) \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle \phi, \mu_i \rangle \right] ,
\end{aligned}$$

where the last step uses the contraction lemma and that σ is 1-Lipschitz.

Now, using the neural network representation of ϕ:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\mu_i) \right] &\leq \delta \mathbb{E}\left[\sup_{\|\nu\|_{TV} \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \int_{\mathbb{R}^d} \int_{\mathbb{S}^d} \sigma(\langle w, \tilde{x}_i \rangle)^2 \nu(dw) \mu_i(dx_i) \right] \\
&\leq \delta \mathbb{E}\left[\sup_{\|w\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \mathbb{E}_{\mu_i}[\sigma(\langle w, \tilde{x}_i \rangle)^2] \right] \\
&\leq \delta \mathbb{E}_{\mu_1, \ldots, \mu_n}\left[\mathbb{E}\left[\sup_{\|w\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^n \epsilon_i \sigma(\langle w, \tilde{x}_i \rangle)^2 \middle| x_1, \ldots, x_n \right] \right] ,
\end{aligned}$$

where the last step uses Jensen's inequality and Fubini's theorem. The conditional expectation is itself a Rademacher complexity, so we may apply the contraction lemma again as the $\sigma(\langle w, \tilde{x}_i \rangle)^2$ activation is $2\sqrt{2}R$-Lipschitz for the domain $\mathbb{I}$ of $\tilde{x}_i$. Using the variational definition of the l_2 norm we have the bound:

$$\mathbb{E}\left[\sup_{\|f\|_{\mathcal{S}_1} \leq \delta} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\mu_i) \right] \leq \frac{4R^2\delta}{\sqrt{n}} .$$

The high probability bound then follows from McDiarmid's inequality.

□

A.3. Proof of Proposition 5.2

Proof:

We appeal to the following concentration inequality for empirical measures under the Wasserstein metric:

Theorem A.1 (Theorem 1 in (Fournier & Guillin, 2015)) *Let $\hat{\mu}_N = \frac{1}{N}\sum_{j=1}^{N}\delta_{X_j}$ where $X_i \sim \mu \in \mathcal{P}(\mathbb{I})$ iid. Then $\mathbb{E}[W_1(\hat{\mu}_N, \mu)] \lesssim N^{-1/d}$ where $d > 2$ is the dimension of $\mathbb{I}$.*

It's easy to see that any $\phi \in \mathcal{A}_2$ has Lipschitz constant bounded above by $2\sqrt{2}R$, and therefore $\sup_{\phi\in\mathcal{A}_2}|\langle\phi, \mu - \mu^*\rangle| \leq 2\sqrt{2}RW_1(\mu, \mu^*)$. Therefore

$$\begin{aligned}
\mathbb{E}\left[\sup_{\phi\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\phi,\mu_i\rangle\right] &\leq \mathbb{E}\left[\sup_{\phi\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\phi,\mu^*\rangle\right] + \mathbb{E}\left[\sup_{\phi\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\phi,(\mu^*-\mu_i)\rangle\right] \\
&\leq 2R^2\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right|\right] + 2\sqrt{2}R\mathbb{E}[W_1(\mu_i,\mu^*)] \\
&\lesssim R(n^{-1/2} + R\mathbb{E}_{N\sim\Omega}[N^{-1/d}])\,.
\end{aligned}$$

The conclusion then follows from the same Rademacher decomposition as in Proposition 5.1.

□

A.4. Proof of Theorem 4.1

For simplicity, we consider spherical inputs rather than Euclidean inputs, so we consider $k(x,y) = \int_{\mathbb{S}^d}\sigma(\langle w,x\rangle)\sigma(\langle w,y\rangle)\kappa(dw)$ without the $\tilde{x}$ bias terms, and assume $x \in \mathbb{S}^d$. Note that the Euclidean inputs may be seen as a restriction of the spherical inputs to an appropriate spherical cap, see (Bach, 2017a) for details of this construction.

A.4.1. Spherical Harmonics and Kernel Norm Background

We'll use $\simeq$ to denote equality up to universal constants. To understand functions in $\mathcal{A}_2$, we require the following details of spherical harmonics (Efthimiou & Frye, 2014).

A basis on $\mathbb{S}^d$ is given by the orthogonal polynomials $Y_{k,j}$, where $k \geq 0$ and $1 \leq j \leq N(d,k)$ where

$$\begin{aligned}
N(d,k) &\simeq \frac{k+d}{k}\frac{\Gamma(k+d-1)}{\Gamma(d)\Gamma(k)} \\
&\simeq \frac{k+d}{k}\frac{(k+d)^{k+d-3/2}}{d^{d-1/2}k^{k-1/2}}
\end{aligned}$$

The Legendre polynomials $P_k(t)$ act on one dimensional real inputs and satisfy the addition formula

$$\sum_{j=1}^{N(d,k)} Y_{k,j}(x)Y_{k,j}(y) = N(d,k)P_k(\langle x,y\rangle)$$

Finally, given a function $g : \mathbb{S}^d \to \mathbb{R}$, the kth spherical harmonic of g is the degree k component of g in the orthogonal basis, equivalently written as

$$g_k(x) = \int_{\mathbb{S}^d} g(y)P_k(\langle x,y\rangle)\kappa(dy)$$

We also require several calculations on functions with bounded functional norm and projections (Bach, 2017a), where we remind that we're using the activation $\sigma(x)^2$. For $g \in \mathcal{A}_2$ or $g(x) = \sigma(\langle w, x\rangle)^2$ for any $w \in \mathbb{S}^d$, we have that $g_{2k} = 0$ for all $k \geq 2$.

For $g \in \mathcal{A}_2$, the norm of each harmonic satisfies $\|g_k\|_2^2 = \lambda_k^2 N(d,k)$, and the kernel norm can be calculated explicitly as

$$\gamma_2(g)^2 = \sum_{k=0, \lambda_k \neq 0}^{\infty} \lambda_k^{-2} \|g_k\|_{L_2}^2$$

We have that $\lambda_1 \simeq d^{-1}$, $\lambda_k = 0$ for $k \geq 3$ and k even, and for $k \geq 3$ and k odd:

$$\lambda_k \simeq \pm \frac{d^{d/2+1/2} k^{k/2-3/2}}{(d+k)^{k/2+d/2+1}} \tag{6}$$

A.4.2. Separation of $\mathcal{S}_1$ and $\mathcal{S}_2$

Let $g(x) = \sigma(\langle x, w\rangle)^2$ for an arbitrary $w \in \mathbb{S}^d$, we have that $\|g_k\|_2^2 = \lambda_k^2 N(d,k)$. Define $\tilde{g} = g - \sum_{i=0}^{d^2-1} g_i$.

The following lemmas capture that $\tilde{g}$ has high correlation with g and exponentially small correlation with functions in $\mathcal{A}_2$.

Lemma A.2 *The correlation lower bound $\langle g, \tilde{g}\rangle \gtrsim d^{-21/2}$ holds.*

Proof:

Note that

$$\langle g, \tilde{g}\rangle = \sum_{k=d^2} \|g_k\|_2^2 = \sum_{k=d^2} \lambda_k^2 N(d,k) \tag{7}$$

We can calculate, because $k + d \leq 2k$:

$$\begin{aligned}
\lambda_k^2 N(d,k) &\simeq \frac{d^{d+1} k^{k-3}}{(d+k)^{k+d+2}} \cdot \frac{k+d}{k} \frac{(k+d)^{k+d-3/2}}{d^{d-1/2} k^{k-1/2}} \\
&\simeq d^{3/2} k^{-7/2} (k+d)^{-7/2} \\
&\gtrsim d^{3/2} k^{-7}
\end{aligned}$$

And therefore

$$\langle g, \tilde{g}\rangle \gtrsim \sum_{k=d^2}^{\infty} d^{3/2} k^{-7} \geq d^{3/2} \int_{d^2}^{\infty} k^{-7} dk \simeq d^{3/2} (d^2)^{-6}$$

which yields the desired lower bound.

□

Lemma A.3 *The value of the optimization problem*

$$\begin{aligned}
\max_{\phi} \quad & \langle \phi, \tilde{g}\rangle_{L_2} \\
\textit{s.t.} \quad & \gamma_2(\phi)^2 \leq \delta^2
\end{aligned}$$

is upper bounded by $\delta \cdot d^{1/2-d/3}$

Proof: By orthogonality we may assume $\phi_k = \alpha_k \tilde{g}_k = \alpha_k g_k$, where $\alpha_k = 0$ for $k < d^2$. Then the problem is equivalently

$$\begin{aligned} \min_{\alpha} \quad & -\sum_{k=d^2}^{\infty} \alpha_k \|g_k\|_2^2 \\ \text{s.t.} \quad & \sum_{k=d^2}^{\infty} \alpha_k^2 \lambda_k^{-2} \|g_k\|_2^2 \leq \delta^2 \end{aligned}$$

Taking λ as a Lagrangian multiplier yields the optimality condition $\alpha_k = (2\lambda)^{-1}\lambda_k^2$.

Plugging this into the constraint and introducing notation S yields

$$(2\lambda)^{-2} S := (2\lambda)^{-2} \sum_{k=d^2}^{\infty} \lambda_k^2 \|g_k\|_2^2 \leq \delta^2$$

Then the objective (returned to a maximum) obeys the bound

$$\begin{aligned} \sum_{k=d^2} (2\lambda)^{-1} \lambda_k^2 \|g_k\|_2^2 &= (2\lambda)^{-1} S \\ &\leq \delta \sqrt{S} \end{aligned}$$

So it remains to calculate S. Plugging in the value of $\|g_k\|_2^2$ gives

$$S = \sum_{k=d^2}^{\infty} \lambda_k^4 N(d,k)$$

We can give the form of each term, using that $k \geq d^2$:

$$\begin{aligned} \lambda_k^4 N(d,k) &\lesssim d^{3/2} k^{-7} \frac{d^{d+1} k^{k-3}}{(d+k)^{k+d+2}} \\ &\lesssim d^{3/2} k^{-7} \frac{d^{d+1} k^{k-3}}{k^{k+d+2}} \\ &\lesssim d^{5/2} k^{-12} \left(\frac{d}{k}\right)^d \\ &\lesssim d^{5/2} k^{-12} \left(\frac{d}{k^{1/2}} \cdot \frac{1}{k^{1/2}}\right)^d \\ &\lesssim d^{5/2} k^{-12} k^{-d/2} \end{aligned}$$

For sufficiently large d, we may ignore the lower terms and reduce the exponential term to $k^{-d/3}$, then:

$$S \lesssim \sum_{k=d^2}^{\infty} k^{-d/3} \simeq \int_{d^2}^{\infty} k^{-d/3} \simeq d^{-1} (d^2)^{1-d/3}$$

The bound follows.

□

Let $h = g - g_0 - g_2$, and define $f_1(\mu) = d^{-1}\sigma(\langle h, \mu\rangle)$, remembering that we're using the regular ReLU for the measure network activation.

Lemma A.4 $\|f_1\|_{\mathcal{S}_1} \lesssim 1$.

Proof: It suffices to bound $\gamma_1(h)$, remembering that our test functions are defined using networks with the squared ReLU activation. Clearly $\gamma_1(g) \leq 1$ as it itself a single neuron. For the other terms, we can write the harmonics explicitly, using the fact that $P_0(t) = 1$ and $P_2(t) = \frac{(d+1)t^2-1}{d}$. Starting with the constant term g_0:

$$\begin{aligned} g_0(x) &= \int_{\mathbb{S}^d} g(y)\kappa(dy) \\ &= \int_{\mathbb{S}^d} \sigma(\langle w, y\rangle)^2\kappa(dy) \\ &= \int_{\mathbb{S}^d} \sigma(y_1)^2\kappa(dy) \\ &= \frac{1}{2(d+1)} \end{aligned}$$

Note that $\sigma(z)^2 + \sigma(-z)^2 = z^2$, so we can represent a constant function as a neural network via:

$$\begin{aligned} \sum_{i=1}^{d+1} \sigma(\langle e_i, x\rangle)^2 + \sigma(\langle -e_i, x\rangle)^2 &= \sum_{i=1}^{d+1} \langle e_i, x\rangle^2 \\ &= \|x\|_2 = 1 \end{aligned}$$

So we have $\gamma_1(g_0) \leq 1$.

The second spherical harmonic is given as:

$$\begin{aligned} g_2(x) &= N(d,2)\int_{\mathbb{S}^d} g(y)\frac{(d+1)\langle x, y\rangle^2 - 1}{d}\kappa(dy) \\ &= \frac{N(d,2)}{d}\left((d+1)\int_{\mathbb{S}^d} g(y)\langle x, y\rangle^2\kappa(dy) - \int_{\mathbb{S}^d} g(y)\kappa(dy)\right) \end{aligned}$$

We can represent the constant term as above, and the first integral as

$$\begin{aligned} \int_{\mathbb{S}^d} \sigma(\langle w, y\rangle)^2\langle x, y\rangle^2\kappa(dy) &= \int_{\mathbb{S}^d} \sigma(\langle w, y\rangle)^2(\sigma(\langle x, y\rangle)^2 + \sigma(\langle x, -y\rangle)^2)\kappa(dy) \\ &= \int_{\mathbb{S}^d} \sigma(\langle x, y\rangle)^2(\sigma(\langle w, y\rangle)^2 + \sigma(\langle w, -y\rangle)^2)\kappa(dy) \\ &= \int_{\mathbb{S}^d} \sigma(\langle x, y\rangle)^2\langle w, y\rangle^2\kappa(dy) \end{aligned}$$

This last line is a convex neural network representation using the squared ReLU activation, and thus we have $\gamma_1\left(\int_{\mathbb{S}^d} g(y)\langle x, y\rangle^2\kappa(dy)\right) \leq \int_{\mathbb{S}^d}\langle w, y\rangle^2\kappa(dy) = \frac{1}{d+1}$.

Thus, $\gamma_1(g_2) \leq \frac{N(d,2)}{d}(1+1) \lesssim d$. And all together, $\gamma_1(h) \leq \gamma_1(g) + \gamma_1(g_0) + \gamma(g_2) \lesssim d$.

So by homogeniety the bound on $\|f\|_{\mathcal{S}_1}$ follows.

□

Our choice of f_1 induces a separation between S_1 and S_2.

Theorem A.5 *We have that* $\|f_1\|_{\mathcal{S}_1} \lesssim 1$*, and*

$$\inf_{\|f\|_{\mathcal{S}_2} \leq \delta} \|f - f_1\|_\infty \gtrsim |d^{-11} - d^{-d/3}\delta| \tag{8}$$

Proof:

Because we've subtracted out the 0th and 2nd harmonics, and all other even harmonics are zero, $\tilde{g}$ and h are odd functions.

Consider the signed measure $\nu(dx) := \frac{2\tilde{g}(x)}{\|\tilde{g}\|_{L_1}}\kappa(dx)$, with Jordan decomposition $\nu = \nu^+ - \nu^-$ with the positive measures $\nu^+(dx) := \frac{2\sigma(\tilde{g}(x))}{\|\tilde{g}\|_{L_1}}\kappa(dx)$ and $\nu^-(dx) := \frac{2\sigma(-\tilde{g}(x))}{\|\tilde{g}\|_{L_1}}\kappa(dx)$.

Note that from the oddness of $\tilde{g}$ and symmetry of κ:

$$\begin{aligned}
TV(\nu^-) &= \frac{2}{\|\tilde{g}\|_{L_1}} \int_{\mathbb{S}^d} \sigma(-\tilde{g}(x))\kappa(dx) \\
&= \frac{2}{\|\tilde{g}\|_{L_1}} \int_{\mathbb{S}^d} \sigma(\tilde{g}(-x))\kappa(dx) \\
&= \frac{2}{\|\tilde{g}\|_{L_1}} \int_{\mathbb{S}^d} \sigma(\tilde{g}(x))\kappa(dx) \\
&= TV(\nu^+)
\end{aligned}$$

Because $TV(\nu^+) + TV(\nu^-) = TV(\nu) = 2$, we conclude ν^+ and ν^- are both probability measures. We'll use these measures to separate f and f_1. By Lipschitz continuity of σ:

$$\begin{aligned}
|f(\nu^+) - f(\nu^-)| &= \left| \int_{\mathbb{S}^d} \sigma(\langle \phi, \nu^+ \rangle) - \sigma(\langle \phi, \nu^- \rangle)\chi(d\phi) \right| \\
&\leq \int_{\mathbb{S}^d} |\sigma(\langle \phi, \nu + \nu^- \rangle) - \sigma(\langle \phi, \nu^- \rangle)|\chi(d\phi) \\
&\leq \sup_{\gamma_2(\phi) \leq 1} |\langle \phi, \nu \rangle| \|f\|_{\mathcal{S}_2} \\
&\leq \frac{2}{\|\tilde{g}\|_{L_1}} \sup_{\gamma_2(\phi) \leq 1} |\langle \phi, \tilde{g} \rangle| \|f\|_{\mathcal{S}_2} \\
&\lesssim \frac{2}{\|\tilde{g}\|_{L_1}} d^{1/2 - d/3}\delta
\end{aligned}$$

where in the last line we use Lemma A.3.

Concerning the function f_1, we first use oddness again to notice:

$$\begin{aligned}
\langle h, \nu^- \rangle &= \frac{2}{\|\tilde{g}\|_{L_1}} \int_{\mathbb{S}^d} h(x)\sigma(-\tilde{g}(x))\kappa(dx) \\
&= \frac{2}{\|\tilde{g}\|_{L_1}} \int_{\mathbb{S}^d} h(x)\sigma(\tilde{g}(-x))\kappa(dx) \\
&= \frac{2}{\|\tilde{g}\|_{L_1}} \int_{\mathbb{S}^d} h(-x)\sigma(\tilde{g}(x))\kappa(dx) \\
&= -\langle h, \nu^+ \rangle
\end{aligned}$$

So $\langle h, \nu \rangle = \langle h, \nu^+ - \nu^- \rangle = 2\langle h, \nu^+ \rangle$, and therefore from Lemma A.2 with $\alpha = 2$,

$$\begin{aligned} d^{-21/2} \lesssim \langle g, \tilde{g} \rangle &= \langle h, \tilde{g} \rangle \\ &= \frac{\|\tilde{g}\|_{L_1}}{2} \langle h, \nu \rangle \\ &= \|\tilde{g}\|_{L_1} \langle h, \nu^+ \rangle \end{aligned}$$

So $\langle h, \nu^+ \rangle \gtrsim \frac{d^{-21/2}}{\|\tilde{g}\|_{L_1}}$, and we conclude

$$\begin{aligned} |f_1(\nu^+) - f_1(\nu^-)| &= d^{-1} |\sigma(\langle h, \nu^+ \rangle) - \sigma(\langle h, \nu^- \rangle)| \\ &= d^{-1} \sigma(\langle h, \nu^+ \rangle) \\ &\gtrsim \frac{d^{-23/2}}{\|\tilde{g}\|_{L_1}} \end{aligned}$$

Now, suppose $\|f - f_1\|_\infty \leq \epsilon$. Then

$$\begin{aligned} \frac{d^{-23/2}}{\|\tilde{g}\|_{L_1}} &\lesssim |f_1(\nu^+) - f_1(\nu^-)| \\ &\leq |f_1(\nu^+) - f(\nu^+)| + |f(\nu^+) - f(\nu^-)| + |f(\nu^-) - f_1(\nu^-)| \\ &\lesssim \epsilon + \frac{2}{\|\tilde{g}\|_{L_1}} d^{1/2 - d/3} \delta + \epsilon \end{aligned}$$

So for sufficiently large d, we have $\frac{|d^{-23/2} - d^{1/2-d/3}\delta|}{\|\tilde{g}\|_{L_1}} \lesssim \epsilon$. Finally, note by Jensen's inequality and spherical harmonic orthogonality that $\|\tilde{g}\|_{L_1} \leq \|\tilde{g}\|_{L_2} \leq \|g\|_{L_2} \lesssim d^{-1/2}$.

□

A.4.3. Separation of $\mathcal{S}_2$ and $\mathcal{S}_3$

In order to instantiate the class $\mathcal{S}_3$, we must fix τ, the base probability measure over test functions in $\mathcal{A}_2$. Consider some probability distribution ζ over the square-summable sequences $l_2(\mathbb{R}^+)$ such that for $c \in supp(\zeta)$, $\sum_{k=0}^{\infty} c_k^2 = 1$. Furthermore, we will make the simplyfing assumption that $c_0 = 0$. For each k let κ_k be uniform over $\mathbb{S}^{N(d,k)-1}$, and note that $N(d,1) = d+1$ so $\kappa = \kappa_1$.

Then we sample $\phi \sim \tau$ as $\phi = \sum_{k=1}^{\infty} \sum_{j=0}^{N(d,k)} \lambda_k c_k \alpha_{kj} Y_{kj}$ where $c \sim \zeta$ and $\alpha_k \sim \kappa_k$. Observe that

$$\gamma_2(\phi)^2 = \sum_{k=1, \lambda_k \neq 0}^{\infty} \sum_{j=1}^{N(d,k)} \lambda_k^{-2} \lambda_k^2 c_k^2 \alpha_{kj}^2 = 1$$

so τ indeed samples functions from $\mathcal{A}_2$.

We define $f_2(\mu) = \sigma(\langle g, \mu \rangle)$ where $g = \lambda_1 Y_{1,1}$. Clearly $\gamma_2(g)^2 = \lambda_1^{-2} \lambda_1^2 \|Y_{1,1}\|_{L_2}^2 = 1$, so $\|f_2\|_{\mathcal{S}_2} \leq 1$.

Theorem A.6 *We have that $\|f_2\|_{\mathcal{S}_2} \leq 1$, and*

$$\inf_{\|f\|_{\mathcal{S}_3} \leq \delta} \|f - f_2\|_\infty \gtrsim d^{-2} \delta^{-5/d} \tag{9}$$

Proof: Consider the function $h(x) = \sum_{j=1}^{N(d,1)} \beta_{1,j} Y_{1,j}$ and probability measure $\mu_\beta^*(dx) = \frac{h(x) + \|h\|_\infty}{\|h + \|h\|_\infty\|_{L_1}} \kappa(dx)$. Observe that

$$f_2(\mu_\beta^*) = \frac{\lambda_1}{\|h + \|h\|_\infty\|_{L_1}} \sigma(\langle e_1, \beta \rangle)$$

For a function $f \in \mathcal{S}_3$ with density q with respect to τ, we have:

$$\begin{aligned} f(\mu_\beta^*) &= \int_{\mathcal{A}_2} \sigma(\langle \phi, \mu_\beta^* \rangle) q(\phi) \tau(d\phi) \\ &= \frac{\lambda_1}{\|h + \|h\|_\infty\|_{L_1}} \int_{l_2(\mathbb{R}^+)} \int_{\mathbb{S}^d} \sigma(\langle c_1 \alpha_1, \beta \rangle) \hat{q}(c, \alpha_1) \kappa(d\alpha_1) \zeta(dc) \\ &= \frac{\lambda_1}{\|h + \|h\|_\infty\|_{L_1}} \int_{\mathbb{S}^d} \sigma(\langle \alpha_1, \beta \rangle) \left[\int_{l_2(\mathbb{R}^+)} c_1 \hat{q}(c, \alpha_1) \zeta(dc) \right] \kappa(d\alpha_1) \end{aligned}$$

where $\hat{q}$ marginalizes out all other α_k terms. Let $\tilde{q}(\alpha_1) = \int_{l_2(\mathbb{R}^+)} c_1 \hat{q}(c, \alpha_1) \zeta(dc)$. From the fact that $c_1 \leq 1$, and by Jensen's inequality, $\|\tilde{q}\|_{L_2(\kappa)} \leq \|\hat{q}\|_{L_2(\kappa \times \zeta)} \leq \|q\|_{L_2(\tau)}$.

Now we may appeal to a separation of test function representations acting on spherical inputs. From D.5 in (Bach, 2017a), there exists some $\beta \in \mathbb{S}^d$ such that

$$|\sigma(\langle e_1, \beta \rangle) - \int_{\mathbb{S}^d} \sigma(\alpha_1, \beta) \tilde{q}(\alpha_1) \kappa(d\alpha_1)\| \gtrsim \|\tilde{q}\|_{L_2}^{-5/d} \geq \|q\|_{L_2}^{-5/d}$$

Therefore

$$|f_2(\mu_\beta^*) - f(\mu_\beta^*)| \gtrsim \frac{\lambda_1}{\|h + \|h\|_\infty\|_{L_1}} \|q\|_{L_2}^{-5/d}$$

Finally, note that $\lambda_1 \simeq d^{-1}$, and by the addition formula and the fact $P_k(1) = 1$ for all k:

$$\begin{aligned} \|h + \|h\|_\infty\|_{L_1} &\leq 2\|h\|_\infty \\ &= 2 \max_{x \in \mathbb{S}^d} \sum_{j=1}^{N(d,1)} \beta_{1,j} Y_{1,j}(x) \\ &\leq 2 \max_{x \in \mathbb{S}^d} \|\beta\|_2 \sqrt{\sum_{j=1}^{N(d,1)} Y_{1,j}(x)^2} \\ &\leq 2N(d,1) \\ &\lesssim d \end{aligned}$$

So we arrive at the desired bound.

□

B. Experimental Details and Additional Data

Synthetic Details: For all experiments we use the same architecture. Namely, for an input set $x = (x_1, \ldots, x_N)$, the network is defined as $f_N(x) = w_3^T \sigma(W_2 \frac{1}{N} \sum_{i=1}^N \sigma(W_1 \tilde{x}_i))$, where we choose the architecture as $W_1 \in \mathbb{R}^{h_1 \times d}$,

$W_2 \in \mathbb{R}^{h_2 \times h_1}$, and $w_3 \in \mathbb{R}^{h_2}$. Here, $h_1, h_2 = 100$ for $\mathcal{S}_1$, $h_1 = 100$ and $h_2 = 1000$ for $\mathcal{S}_2$, and $h_1 = h_2 = 1000$ for $\mathcal{S}_3$. The weights are initialized with the uniform Kaiming initialization (He et al., 2015) and frozen as described in Table 1.

We relax the functional norm constraints to penalties, by introducing regularizers of the form $\lambda \|f_N\|_{\mathcal{S}_i}$ for λ a hyperparameter. Let $K(\cdot)$ map a matrix to the vector of row-wise squared norms, and let $|\cdot|$ denote the element-wise absolute value of a matrix. Then we calculate the functional norms via the path norm as follows:

- For $\mathcal{S}_1$, $\|f_N\|_{\mathcal{S}_1} = |w_3|^T |W_2| K(W_1)$
- For $\mathcal{S}_2$, we explicitly normalize the frozen matrix W_1 to have all row-wise norms equal to 1, then $\|f_N\|_{\mathcal{S}_2} = |w_3|^T K(W_2)$
- For $\mathcal{S}_3$, we normalize the rows of W_1 and W_2, which simply implies $\|f_N\|_{\mathcal{S}_3} = \|w_3\|_2$

We optimized via Adam (Kingma & Ba, 2014) with an initial learning rate of 0.0005, for 5000 iterations. Under this architecture, all $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ functions achieved less than 10^{-15} training error without regularization on all objective functions (listed below) on training sets of 100 samples.

We use the following symmetric functions for our experiments:

- $f_N^*(x) = \max_i(\|x_i\|_2^{-1})$
- $f_N^*(x) = \lambda \log\left(\sum_{i=1}^N \exp(\|x_i\|_2^{-1}/\lambda)\right)$ for $\lambda = 0.1$
- $f_N^*(x) = \text{median}(\{\|x_i\|_2^{-1}\}_{i=1}^N)$
- $f_N^*(x) = \text{second}_i(\|x_i\|_2^{-1})$ i.e. the second largest value in a given set
- $f_N^*(x) = \frac{1}{N}\sum_{i=1}^N (\|x_i\|_2^{-1})$
- $f_N^*(x) = \frac{2}{N(N-1)} \sum_{i<j} \frac{1}{\|x_i - x_j\|_2}$
- $f_N^*(x)$ is an individual neuron, parameterized the same as f_N but with different hidden layer sizes. For the neuron, $h_1 = h_2 = 1$, for the smooth_neuron, $h_1 = 100$ and $h_2 = 1$. Additionally, the proof of Theorem A.5 dictates that we must choose the neuron's test function to have large kernel norm, so we initialize W_1 elementwise from the Gaussian mixture with density $0.5 * \mathcal{N}(1, 0.5) + 0.5 * \mathcal{N}(-1, 0.5)$.

Note that in order to guarantee the "smooth_neuron" is representable by our finite-width networks, we explicitly set W_1 in the $\mathcal{S}_2$ and $\mathcal{S}_3$ models to equal the W_1 matrix of the "smooth_neuron".

For each model in each experiment, λ was determined through cross validation over $\lambda \in [0, 10^{-6}, 10^{-4}, 10^{-2}]$ using fresh samples of training data, and choosing the value of λ with lowest generalization error, which was calculated from another 1000 sampled points.

Then, with determined λ, each model was trained from scratch over 10 runs with independent random initializations. The mean and standard deviation of the generalization error, testing on varying values of N, are plotted in Figure 1.

Application Details: For the MNIST experiment with results given in Table 4, we follow a similar setup to (De Bie et al., 2019). From an image in $\mathbb{R}^{28\times 28}$, we produce a point cloud by considering a set of tuples of the form (r, c, t), which are the row, column and intensity respectively for each pixel. We restrict to pixels where $t > 0.5$, and select the pixels with the top 200 intensities to comprise the point cloud (if there are fewer than 200 pixels remaining after thresholding, we resample among them). Furthermore, we normalize the row and column values among all the points in the cloud. This process maps an image to a set $S \subseteq \mathbb{R}^3$ such that $|S| = 200$.

For this dataset we consider $h_1 = 500$ and $h_2 = 500$ for our $\mathcal{S}_i$ finite-width architectures.

We perform cross-validation by setting aside 10% of the data as a validation set, and calculate the mean and standard deviation of the generalization error over five runs. In order to study generalization in this setting, we test on point clouds of different size, 100 and 200, and show the results in Table 4. The starting learning rate is 0.001. Otherwise, all other experimental details are the same as above.

	Error ($N = 100$)	Error ($N = 200$)
$\mathcal{S}_1$	8.03	5.62
$\mathcal{S}_2$	8.25	5.78
$\mathcal{S}_3$	14.45	10.80

Table 4: Classification test error on Pointcloud MNIST in percent, after images are compressed into sets of size N, trained with $N = 200$.

Robust Mean Details: We use the regular ReLU activation in the first layer for training stability. Each network is trained on a batch of 5000 input sets sampled as above, as the task of robust estimation appears more susceptible to overfitting than the simpler symmetric objectives learned in the previous section. All networks are trained for 30000 iterations, and all other details of training are kept consistent with the previous section (including the larger number of random kernel features).

The hyperparameters required for the adversarial estimator in (Diakonikolas et al., 2017) are τ and "cher", which both control the thresholding of which vectors are discarded according to the projection on the maximal eigenvector of the empirical covariance. Cross validation over the sets $[0.1, 0.15, 0.2]$ and $[1.5, 1.8, 2.0, 2.3]$ yielded the choices $\tau = 0.1$ and "cher" $= 1.5$.

Additional Experiments In Figure 4 we consider higher dimensional vectors for our set inputs to the symmetric models. In Figure 5 we consider training over multiple set sizes as well, with the input size sampled uniformly from $4, 5, 6$.

We consider the Pointcloud MNIST dataset, after mapping our image to sets. This dataset is substantially more difficult than regular MNIST, as the induced sets obfuscate the geometric structure of the original images. The results on Pointcloud MNIST, across differently-sized set representations of images, are given in Table 4. The fact that we only consider three-layer networks limits the ability of the model to reconstruct the original image representation and perform comparably to a model acting on regular MNIST. Nevertheless, we still observe the expected ordering of our functional spaces. When testing on smaller sets than training, the generalization error increases faster for $\mathcal{S}_3$ than for $\mathcal{S}_1$ and $\mathcal{S}_2$.

In Table 5 we consider the robust mean experiment, using the same hyperparameters except training on sets of larger size ($N = 60$) and plotting MSE on sets of varying size. As with the smaller scale experiment, we observe that $\mathcal{S}_1$ enjoys a slight advantage over the other methods when restricting attention to the in-distribution generalization setting of $N = 60$, but outside that range the performance is comparable to the naive sample mean, suggesting that out-of-distribution generalization for the robust mean is not easily attainable for these networks.

	$N = 20$	$N = 40$	$N = 60$	$N = 80$	$N = 100$
$\mathcal{S}_1$	0.149 ± 0.039	0.073 ± 0.023	0.043 ± 0.004	0.034 ± 0.004	0.028 ± 0.003
$\mathcal{S}_2$	0.151 ± 0.039	0.076 ± 0.023	0.045 ± 0.004	0.036 ± 0.004	0.030 ± 0.003
$\mathcal{S}_3$	0.159 ± 0.039	0.081 ± 0.023	0.050 ± 0.004	0.040 ± 0.004	0.034 ± 0.003
Sample Mean	0.152 ± 0.069	0.066 ± 0.029	0.055 ± 0.025	0.034 ± 0.015	0.026 ± 0.012
Geometric Median	0.137 ± 0.062	0.063 ± 0.028	0.047 ± 0.021	0.032 ± 0.014	0.025 ± 0.011
Adversarial Estimator	0.472 ± 0.545	0.386 ± 0.555	0.346 ± 0.546	0.282 ± 0.521	0.206 ± 0.455

Table 5: Mean squared test error for robust mean estimation among the finite model instantiations and baselines.

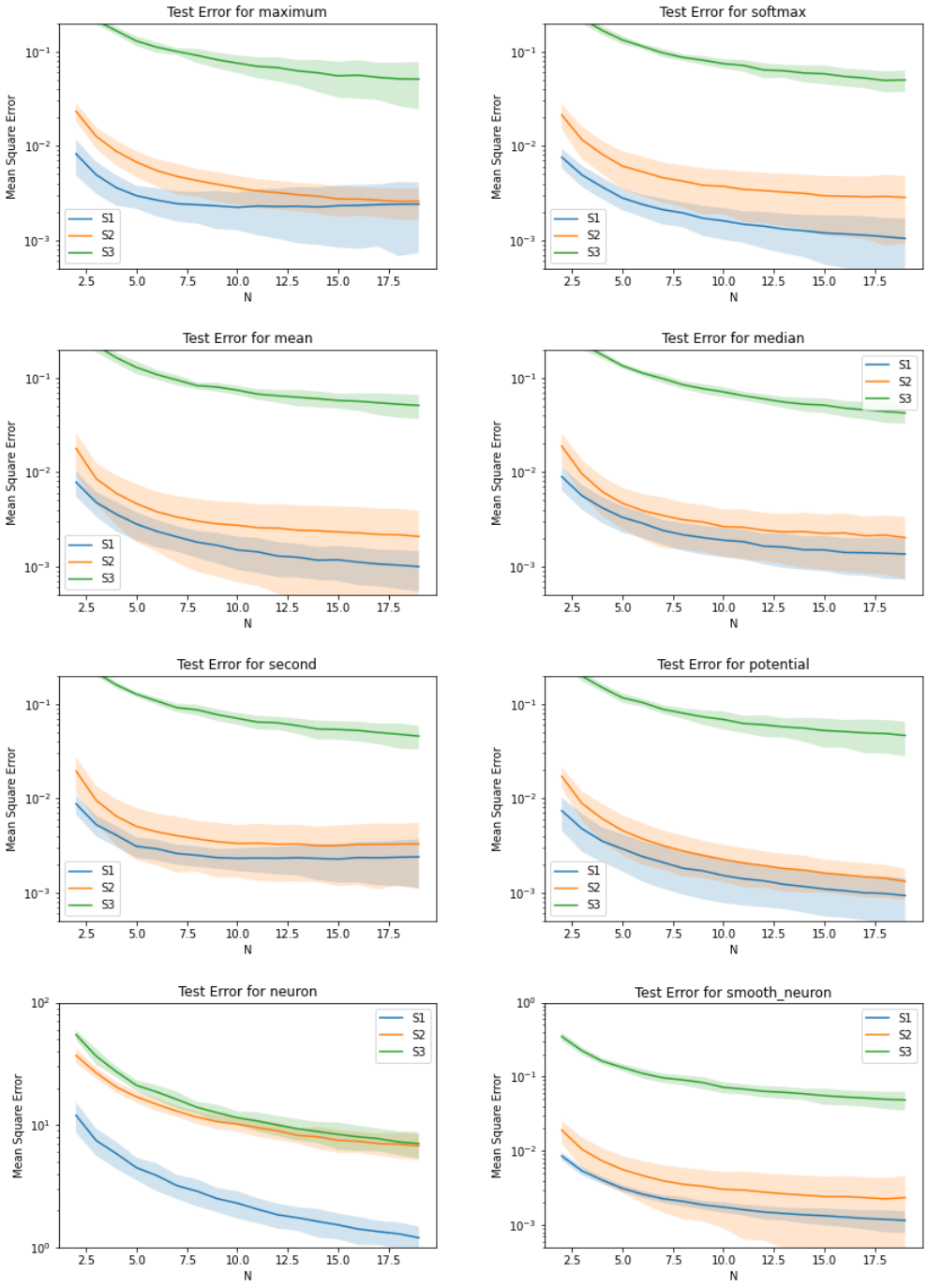

Figure 4: Test Error for $d = 20$

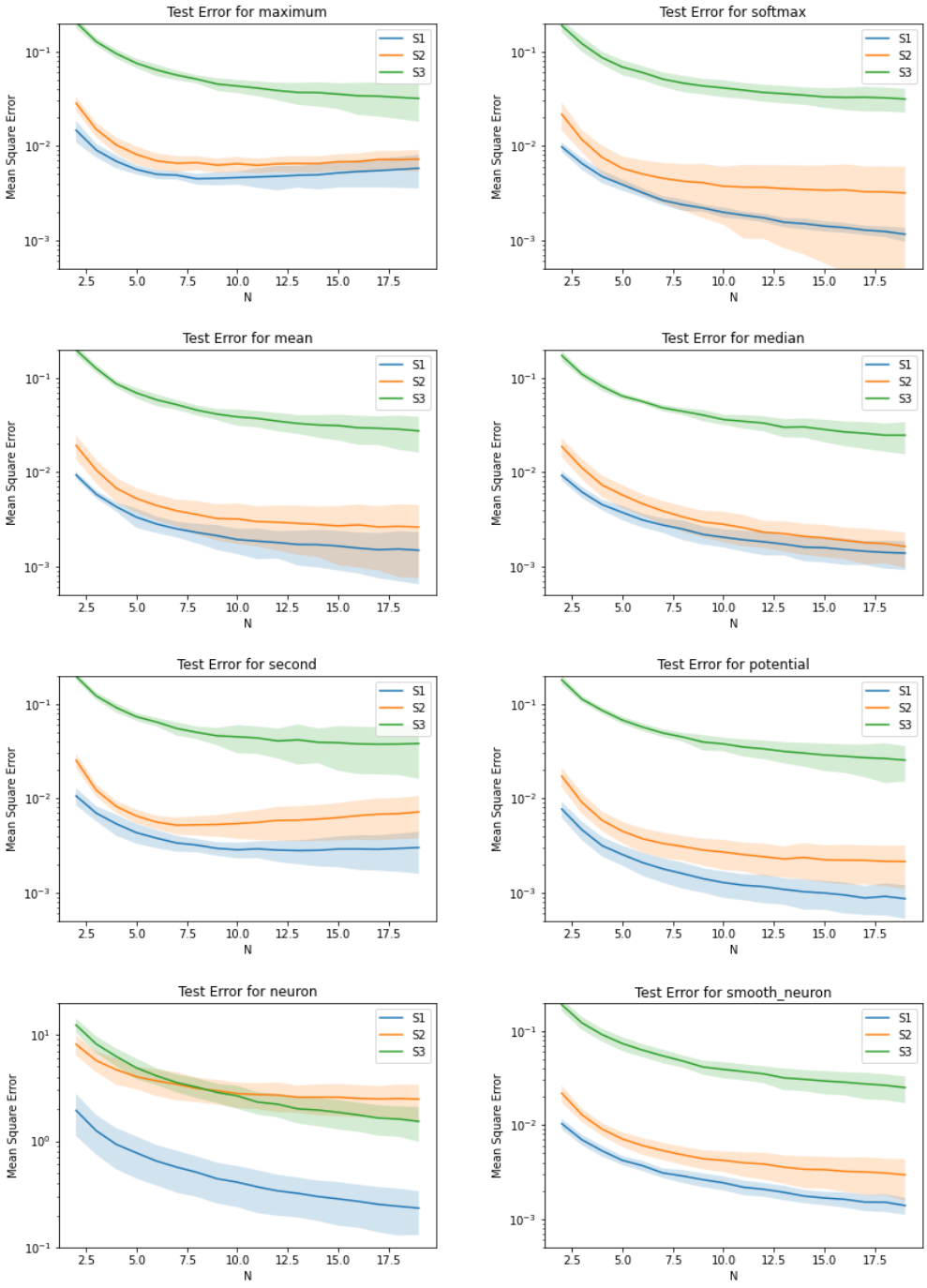

Figure 5: Test Error for varied input size training