SPARSE DEEP NEURAL NETWORKS USING $L_{1,\infty}$-WEIGHT NORMALIZATION

University of Science and Technology of China and Purdue University

Supplementary Material

S1 TECHNICAL LEMMAS

Lemma 1. *(Ledoux and Talagrand, 2013) Assume that the hypothesis class* $\mathcal{F} \subseteq \{f|f : \mathcal{X} \to \mathbb{R}\}$*, and* $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathcal{X}$*. Let* $G : \mathbb{R} \to \mathbb{R}$ *be convex and increasing. Assume that the function* $\phi : \mathbb{R} \to \mathbb{R}$ *is L-Lipschitz continuous, and satisfies that* $\phi(0) = 0$*. We have:*

$$\mathbb{E}_\epsilon \left[G\left(\sup_{f\in\mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(f(\boldsymbol{x}_i)) \right) \right) \right] \leq \mathbb{E}_\epsilon \left[G\left(L \sup_{f\in\mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(\boldsymbol{x}_i) \right) \right) \right]$$

The lemma below is a more general version of (Mohri et al., 2012, Theorem 3.1), where they assume $a = 0$, and the proof is very similar to the original one.

Lemma 2. *Let* z *be a random variable of support* $\mathcal{Z}$ *and distribution* $\mathcal{D}$*. Let* $S = \{z_1 \dots z_n\}$ *be a data set of* n *i.i.d. samples drawn from* $\mathcal{D}$*. Let* $\mathcal{F}$ *be a hypothesis class satisfying* $\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to [0, A_0]\}$*. Fix* $\delta \in (0, 1)$*. With probability at least* $1 - \delta$ *over the choice*

of S, the following holds for all $h \in \mathcal{F}$:

$$\left|\mathbb{E}_{\mathcal{D}}[h] - \widehat{\mathbb{E}}_S[h]\right| \leq 2\mathfrak{R}_n(\mathcal{F}) + A_0\sqrt{\frac{\log(2/\delta)}{2n}}$$

Proof. According to Mohri et al. (2012)[Theorem 3.1], fix $\delta \in (0,1)$. With probability at least $1 - \frac{\delta}{2}$ over the choice of S, the following holds for all $h \in \mathcal{F}$:

$$\mathbb{E}_{\mathcal{D}}[h/A_0] - \widehat{\mathbb{E}}_S[h/A_0] \leq 2\mathfrak{R}_n(\mathcal{F}/A_0) + \sqrt{\frac{\log(2/\delta)}{2n}}$$

With probability at least $1 - \frac{\delta}{2}$ over the choice of S, the following holds for all $h \in \mathcal{F}$:

$$\mathbb{E}_{\mathcal{D}}[-h/A_0] - \widehat{\mathbb{E}}_S[-h/A_0] \leq 2\mathfrak{R}_n(-\mathcal{F}/A_0) + \sqrt{\frac{\log(2/\delta)}{2n}}$$

By the definition of Rademacher complexity, $\mathfrak{R}_n(-\mathcal{F}/A_0) = \mathfrak{R}_n(\mathcal{F}/A_0) = \mathfrak{R}_n(\mathcal{F})/A_0$. Thus we complete the proof. □

Lemma 3. *For any $f \in \mathcal{SN}_{c,\boldsymbol{o}}^{k,\boldsymbol{d},\sigma}$ and $\boldsymbol{x} \in \mathcal{X}$,*

$$\|f(\boldsymbol{x})\|_\infty \leq \|\boldsymbol{o}\|_\infty \max\left(1, (c\rho_\sigma)^k\right).$$

Proof. We instead prove the result for any

$$f \in \mathcal{D}_{c,r}^{k,\boldsymbol{d},\sigma} \triangleq \{g : g \in \mathcal{SN}_{c,\boldsymbol{o}}^{k,\boldsymbol{d},\sigma}, \forall \boldsymbol{o} : \|\boldsymbol{o}\|_\infty \leq r\},$$

and complete the proof by induction on depth $k+1$. When $k=0$,

$$\begin{aligned}
\sup_{f\in\mathcal{D}^{0,}_{c,r}\boldsymbol{d},\sigma} \|f(\boldsymbol{x})\|_\infty &= \sup_{f\in\mathcal{D}^{0,}_{c,r}\boldsymbol{d},\sigma} \left\|\tilde{\mathbf{V}}_1^T(1, f_0^T(\boldsymbol{x}))^T\right\|_\infty \\
&= r \sup_{f\in\mathcal{D}^{0,}_{c,r}\boldsymbol{d},\sigma} \frac{\left\|\left(\tilde{\mathbf{V}}_1^T(1, f_0^T(\boldsymbol{x}))^T\right)\right\|_{p^*}}{\left\|\tilde{\mathbf{V}}_1\right\|_{1,\infty}} \\
&\le r \sup_{f\in\mathcal{D}^{0,}_{c,r}\boldsymbol{d},\sigma} \frac{1}{\left\|\tilde{\mathbf{V}}_1\right\|_{1,\infty}} \left\|\tilde{\mathbf{V}}_1^T(1, \boldsymbol{x}^T)^T\right\|_\infty \\
&\le r\max(1, \|\boldsymbol{x}\|_\infty) \\
&\le r.
\end{aligned}$$

Define $\boldsymbol{d}_{k+} = (d_0, \cdots, d_{k-1}, d_k+1)$.

$$\begin{aligned}
\sup_{f\in\mathcal{D}^{k,}_{c,r}\boldsymbol{d},\sigma} \|f(\boldsymbol{x})\|_\infty &= \sup_{f\in\mathcal{D}^{k,}_{c,r}\boldsymbol{d},\sigma} \left\|\tilde{\mathbf{V}}_{k+1}^T(1, \sigma\circ f_k^T(\boldsymbol{x}))^T\right\|_\infty \\
&= r \sup_{f\in\mathcal{D}^{k,}_{c,r}\boldsymbol{d},\sigma} \frac{\left\|\left(\tilde{\mathbf{V}}_{k+1}^T(1, \sigma\circ f_k^T(\boldsymbol{x}))^T\right)\right\|_\infty}{\left\|\tilde{\mathbf{V}}_{k+1}\right\|_{1,\infty}} \\
&\le r \sup_{f\in\mathcal{D}^{k,}_{c,r}\boldsymbol{d},\sigma} \frac{1}{\left\|\tilde{\mathbf{V}}_{k+1}\right\|_{1,\infty}} \left\|\tilde{\mathbf{V}}_{k+1}^T(1, \sigma\circ f_k^T(\boldsymbol{x}))^T\right\|_\infty \\
&= r \sup_{f\in\mathcal{D}^{k,}_{p,q,c,r}\boldsymbol{d},\sigma} \frac{1}{\|\boldsymbol{v}\|_1} \left|\langle \boldsymbol{v}, (1, \sigma\circ f_k^T(\boldsymbol{x}))^T\rangle\right| \\
&\le r \left\|(1, \sigma\circ f_k^T(\boldsymbol{x}))\right\|_\infty \\
&\le r \left\|(1, \rho_\sigma f_k^T(\boldsymbol{x}))\right\|_\infty \\
&\le r\max(1, c\rho_\sigma) \sup_{f\in\mathcal{D}^{k-1,}_{c,1}\boldsymbol{d}_{k+},\sigma} \|f(\boldsymbol{x})\|_\infty
\end{aligned}$$

The penultimate step follows from the fact that

$$(1, \rho_\sigma f_k^T)^T \in \max(1, c\rho_\sigma)\mathcal{D}_{c,1}^{k-1,\boldsymbol{d}_{k+},\sigma}.$$

Finally, the proof is completed by the induction assumption. □

Lemma 4. *Assume A1-A2 hold. In addition, the loss function* $L(f(\boldsymbol{x}), y) : \mathcal{Y} \times \mathcal{Y} \to [0, A_0]$*, is* ρ*-Lipschitz continuous on its first argument. Fix* $\delta \in (0, 1)$ *and* $o > 0$*, then with probability at least* $1 - \delta$ *over the choice of the sample, every* $f \in \mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}$ *satisfies that*

$$\mathcal{E}_L(f) \leq A_0\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2o\rho}{\sqrt{n}}\left[\sqrt{(k+1)\log 16}\left(\sum_{\ell=0}^{k}(c\rho_\sigma)^\ell + (c\rho_\sigma)^k\right) + (c\rho_\sigma)^k\sqrt{2\log(2m_1)}\right].$$

Furthermore, if $c\rho_\sigma \geq 1$*, with probability at least* $1 - \delta$ *over the choice of the sample, every* $f \in \mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}$ *satisfies that*

$$\mathcal{E}_L(f) \leq A_0\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2o\rho}{\sqrt{n}}(c\rho_\sigma)^k(\sqrt{(k+3)\log 4} + \sqrt{2\log(2m_1)}).$$

Proof. First, we upper bound $\mathfrak{R}_n(\mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma})$ by the same bounds in Theorem 1, as Theorem 1 holds for any sample S under our assumptions. Then we could further bound the Rademacher complexity of the corresponding hypothesis class according to Lemma 1 and A2. Finally we get the desired result by Lemma 2. □

Lemma 5. *Assume B1-B2 hold. In addition, the loss function* $L(f(\boldsymbol{x}), y) : \mathcal{Z} \times \mathcal{Y} \to [0, A_0]$*, satisfies that*

$$|L(f_1(\boldsymbol{x}), y) - L(f_2(\boldsymbol{x}), y)| \leq \rho \left\|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})\right\|_2$$

for any $\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$. *Fix* $\delta \in (0,1)$ *and* $\boldsymbol{o} \geq \mathbf{0}$, *then with probability at least* $1-\delta$ *over the choice of the sample, every* $f \in \mathcal{SN}_{c,\boldsymbol{o}}^{k,\boldsymbol{d},\sigma}$ *satisfies that*

$$\mathcal{E}_L(f) \leq A_0\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2\sqrt{2}\rho}{\sqrt{n}}\left(\sum_{j=1}^{m_2} o_j\right)\left[\sqrt{(k+1)\log 16}\left(\sum_{\ell=0}^{k}(c\rho_\sigma)^\ell + (c\rho_\sigma)^k\right) + (c\rho_\sigma)^k\sqrt{2\log(2m_1)}\right].$$

Furthermore, if $c\rho_\sigma \geq 1$, *with probability at least* $1-\delta$ *over the choice of the sample, every* $f \in \mathcal{SN}_{c,\boldsymbol{o}}^{k,\boldsymbol{d},\sigma}$ *satisfies that*

$$\mathcal{E}_L(f) \leq A_0\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{2\sqrt{2}\rho}{\sqrt{n}}\left(\sum_{j=1}^{m_2} o_j\right)(c\rho_\sigma)^k(\sqrt{(k+3)\log 4} + \sqrt{2\log(2m_1)}).$$

Proof. First, we upper bound $\mathfrak{R}_n(\mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma})$ by the same bounds in Theorem 1, as Theorem 1 holds for any sample S under our assumptions. Then we could further bound the Rademacher complexity of the corresponding hypothesis class by (Maurer, 2016, Corollary 1) and B2. Finally get the desired result by Lemma 2. □

S2 DETAILED PROOFS

Proof of Theorem 1

Proof. We first show that

$$\widehat{\mathfrak{R}}_S(\mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}) \leq o\sqrt{\frac{(k+1)\log 16}{n}}\left(\sum_{\ell=0}^{k}(c\rho_\sigma)^\ell + (c\rho_\sigma)^k\right) + o(c\rho_\sigma)^k\sqrt{\frac{2\log(2m_1)}{n}}.$$

The proof has two main steps. Fixing the sample S, and the architecture of the DNN, define a series of random variables $\{Z_0, Z_1, \cdots, Z_k\}$ as

$$Z_0 = \left\| \sum_{i=1}^{n} \epsilon_i \boldsymbol{x}_i \right\|_\infty$$

and

$$Z_j = \sup_{f \in \mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}} \left\| \sum_{i=1}^{n} \epsilon_i \sigma \circ f_j(\boldsymbol{x}_i) \right\|_\infty ,$$

for $j = 1, \cdots, k$, where $\{\epsilon_1, \cdots, \epsilon_n\}$ are n independent Rademacher random variables, and f_j denotes the jth hidden layer of the sparse DNN f. In the first step, we prove by induction that for $j = 1, \cdots, k$ and any $t \in \mathbb{R}$

$$\mathbb{E}_\epsilon \exp(tZ_j) \leq 4^j \exp\left(\frac{t^2 n s_j^2}{2} + t(c\rho_\sigma)^j \sqrt{2n \log(2m_1)} \right),$$

where

$$s_j = \sum_{i=1}^{j} (c\rho_\sigma)^i + (c\rho_\sigma)^j.$$

Note that $s_{j+1} = c\rho_\sigma(s_j + 1)$.

When $j = 0$, by (Kakade et al., 2009, Theorem 3), $E_\epsilon Z_0 \leq \sqrt{2n \log(2m_1)}$. Note that Z_0 is a deterministic function of the i.i.d.random variables $\epsilon_1, \cdots, \epsilon_n$, and satisfies that

$$|Z_0(\epsilon_1, \cdots, \epsilon_i, \cdots, \epsilon_n) - Z_0(\epsilon_1, \cdots, -\epsilon_i, \cdots, \epsilon_n)| \leq 2 \max \|\mathbf{x}_i\|_\infty$$

by Minkowski inequality. By the proof of Theorem 6.2 (Boucheron et al., 2003), Z_0 is subgaussian satisfies that

$$\mathbb{E}_\epsilon \exp\left(tZ_0\right) = \mathbb{E}_\epsilon \exp\left(t(Z_0 - E_\epsilon Z_0)\right) * \exp\left(tE_\epsilon Z_0\right) \leq \exp\left(\frac{t^2 n}{2} + t\sqrt{2n \log(2m_1)} \right)$$

for any $t \in \mathbb{R}$. For the case when $j \geq 1$,

$$\begin{aligned}
\mathbb{E}_\epsilon \exp\left(tZ_j\right) &= \mathbb{E}_\epsilon \exp\left(t \sup_{\substack{\left\|\tilde{\mathbf{V}}_j\right\|_{1,\infty} \leq c \\ f \in \mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}} \left\| \sum_{i=1}^n \epsilon_i \sigma\left(\tilde{\mathbf{V}}_j^T (1, \sigma \circ f_{j-1}(\boldsymbol{x}_i))\right) \right\|_\infty \right) \\
&= \mathbb{E}_\epsilon \exp\left(t \sup_{\substack{\|\boldsymbol{v}\|_1 \leq c \\ f \in \mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}} \left| \sum_{i=1}^n \epsilon_i \sigma\left(\boldsymbol{v}^T (1, \sigma \circ f_{j-1}(\boldsymbol{x}_i))\right) \right| \right) \\
&\leq 2\mathbb{E}_\epsilon \exp\left(t \sup_{\substack{\|\boldsymbol{v}\|_1 \leq c \\ f \in \mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}} \sum_{i=1}^n \epsilon_i \sigma\left(\boldsymbol{v}^T (1, \sigma \circ f_{j-1}(\boldsymbol{x}_i))\right) \right) \qquad \text{(S2.1a)} \\
&\leq 2\mathbb{E}_\epsilon \exp\left(t\rho_\sigma \sup_{\substack{\|\boldsymbol{v}\|_1 \leq c \\ f \in \mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}} \sum_{i=1}^n \epsilon_i \left(\boldsymbol{v}^T (1, \sigma \circ f_{j-1}(\boldsymbol{x}_i))\right) \right) \qquad \text{(S2.1b)} \\
&\leq 2\mathbb{E}_\epsilon \exp\left(tc\rho_\sigma \sup_{f \in \mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}} \left\| \sum_{i=1}^n \epsilon_i (1, \sigma \circ f_{j-1}(\boldsymbol{x}_i)) \right\|_\infty \right) \\
&\leq 2\mathbb{E}_\epsilon \exp\left(tc\rho_\sigma (|\sum_{i=1}^n \epsilon_i| + \sup_{f \in \mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}} \left\| \sum_{i=1}^n \epsilon_i \sigma \circ f_{j-1}(\boldsymbol{x}_i) \right\|_\infty) \right) \\
&\leq 2 \left[2\mathbb{E}_\epsilon \exp\left(r_j tc\rho_\sigma \sum_{i=1}^n \epsilon_i \right) \right]^{\frac{1}{r_j}} \left[\mathbb{E}_\epsilon \exp\left(r_j^* tc\rho_\sigma Z_{j-1} \right) \right]^{\frac{1}{r_j^*}} \qquad \text{(S2.1c)} \\
&\leq 4^j \exp\left(\frac{nt^2c^2\rho_\sigma^2 (1 + s_{j-1})^2}{2} + tc^j \rho_\sigma^j \sqrt{2n \log(2m_1)} \right) \\
&= 4^j \exp\left(\frac{nt^2c^2\rho_\sigma^2 s_j^2}{2} + tc^j \rho_\sigma^j \sqrt{2n \log(2m_1)} \right)
\end{aligned}$$

The step in equation (S2.1a) follows from the observation that

$$\mathbb{E}_\epsilon \exp\left(t \sup_{\|\boldsymbol{v}\|_1 \le c} |\sum_{i=1}^n \epsilon_i \sigma\left(\boldsymbol{v}^T \sigma \circ f_{j-1}^*(\boldsymbol{x}_i)\right)|\right) \le E_\epsilon \exp\left(t \sup_{\|\boldsymbol{v}\|_1 \le c} \sum_{i=1}^n \epsilon_i \sigma\left(\boldsymbol{v}^T (1, \sigma \circ f_{j-1}(\boldsymbol{x}_i))\right)\right) + \mathbb{E}_\epsilon \exp\left(t \sup_{\|\boldsymbol{v}\|_1 \le c} \sum_{i=1}^n (-\epsilon_i) \sigma\left(\boldsymbol{v}^T \sigma \circ f_{j-1}(\boldsymbol{x}_i)\right)\right) \tag{S2.2}$$

The step in equation (S2.1b) follows from Lemma 1. Note that equation (S2.1c) holds for any $r_j > 1$ and $r_j^* = \frac{r_j}{r_j - 1}$ by Hölder's inequality $\mathbb{E}(|XY|) \le \mathbb{E}(|X|_j^r)^{\frac{1}{r_j}} \mathbb{E}(|Y|^{r_j^*})^{\frac{1}{r^*}}$. The step in equation (S2.1c) follows from $\mathbb{E}_\epsilon \exp(|X|) \le \mathbb{E}_\epsilon \exp(X) + \mathbb{E}_\epsilon \exp(-X)$. Note that $\sum_{i=1}^n \epsilon_i$ is also a deterministic function of the i.i.d.random variables $\epsilon_1, \cdots, \epsilon_n$, satisfying that $\mathbb{E}_\epsilon \sum_{i=1}^n \epsilon_i = 0$ and

$$|\sum_{i \ne j} \epsilon_i + \epsilon_j - (\sum_{i \ne j} \epsilon_i - \epsilon_j)| \le 2.$$

Then by the proof of Theorem 6.2 (Boucheron et al., 2003),

$$\mathbb{E}_\epsilon \exp(t \sum_{i=1}^n \epsilon_i) \le \exp(\frac{t^2 n}{2}) \tag{S2.3}$$

for any $t \in \mathbb{R}$. Then we get the desired result by choosing the optimal $r_k = s_{k-1} + 1$ while following the induction assumption.

The second step is based on the idea of (Golowich et al., 2018) using Jensen's inequality. For any $\lambda > 0$,

$$n\widehat{\mathfrak{R}}_S(\mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}} \left(\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i)\right)\right]$$

$$\leq \frac{1}{\lambda}\log \mathbb{E}_\epsilon \exp\left(\lambda \sup_{f\in\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}\left(\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right)\right)$$

$$\leq \frac{1}{\lambda}\log \mathbb{E}_\epsilon \exp\left(\lambda o \sup_{f\in\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}\left\|\sum_{i=1}^{n}\epsilon_i(1,\sigma\circ f_k(\boldsymbol{x}_i))\right\|_\infty\right)$$

$$\leq \frac{1}{\lambda}\left[(k+1)\log 4 + \frac{\lambda^2 o^2 n(s_k+1)^2}{2} + \lambda o c^k \rho_\sigma^k\sqrt{2n\log(2m_1)}\right] \quad \text{(S2.4a)}$$

where the step in equation (S2.4a) is derived using a similar technique as in equation (S2.1a) to equation (S2.1c). By choosing the optimal $\lambda = \frac{\sqrt{(k+1)\log 16}}{o(s_k+1)\sqrt{n}}$, we have

$$\widehat{\mathfrak{R}}_S(\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}) \leq o\sqrt{\frac{(k+1)\log 16}{n}}\left(\sum_{i=0}^{k}(c\rho_\sigma)^i + (c\rho_\sigma)^k\right) + o(c\rho_\sigma)^k\sqrt{\frac{2\log(2m_1)}{n}}$$

We then show that

$$\widehat{\mathfrak{R}}_S(\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}) \leq \frac{1}{\sqrt{n}}o(c\rho_\sigma)^k(\sqrt{(k+3)\log 4} + \sqrt{2\log(2m_1)}).$$

when $c\rho_\sigma \geq 1$. Similar to the general case, the proof has two main steps. In the first step, we prove by induction that for $j = 1, \cdots, k$ and any $t \in \mathbb{R}$

$$\mathbb{E}_\epsilon \exp(tZ_j) \leq \sum_{i=1}^{j} 2^{j-i+2}\exp\left(\frac{nt^2c^{2(j-i+1)}\rho_\sigma^{2j}}{2}\right) + 2^j\exp\left(\frac{nt^2c^{2j}\rho_\sigma^{2j}}{2} + t(c\rho_\sigma)^j\sqrt{2n\log(2m_1)}\right)$$

When $j = 0$, we already have for any $t \in \mathbb{R}$,

$$\mathbb{E}_\epsilon \exp(tZ_0) \leq \exp\left(\frac{t^2 n\max\|\mathbf{x}_i\|_\infty^2}{2} + tA^1_{m_1,S}\right).$$

For the case when $j \geq 1$,

$$
\begin{aligned}
\mathbb{E}_\epsilon \exp\left(tZ_j\right) &= \mathbb{E}_\epsilon \exp\left(t \sup_{\substack{\left\|\tilde{\mathbf{V}}_j\right\|_{1,\infty} \leq c \\ f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma}} \left\| \sum_{i=1}^n \epsilon_i \sigma_j \circ (\tilde{\mathbf{V}}_j^T (1, \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i)) \right\|_\infty \right) \\
&= \mathbb{E}_\epsilon \exp\left(t \sup_{\substack{\|\boldsymbol{v}\|_1 \leq c \\ f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma}} |\sum_{i=1}^n \epsilon_i \sigma_j \left(\boldsymbol{v}^T (1, \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i)) \right)| \right) \\
&\leq 2\mathbb{E}_\epsilon \exp\left(t \sup_{\substack{\|\boldsymbol{v}\|_1 \leq c \\ f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma}} \sum_{i=1}^n \epsilon_i \sigma_j \left(\boldsymbol{v}^T (1, \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i)) \right) \right) \qquad \text{(S2.5a)} \\
&\leq 2\mathbb{E}_\epsilon \exp\left(t\rho_\sigma \sup_{\substack{\|\boldsymbol{v}\|_1 \leq c \\ f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma}} \sum_{i=1}^n \epsilon_i \boldsymbol{v}^T (1, \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i)) \right) \qquad \text{(S2.5b)} \\
&\leq 2\mathbb{E}_\epsilon \exp\left(tc\rho_\sigma \sup_{f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma} \left\| \sum_{i=1}^n \epsilon_i (1, \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i)) \right\|_\infty \right) \\
&\leq 2\mathbb{E}_\epsilon \exp\left(tc\rho_\sigma \max\left(|\sum_{i=1}^n \epsilon_i|, \sup_{f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma} \left\| \sum_{i=1}^n \epsilon_i \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i) \right\|_\infty \right) \right) \\
&= 2\mathbb{E}_\epsilon \max\left(\exp(tc\rho_\sigma |\sum_{i=1}^n \epsilon_i|), \exp\left(\sup_{f \in \mathcal{SN}^k_{c,o}(\boldsymbol{d},\sigma} \left\| \sum_{i=1}^n \epsilon_i \sigma_{j-1} \circ f_{j-1}(\boldsymbol{x}_i) \right\|_\infty \right) \right) \right) \\
&\leq 2\mathbb{E}_\epsilon \exp\left(tc\rho_\sigma |\sum_{i=1}^n \epsilon_i| \right) + 2\mathbb{E}_\epsilon \exp\left(tc\rho_\sigma Z_{j-1} \right) \\
&\leq 2^2 \exp\left(\frac{t^2 c^2 \rho_\sigma^2 n}{2} \right) + 2 \left[\sum_{i=1}^{k-1} 2^{j-i+1} \exp\left(\frac{n(c\rho_\sigma t)^2 (c\rho_\sigma)^{2(j-i)}}{2} \right) + \right.
\end{aligned}
$$

$$2^{j-1}\exp\left(\frac{n(c\rho_{\sigma}t)^2(c\rho_{\sigma})^{2(j-1)}}{2}+(tc\rho_{\sigma})(c\rho_{\sigma})^{j-1}\sqrt{2n\log(2m_1)}\right)\Bigg] \tag{S2.5c}$$

The step in equation (S2.5a) follows from equation (S2.2). The step in equation (S2.5b) follows from Lemma 1. The step in equation (S2.5c) follows from equation S2.3 and the induction hypothesis.

The second step is by Jensen's inequality. For any $\lambda > 0$,

$$\begin{aligned}
n\widehat{\mathfrak{R}}_S(\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}) &= \mathbb{E}_{\epsilon}\left[\sup_{f\in\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}\left(\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right)\right] \\
&\leq \frac{1}{\lambda}\log\mathbb{E}_{\epsilon}\exp\left(\lambda\sup_{f\in\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}\left(\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x}_i)\right)\right) \\
&\leq \frac{1}{\lambda}\log\mathbb{E}_{\epsilon}\exp\left(\lambda o\sup_{f\in\mathcal{SN}^{k,\boldsymbol{d},\sigma}_{c,o}}\left\|\sum_{i=1}^{n}\epsilon_i(1,\sigma\circ f_j(\boldsymbol{x}_i))\right\|_{\infty}\right) \\
&\leq \frac{1}{\lambda}\log\mathbb{E}_{\epsilon}\left[\exp\left(\lambda o|\sum_{i=1}^{n}\epsilon_i|\right)+\exp\left(\lambda o Z_k\right)\right] \\
&\leq \frac{1}{\lambda}\log\left[2\exp(\frac{\lambda^2o^2n}{2})+\sum_{i=1}^{k}2^{k-i+2}\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2(k-i+1)}}{2}\right)+\right. \\
&\left.2^k\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2k}\max_i\|\boldsymbol{x}_i\|_{\infty}^2}{2}+to(c\rho_{\sigma})^k\sqrt{2n\log(2m_1)}\right)\right], \quad \text{(S2.6a)}
\end{aligned}$$

where the step in equation (S2.6a) follows from the first main step. Especially, if $c\rho_{\sigma}\geq 1$,

$$\begin{aligned}
\text{(S2.6a)} &\leq \frac{1}{\lambda}\log\left[\sum_{i=1}^{k+1}2^{k-i+2}\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2k}}{2}\right)+2^k\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2k}}{2}+to(c\rho_{\sigma})^k\sqrt{2n\log(2m_1)}\right)\right] \\
&\leq \frac{1}{\lambda}\log\left[2^{k+2}\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2k}}{2}\right)+2^{k+2}\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2k}}{2}+to(c\rho_{\sigma})^k\sqrt{2n\log(2m_1)}\right)\right] \\
&\leq \frac{1}{\lambda}\log\left[2^{k+3}\exp\left(\frac{no^2\lambda^2(c\rho_{\sigma})^{2k}}{2}+\lambda o(c\rho_{\sigma})^k\sqrt{2n\log(2m_1)}\right)\right]
\end{aligned}$$

$$= \frac{(k+3)\log 2}{\lambda} + \frac{no^2\lambda(c\rho_{\sigma})^{2k}}{2} + o(c\rho_{\sigma})^k\sqrt{2n\log(2m_1)}$$

By choosing the optimal λ, we have

$$\widehat{\mathfrak{R}}_S(\mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}) \leq \frac{1}{\sqrt{n}}o(c\rho_{\sigma})^k(\sqrt{(k+3)\log 4} + \sqrt{2\log(2m_1)})$$

□

Proof of Theorem 2

We provide a general version of Theorem 2 with no assumption on the values of c or ρ_{σ}. Theorem 2 is the direct conclusion of the proposition below.

Proposition 1. *Assume A1-A3 hold. Fix $\delta \in (0,1)$, then with probability at least $1-\delta$ over the choice of the sample, for* **every** *sparse DNN $f_T \in \mathcal{S}_c^{k,\boldsymbol{d},\sigma} = T_{k+1}\circ\sigma\circ T_k\circ\cdots\circ\sigma\circ T_1$, we have*

$$\mathcal{E}_{L_S}(f_T) \leq \sqrt{\frac{\log(\frac{2}{\delta}) + 2\log(\|T_{k+1}\|_1 + 2)}{2n}} + \frac{2}{\sqrt{n}}(\|T_{k+1}\|_1 + 1)\left[\sqrt{(k+1)\log 16}\left(\sum_{\ell=0}^{k}(c\rho_{\sigma})^{\ell} + (c\rho_{\sigma})^k\right) + (c\rho_{\sigma})^k\sqrt{2\log(2m_1)}\right].$$

Furthermore, if $c\rho_{\sigma} \geq 1$, With probability at least $1-\delta$ over the choice of the sample, for **every** *sparse DNN $f_T \in \mathcal{S}_c^{k,\boldsymbol{d},\sigma} = T_{k+1}\circ\sigma\circ T_k\circ\cdots\circ\sigma\circ T_1$, we have*

$$\mathcal{E}_L(f_T) \leq \sqrt{\frac{\log(\frac{2}{\delta}) + 2\log(\|T_{k+1}\|_1 + 2)}{2n}} + \frac{2}{\sqrt{n}}(\|T_{k+1}\|_1 + 1)(c\rho_{\sigma})^k(\sqrt{(k+3)\log 4} + \sqrt{2\log(2m_1)}).$$

Proof. The proof is inspired by (Bartlett et al., 2017). Given a positive integer l, Define

a set

$$\mathcal{B}(l) = \mathcal{SN}_{c,l}^{k,\boldsymbol{d},\sigma}.$$

Correspondingly subdivide δ as

$$\delta(l) = \frac{\delta}{l(l+1)}.$$

Fix any l, we could get the corresponding generalization bounds as an instance of Lemma 4. By $A3$, for any $f \in \mathcal{SN}_{c,o}^{k,\boldsymbol{d},\sigma}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}$, we have

$$\left| \frac{\partial L(f(\boldsymbol{x}), y)}{\partial f(\boldsymbol{x})} \right| = 1 \tag{S2.7}$$

and

$$|L(f(\boldsymbol{x}), \boldsymbol{y})| \leq 1. \tag{S2.8}$$

Thus for the mean square error, we could replace ρ and A_0 in Lemma 4 with equations (S2.7) and (S2.8), respectively, and get the corresponding generalization bound.

As $\sum_{l \in \mathcal{N}_+} \delta(l) = \delta$, the preceding bound holds simultaneously for all functions in the union $\cup\{\mathcal{B}(l) : l \in \mathcal{N}_+\}$ with probability at least $1-\delta$. Thus given f_T, choose the smallest l such that $f_T \in \mathcal{B}(l)$. As $T_{k+1}(\mathbf{u}) = \tilde{V}_{k+1}^T(1, \mathbf{u}^T)^T$, then the smallest l satisfies that

$$l \leq \|T_{k+1}\|_1 + 1.$$

Further replace the l's with $\|T_{k+1}\|_1 + 1$, thus we get the desired result. □

Proof of Theorem 3

We provide a general version of Theorem 3 with no assumption on the values of c or ρ_σ. Theorem 3 is the direct conclusion of the proposition below.

Proposition 2. *Assume B1-B2 hold. Fix $\delta \in (0,1)$, $c > 0$, the number of hidden layers $k \in [0,\infty)$, and widths $\boldsymbol{d} \in \mathbb{N}_+^{k+2}$ with $d_0 = m_1$ and $d_{k+1} = 1$. With probability at least $1-\delta$ over the choice of the sample, for* **every** *sparse DNN $f_T \in \mathcal{S}_c^{k,\boldsymbol{d},\sigma} = T_{k+1} \circ \sigma \circ T_k \circ \cdots \circ \sigma \circ T_1$, we have*

$$\mathcal{E}_{L_C}(f_T) \leq \left(2(\|T_{k+1}\|_{1,\infty} + \frac{1}{m_2}) \max\left(1, (c\rho_\sigma)^k\right) + \log m_2\right) \sqrt{\frac{\log\sqrt{\frac{2}{\delta}} + \sum_{j=1}^{m_2} \log(m_2 \|T_{k+1}[j]\|_1 + 2)}{n}}$$
$$+ \frac{2\sqrt{2}}{\sqrt{n}} \left(\|T_{k+1}\|_{1,1} + 1\right) \left(1 + \frac{\sqrt{m_2 - 1}}{1 + (m_2 - 1)\exp(-2(\|T_{k+1}\|_{1,\infty} + \frac{1}{m_2}) \max\left(1, (c\rho_\sigma)^k\right))}\right) *$$
$$\left[\sqrt{(k+1)\log 16} \left(\sum_{\ell=0}^{k} (c\rho_\sigma)^\ell + (c\rho_\sigma)^k\right) + (c\rho_\sigma)^k \sqrt{2\log(2m_1)}\right].$$

Furthermore, if $c\rho_\sigma \geq 1$, With probability at least $1-\delta$ over the choice of the sample, for **every** *sparse DNN $f_T \in \mathcal{S}_c^{k,\boldsymbol{d},\sigma} = T_{k+1} \circ \sigma \circ T_k \circ \cdots \circ \sigma \circ T_1$, we have*

$$\mathcal{E}_{L_C}(f_T) \leq \left(2(\|T_{k+1}\|_{1,\infty} + \frac{1}{m_2})(c\rho_\sigma)^k + \log m_2\right) \sqrt{\frac{\log\sqrt{\frac{2}{\delta}} + \sum_{j=1}^{m_2} \log(m_2 \|T_{k+1}[j]\|_1 + 2)}{n}}$$
$$+ \frac{2\sqrt{2}}{\sqrt{n}} \left(\|T_{k+1}\|_{1,1} + 1\right) \left(1 + \frac{\sqrt{m_2 - 1}}{1 + (m_2 - 1)\exp\left(-2(\|T_{k+1}\|_{1,\infty} + \frac{1}{m_2})(c\rho_\sigma)^k\right)}\right) *$$
$$(c\rho_\sigma)^k (\sqrt{(k+3)\log 4} + \sqrt{2\log(2m_1)}).$$

Proof. The proof is inspired by (Bartlett et al., 2017). Given positive integers $\boldsymbol{l} =$

$(l_1, \cdots, l_{m_2})$, define a set

$$\mathcal{B}(\boldsymbol{l}) = \mathcal{SN}_{c,\boldsymbol{l}/m_2}^{k,\boldsymbol{d},\sigma}.$$

Correspondingly subdivide δ as

$$\delta(\boldsymbol{l}) = \frac{\delta}{l_1(l_1+1)\cdots l_{m_2}(l_{m_2}+1)}.$$

Fix any $\boldsymbol{l}$, we get the corresponding generalization bound as an instance of Lemma 5. Consider $f \in \mathcal{SN}_{c,\boldsymbol{O}}^{k,\boldsymbol{d},\sigma}, \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$. For $j^{'} \neq y$,

$$\left|\frac{\partial L_C(f(\boldsymbol{x}),y)}{\partial f(\boldsymbol{x})[j']}\right| \leq 1/(1+\sum_{j\neq j'} \exp\left(-(o_j+o_{j'})\max\left(1,(c\rho_{\sigma})^k\right)\right).$$

For y,

$$\begin{aligned}\left|\frac{\partial L_C(f(\boldsymbol{x}),y)}{\partial f(\boldsymbol{x})[y]}\right| &= \left|1-\frac{1}{1+\sum_{j\neq y}\exp\left(f(\boldsymbol{x})[j]-f(\boldsymbol{x})[y]\right)}\right| \\ &\leq \left|1-\frac{1}{1+\sum_{j\neq y}\exp\left((o_j+o_y)\max\left(1,(c\rho_{\sigma})^k\right)\right)}\right|.\end{aligned}$$

Additionally,

$$|L_C(f(\boldsymbol{x}),y)| \leq \max_{j'} \log\left(\sum_{j=1}^{m_2}\exp\{(o_j+o_j^{'})\max\left(1,(c\rho_{\sigma})^k\right)\}\right).$$

For simplicity, we assume $o_j \leq o_0$ for $j = 1, \cdots, m_2$, then

$$\left\|\frac{\partial L_C(f(\boldsymbol{x}),y)}{\partial f(\boldsymbol{x})}\right\|_2 \leq 1+\frac{\sqrt{m_2-1}}{1+(m_2-1)\exp(-2o_0\max\left(1,(c\rho_{\sigma})^k\right))} \tag{S2.9}$$

and

$$|L_C(f(\boldsymbol{x}),y)| \leq 2o_0\max\left(1,(c\rho_{\sigma})^k\right)+\log m_2. \tag{S2.10}$$

We could replace ρ and A_0 in Lemma 5 with equations (S2.9) and (S2.10), respectively, and get the corresponding generalization bound for $\mathcal{SN}_{c,\boldsymbol{o}}^{k,\boldsymbol{d},\sigma}$.

As $\sum_{\boldsymbol{l}\in\mathcal{N}_+^{m_2}} \delta(\boldsymbol{l}) = \delta$, the preceding bound holds simultaneously for all functions in the union $\cup\{\mathcal{B}(\boldsymbol{l}) : \boldsymbol{l} \in \mathcal{N}_+^{m_2}\}$ with probability at least $1-\delta$. Thus given f_T, choose the smallest $\boldsymbol{l}$ such that $f_T \in \mathcal{B}(\boldsymbol{l})$. As $T_{k+1}(\mathbf{u}) = \tilde{V}_{k+1}^T(1, \mathbf{u}^T)^T$, then the smallest $\boldsymbol{l}$ satisfies that

$$l_j \leq m_2 \left\|T_{k+1}[j]\right\|_1 + 1, \forall j.$$

Therefore

$$\sum_{j=1}^{m_2} \frac{l_j}{m_2} \leq \left\|T_{k+1}\right\|_{1,1} + 1, \quad \max_j l_j \leq m_2 \left\|T_{k+1}\right\|_{1,\infty} + 1.$$

Therefore we get the desired result. □

S3 ADDITIONAL EXPERIMENTS

We extend the classification experiment in Section 5.2.

Firstly, we examine the effect of the sample size on generalization. As shown in Table 1, when the sample size increases, the generalization error becomes smaller, while having the normalization constant c fixed.

	size=500	size=1000	size=1500	size=2000	size=2500
$c = \infty$	1.674/69.90	1.576/71.00	1.528/72.20	1.508/75.70	1.489/76.60
$c = 0.16$	0.441/87.03	0.343/88.10	0.258/89.30	0.208/91.50	0.199/92.74
$c = 0.13$	0.376/87.23	0.334/87.80	0.252/89.47	0.171/91.76	0.171/92.80
$c = 0.10$	0.324/87.78	0.280/87.60	0.223/90.30	0.176/90.70	0.169/90.80
$c = 0.07$	0.260/88.34	0.241/87.80	0.189/90.80	0.162/91.21	0.133/91.86
$c = 0.04$	0.134/89.57	0.112/89.94	0.102/91.31	0.084/91.72	0.073/92.24
$c = 0.01$	0.068/88.48	0.079/89.15	0.034/90.30	0.036/91.00	0.024/91.47

Table 1: Generalization error/test accuracy for the classification experiment with different values of c and sample sizes.

Secondly, we check the relationship between the depth of the neural network and the generalization error. The result is shown in Table 2. When c is relatively large, the generalization error increases, as the neural network grows deeper. On the contrary, when $c = 0.04, 0.01$, the generalization error might even decrease, as the depth increases. This might be caused by the shrinkage of the term (c^k).

	100-20-2	100-50-20-2	100-100-50-20-2
∞	1.535/70.30	1.674/69.90	1.710/69.10
$c = 0.50$	0.461/83.14	0.478/84.78	0.542/82.42
$c = 0.16$	0.351/84.67	0.441/87.03	0.456/84.41
$c = 0.13$	0.322/85.35	0.376/87.23	0.431/84.89
$c = 0.10$	0.312/86.10	0.324/87.78	0.383/86.03
$c = 0.07$	0.245/88.42	0.260/88.34	0.274/87.98
$c = 0.04$	0.103/89.12	0.134/89.57	0.131/88.33
$c = 0.01$	0.072/87.74	0.068/88.48	0.094/87.52

Table 2: Generalization error/test accuracy for the classification experiment in Section 5.2 with different network structures and sample sizes.

γ_0	0.050	0.045	0.040	0.035	0.030
Training error (%)	90.15	90.04	90.12	90.07	90.10

Table 3: Effect of the initial step size γ_0 on the algorithm.

Thirdly, we show that the projection gradient descent algorithm is not sensitive to the initial step size γ_0, as shown in Table 3.

Bibliography

Bartlett, P. L., D. J. Foster, and M. J. Telgarsky (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6241–6250.

Boucheron, S., G. Lugosi, and O. Bousquet (2003). Concentration inequalities. In *Summer School on Machine Learning*, pp. 208–240.

Golowich, N., A. Rakhlin, and O. Shamir (2018). Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference On Learning Theory*.

Kakade, S. M., K. Sridharan, and A. Tewari (2009). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, pp. 793–800.

Ledoux, M. and M. Talagrand (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of machine learning*. MIT press.