Supplementary Materials

A Analysis

Here, we present the complete proof of our main result. For reader's conveninence, we restate the theorem, lemmas and propositions.

We condition on the initialization $W(0)$ and the outer weight a. The expectation $\mathbb{E}[\cdot]$ is taken over the randomness of the samples drawn at iterations, unless specified otherwise.

Theorem 1. *Suppose the step size $\eta_t \le \frac{\theta}{t+1}$ with $\theta < \frac{1}{4}$. For any $T < \infty$, if*

$$m \ge c\left(d^2 + \max\left\{\left(\frac{(T+1)^{2\theta}}{\theta}\right)^9, \left(\frac{\theta\log(T)}{\delta}\right)^9\right\}\right)$$

for some universal constant $c > 0$, then with probability at least $1 - 2\exp(-2m^{1/3}) - \delta$,

$$\mathbb{E}\left[\|\Delta_t\|_2\right] \le \inf_{\ell}\left\{\prod_{k=0}^{t-1}(1-\eta_k\lambda_\ell)\|\Delta_0\|_2 + \mathcal{R}(\Delta_0,\ell)\right\} + 2c_1, \quad \forall 0 \le t \le T, \tag{1}$$

where $c_1 = \sigma_0\sqrt{\frac{e^{4\theta}\theta^2(2-4\theta)}{1-4\theta}}$.

A.1 Proof Overview

We prove (1) via induction over iteration t.

The base case $t = 0$ trivially holds as $\|\Delta_0\|_2 \le \|\Delta_0\|_2 + 2c_1$. Assume (1) holds for any $s \le t \le T$, we show $\mathbb{E}\left[\|W(s+1) - W(0)\|_{\mathrm{F}}\right]$ is small for any $s \le t$.

Lemma A.1. *For any $t \ge 0$,*

$$\mathbb{E}\left[\|W(t+1) - W(0)\|_{\mathrm{F}}\right] \le \sum_{s=0}^{t}\eta_s\left(\mathbb{E}\left[\|\Delta_s\|_2\right] + \tau\right). \tag{2}$$

Proof. By the SGD update,

$$W_j(t+1) - W_j(t) = \frac{\eta_t a_j}{\sqrt{m}}\left[f^*(X_t) + e_t - f\left(X_t; W(t)\right)\right]\mathbf{1}_{\{\langle W_j(t), X_t\rangle \ge 0\}}X_t, \tag{3}$$

where $X_t \in \mathbb{R}^d$ is the fresh sample drawn at iteration t and e_t is the random noise.

In view of (3), for any s,

$$\|W(s+1) - W(s)\|_{\mathrm{F}} = \frac{\eta_s}{\sqrt{m}}\left|\Delta_s(X_s) + e_s\right|\left\|D_s a X_s^\top\right\|_{\mathrm{F}}, \tag{4}$$

where $D_s \in \mathbb{R}^{m\times m}$ is a diagonal matrix with diagonal entries given by $\{\mathbf{1}_{\{\langle W_1(s), X_s\rangle \ge 0\}}, \cdots, \{\mathbf{1}_{\{\langle W_m(s), X_s\rangle \ge 0\}}\}$, $a \in \mathbb{R}^m$ is the outer weight, and $\Delta_s(X_s) \in \mathbb{R}$ is the prediction error at iteration s given input X_s.

Note that $D_s a X_s^\top$ is a rank-one matrix and thus $\|D_s a X_s^\top\|_{\mathrm{F}} = \|D_s a\|_2 \|X_s\|_2 \le \sqrt{m}$, where the last inequality holds since $\|D_s\|_2 \le 1$, $\|a\|_2 = \sqrt{m}$, and $\|X_s\|_2 = 1$. Thus, by triangle inequality,

$$\|W(t+1) - W(0)\|_{\mathrm{F}} \le \sum_{s=0}^{t}\|W(s+1) - W(s)\|_{\mathrm{F}} \le \sum_{s=0}^{t}\eta_s\left|\Delta_s(X_s) + e_s\right|.$$

Taking expectation on both hand sides, we have

$$\begin{aligned}\mathbb{E}\left[\|W(t+1)-W(0)\|_{\mathrm{F}}\right] &\leq \sum_{s=0}^{t}\eta_s\mathbb{E}\left[|\Delta_s(X_s)+e_s|\right]\\ &\overset{(a)}{\leq} \sum_{s=0}^{t}\eta_s\mathbb{E}\left[\sqrt{\mathbb{E}_{X_s,e_s}\left[(\Delta_s(X_s)+e_s)^2\right]}\right]\\ &\overset{(b)}{\leq} \sum_{s=0}^{t}\eta_s\left(\mathbb{E}\left[\|\Delta_s\|_2\right]+\tau\right) \end{aligned} \tag{5}$$

where (a) holds by Cauchy-Schwartz inequality; (b) holds by independence of X_s and e_s.

□

We now claim that for any $s \leq t$,

$$\mathbb{E}\left[\|\Delta_s\|_2\right] \leq \|\Delta_0\|_2 + 2c_1. \tag{6}$$

To see this, note for any $\varepsilon > 0$, $\mathcal{R}(\Delta_0, \ell) < \varepsilon$ for sufficiently large ℓ. Thus,

$$\mathbb{E}\left[\|\Delta_s\|_2\right] \leq \prod_{k=0}^{s-1}\left(1-\eta_k\lambda_\ell\right)\|\Delta_0\|_2 + \epsilon + 2c_1 \leq \|\Delta_0\|_2 + \epsilon + 2c_1.$$

Since ϵ can be arbitrarily small, (6) holds.

Pugging (6) into (2), when $\eta_s \leq \frac{\theta}{s+1}$, we get

$$\mathbb{E}\left[\|W(s+1)-W(0)\|_{\mathrm{F}}\right] \leq \left[\theta\left(\log(T)+1\right)\right]\left(\|\Delta_0\|_2+\tau+2c_1\right). \tag{7}$$

The induction is then completed by the following proposition.

Proposition A.2. *Suppose the conditions in Theorem 1 hold. If (7) holds for any $s \leq t \leq T-1$, then (1) holds for $t+1$ with probability at least $1-2\exp(-m^{1/3})-\delta$ over the initialization $W(0)$ and the outer weight a.*

In Section A.2, we present the proof of Proposition A.2 in details. As a brief overview, we first follow [Su and Yang, 2019] to derive a recursive relation of Δ_t. Afterwards, we recursively replace Δ_t and bound $\|\Delta_t\|_2$ by the sum of four terms. We then carefully analyze each of the four terms to complete the proof.

A.2 Proof of Proposition A.2

Following [Su and Yang, 2019], we first analyze how the prediction values evolve over iterations. Denote $A = \{j : a_j = 1\}$ and $B = \{j : a_j = -1\}$. By definition,

$$\begin{aligned}f(x;W(t+1)) - f(x;W(t)) &= \frac{1}{\sqrt{m}}\sum_{j\in A}\left[\sigma\left(\langle W_j(t+1),x\rangle\right)-\sigma\left(\langle W_j(t),x\rangle\right)\right]\\ &\quad-\frac{1}{\sqrt{m}}\sum_{j\in B}\left[\sigma\left(\langle W_j(t+1),x\rangle\right)-\sigma\left(\langle W_j(t),x\rangle\right)\right]. \end{aligned}\tag{8}$$

We now bound (8) from both above and below. By the SGD update,

$$W_j(t+1) - W_j(t) = \frac{\eta_t a_j}{\sqrt{m}}\left[f^*(X_t)+e_t-f\left(X_t;W(t)\right)\right]\mathbf{1}_{\{\langle W_j(t),X_t\rangle\geq 0\}}X_t, \tag{9}$$

where $X_t \in \mathbb{R}^d$ is the fresh sample drawn at iteration t and e_t is the random noise. Since $\mathbf{1}_{\{v\geq 0\}}(u-v) \leq \sigma(u)-\sigma(v) \leq \mathbf{1}_{\{u\geq 0\}}(u-v)$ for $u, v \in \mathbb{R}$, it follows that

$$\sigma\left(\langle W_j(t+1),x\rangle\right)-\sigma\left(\langle W_j(t),x\rangle\right) \leq \frac{\eta_t a_j}{\sqrt{m}}\left[f^*(X_t)+e_t-f\left(X_t;W(t)\right)\right]\langle X_t,x\rangle\mathbf{1}_{\{\langle W_j(0),X_t\rangle\geq 0\}}\mathbf{1}_{\{\langle W_j(t+1),x\rangle\geq 0\}}$$

$$\sigma(\langle W_j(t+1),x\rangle)-\sigma(\langle W_j(t),x\rangle) \geq \frac{\eta_t a_j}{\sqrt{m}}\left[f^*(X_t)+e_t-f\left(X_t;W(t)\right)\right]\langle X_t,x\rangle\mathbf{1}_{\{\langle W_j(t),X_t\rangle\geq 0\}}\mathbf{1}_{\{\langle W_j(t),x\rangle\geq 0\}}.$$

For notation simplicity, define the following functions:

$$\Phi_t^+(x,\widetilde{x}) = \frac{1}{m}\sum_{j\in A}\langle x,\widetilde{x}\rangle \mathbf{1}_{\{\langle W_j(t),\widetilde{x}\rangle\geq 0\}}\mathbf{1}_{\{\langle W_j(t),x\rangle\geq 0\}},$$
$$\Psi_t^+(x,\widetilde{x}) = \frac{1}{m}\sum_{j\in A}\langle x,\widetilde{x}\rangle \mathbf{1}_{\{\langle W_j(t),\widetilde{x}\rangle\geq 0\}}\mathbf{1}_{\{\langle W_j(t+1),x\rangle\geq 0\}}.$$

Similarly we define Φ_t^- and Ψ_t^- in terms of the summation over B. Then $H_t = \Phi_t^+ + \Phi_t^-$. Define $M_t = \Psi_t^- - \Phi_t^-$ and $L_t = \Psi_t^+ - \Phi_t^+$.

With the above notation, we obtain the following upper bound:

$$\begin{aligned}
&f(x;W(t+1)) - f(x;W(t))\\
&\leq \eta_t\Psi_t^+(x,X_t)\left(f^*\left(X_t\right)+e_t - f\left(X_t;W(t)\right)\right) + \eta_t\Phi_{t+1}^-(x,X_t)\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right]\\
&= \eta_t\left(\Psi_t^+(x,X_t)+\Phi_t^-(x,X_t)\right)\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right]\\
&= \eta_t\left[H_t(x,X_t)+L_t(x,X_t)\right]\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right]. \qquad (10)
\end{aligned}$$

Similarly, we can obtain a lower bound as

$$\begin{aligned}
f(x;W(t+1)) - f(x;W(t)) &\geq \eta_t\left(\Psi_t^-(x,X_t)+\Phi_t^+(x,X_t)\right)\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right]\\
&= \eta_t\left[H_t(x,X_t)+M_t(x,X_t)\right]\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right]. \qquad (11)
\end{aligned}$$

In view of (10) and (11), if M_t and L_t are small, then the evolution of the prediction values is mainly determined by the kernel function H_t. To capture this idea, define

$$\epsilon_t(x,x';W(t)) \triangleq f\left(x;W(t)\right) - f\left(x;W\left(t+1\right)\right) + \eta_t H_t(x,x')\left[f^*(x')+e_t - f(x';W(t))\right]. \qquad (12)$$

For simplicity, we use $\epsilon_t(x,x')$ to denote $\epsilon_t(x,x';W(t))$. Then from the definition of ϵ_t, we have that

$$f^*(x) - f(x;W(t+1)) = f^*(x) - f(x;W(t)) - \eta_t H_t(x,X_t)\left[f^*(X_t)+e_t - f(X_t;W(t))\right] + \epsilon_t(x,X_t). \qquad (13)$$

Moreover, by (10) and (11),

$$-\eta_t L_t(x,X_t)\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right] \leq \epsilon_t(x,X_t) \leq -\eta_t M_t(x,X_t)\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right]. \qquad (14)$$

Thus, we get

$$\Delta_{t+1}(x) = (\mathsf{I}-\eta_t\mathsf{H}_t)\circ\Delta_t(X_t) - v_t(x,X_t) + \epsilon_t(x,X_t), \qquad (15)$$

where

$$\begin{aligned}
v_t(x,X_t) &\equiv v_t\left(x,X_t;W(t)\right)\\
&\triangleq \eta_t H_t(x,X_t)\left[f^*(X_t)+e_t - f\left(X_t;W(t)\right)\right] - \eta_t\mathbb{E}_{X_t}\left[H_t(x,X_t)\left(f^*(X_t) - f\left(X_t;W(t)\right)\right)\right]
\end{aligned}$$

characterizes the deviation of the stochastic gradient from its expectation.

For notation simplicity, we define operators:

$$\mathsf{K}_t = \mathsf{I} - \eta_t\Phi, \quad \mathsf{Q}_t = \mathsf{I} - \eta_t\mathsf{H}_t, \quad \mathsf{D}_t = \mathsf{Q}_t - \mathsf{K}_t.$$

Note that $\|\mathsf{D}_t\|_2 = \|\mathsf{Q}_t - \mathsf{K}_t\|_2 \leq \eta_t\|\Phi - H_t\|_\infty$. Since H_t is positive semi-definite and $\|H_t\|_\infty \leq 1$, we get that $0\leq\gamma_j\leq 1$ for all j, where γ_i is the i-th largest eigenvalue of H_t. Therefore, as $0\leq\eta_t\leq 2$,

$$\|\mathsf{Q}_t\|_2 \leq \|\mathsf{Q}_t\|_\infty \leq \sup_{1\leq i<\infty}|1-\eta_t\gamma_i| \leq 1. \qquad (16)$$

Similarly, we can get that $\|\mathsf{K}_t\|_2 \leq 1$.

With the above notation, we can simplify (15) as

$$\Delta_{t+1} = \mathsf{Q}_t \circ \Delta_t - v_t + \epsilon_t. \tag{17}$$

It follows that

$$\Delta_{t+1} = \prod_{s=0}^{t} \mathsf{Q}_s \circ \Delta_0 - \sum_{r=0}^{t} \prod_{s=r+1}^{t} \mathsf{Q}_s \circ v_r + \sum_{r=0}^{t} \prod_{s=r+1}^{t} \mathsf{Q}_s \circ \epsilon_r. \tag{18}$$

Here Q_s is random due to the randomness of H_s. We want to decompose (18) into deterministic terms which involve K_s and the remaining part. Intuitively, we want to show the remaining part is small so the dynamic of the prediction error is mainly determined by K_s. Note $\mathsf{Q}_s = \mathsf{K}_s + \mathsf{D}_s$ by definition. For any t, by recursively replacing Q_s by $\mathsf{K}_s + \mathsf{D}_s$ from $s = 0$ to $s = t$, we get that $\prod_{s=0}^{t} \mathsf{Q}_s = \prod_{s=0}^{t} \mathsf{K}_s + \sum_{r=0}^{t} \prod_{i=r+1}^{t} \mathsf{Q}_i \mathsf{D}_r \prod_{j=0}^{r-1} \mathsf{K}_j$. Thus,

$$\Delta_{t+1} = \prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 + \sum_{r=0}^{t} \left(\prod_{i=r+1}^{t} \mathsf{Q}_i \mathsf{D}_r \prod_{j=0}^{r-1} \mathsf{K}_j \circ \Delta_0 \right) + \sum_{r=0}^{t} \left(\prod_{s=r+1}^{t} \mathsf{Q}_s \circ (\epsilon_r - v_r) \right).$$

Taking the L_2 norm over both hand sides and using the triangle inequality, we get

$$\begin{aligned} \|\Delta_{t+1}\|_2 &\le \left\| \prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 \right\|_2 + \sum_{r=0}^{t} \left\| \prod_{i=r+1}^{t} \mathsf{Q}_i \mathsf{D}_r \prod_{j=0}^{r-1} \mathsf{K}_j \circ \Delta_0 \right\|_2 + \left\| \sum_{r=0}^{t} \prod_{s=r+1}^{t} \mathsf{Q}_s \circ v_r \right\|_2 + \sum_{r=0}^{t} \left\| \prod_{s=r+1}^{t} \mathsf{Q}_s \circ \epsilon_r \right\|_2 \\ &\le \left\| \prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 \right\|_2 + \sum_{r=0}^{t} \|\mathsf{D}_r\|_2 \|\Delta_0\|_2 + \left\| \sum_{r=0}^{t} \prod_{s=r+1}^{t} \mathsf{Q}_s \circ v_r \right\|_2 + \sum_{r=0}^{t} \|\epsilon_r\|_2, \end{aligned} \tag{19}$$

where the last inequality holds due to $\|\mathsf{Q}_s\|_2 \le 1$ and $\|\mathsf{K}_s\|_2 \le 1$.

Note that the first term in (19) does not depend on the sample drawn in SGD. The second term corresponds to the approximation error of using K_s instead of Q_s. The third term measures the accumulation of the noise brought by stochastic gradient descent. The last term measures the accumulation of the approximation error of using kernel functions H_t shown in (13).

We will analyze (19) term by term, and then combine them to prove Proposition A.2.

First term: Recall $\lambda_1 \ge \lambda_2 \cdots$ are the eigenvalues of Φ with corresponding eigenfunction ϕ_i and $\mathcal{R}(g, \ell) = \sum_{i \ge \ell+1} \langle g, \phi_i \rangle^2$ is the L_2 norm of the projection of function g onto the space spanned by the $l+1, l+2, \cdots$ eigenfunctions of Φ.

The following lemma derives an upper bound of the first term of (19) via the eigendecomposition of Φ.

Lemma A.3. *Suppose $\eta_s \lambda_1 < 1$ for any $s \le t$, then,*

$$\left\| \prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 \right\|_2 \le \inf_r \left\{ \prod_{s=0}^{t} (1 - \eta_s \lambda_r) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) \right\}.$$

Proof. Fix any t. By the eigendecomposition of Φ, we know $\prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 = \sum_{i=1}^{\infty} \rho_i(t) \langle \Delta_0, \phi_i \rangle \phi_i$, where $\rho_i(t) \triangleq \prod_{s=0}^{t} (1 - \eta_s \lambda_i)$. Thus, for arbitrary $r \in \mathbb{N}$, we have

$$\begin{aligned} \left\| \prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 \right\|_2^2 &= \sum_{i=1}^{\infty} \rho_i^2(t) \langle \Delta_0, \phi_i \rangle^2 \\ &\overset{(a)}{\le} \sum_{i=1}^{r} \rho_r^2(t) \langle \Delta_0, \phi_i \rangle^2 + \sum_{i=r+1}^{\infty} \langle \Delta_0, \phi_i \rangle^2 \\ &\le \rho_r^2(t) \|\Delta_0\|_2^2 + \mathcal{R}^2(\Delta_0, r), \end{aligned}$$

where (a) holds by $\rho_i(t) \le 1$ and the fact that $\rho_i(t) \le \rho_r(t)$ for any t. The conclusion then follows.

$\square$

Second term: To bound the second term of (19), it remains to bound $\sum_{r=0}^{t} \|\mathsf{D}_r\|_2$. Note that $\|\mathsf{D}_r\|_2 = \|\mathsf{Q}_r - \mathsf{K}_r\|_2 \le \eta_r \|H_r - \Phi\|_\infty$. Lemma A.4 and Lemma A.5 below together provide an upper bound of $\|H_r - \Phi\|_\infty$ under event $\Omega_1 \cap \Omega_2$, where

$$\Omega_1 = \left\{ \sup_{x,R} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{|\langle W_i(0), x\rangle| \le R\}} - \mathbb{E}_{w\sim N(0,I_d)} \left[\mathbf{1}_{\{|\langle w, x\rangle| \le R\}} \right] \right| \le \frac{1}{m^{1/3}} + C_2 \sqrt{\frac{d}{m}} \right\}$$

and

$$\Omega_2 = \left\{ \sup_{x,\widetilde{x}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{\langle W_i(0), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(0), \widetilde{x}\rangle \ge 0\}} - \mathbb{E}_{w\sim N(0,I_d)} \left[\mathbf{1}_{\{\langle w, x\rangle \ge 0\}} \mathbf{1}_{\{\langle w, \widetilde{x}\rangle \ge 0\}}) \right] \right| \le \frac{1}{m^{1/3}} + C_3 \sqrt{\frac{d}{m}} \right\}.$$

for some universal constants C_2 and C_3.

Both events are defined with respect to the initial randomness $W(0)$, and require the sample mean of some function of $W_i(0)$ to be close to the expectation. Since $W_i(0)$'s are *i.i.d.* Gaussian, using uniform concentration inequalities, we will show later in Lemma A.9 that both Ω_1 and Ω_2 occur with high probability when m is large.

Denote

$$O_t(x) = \{i : \operatorname{sgn}(\langle W_i(t), x\rangle) \ne \operatorname{sgn}(\langle W_i(0), x\rangle)\}$$

as the set of neurons that have sign flips at iteration t when the input data is x. Denote $S_t(x)$ as the cardinality of $O_t(x)$.

Lemma A.4. *Under Ω_2, for any $t \ge 0$,*

$$\|H_t - \Phi\|_\infty \le \frac{2}{m} \|S_t\|_\infty + C_3 \sqrt{\frac{d}{m}} + \frac{1}{m^{1/3}}.$$

Proof. We first show $\|H_t - H_0\|_\infty \le \frac{2}{m} \|S_t\|_\infty$ and then show $\|H_0 - \Phi\|_\infty \le \frac{1}{m^{1/3}} + C_3\sqrt{\frac{d}{m}}$. The conclusion follows by the triangle inequality.

To see $\|H_t - H_0\|_\infty \le \frac{2}{m} \|S_t\|_\infty$, note

$$\begin{aligned}
|H_t(x,\widetilde{x}) - H_0(x,\widetilde{x})| &= \left| \langle x, \widetilde{x}\rangle \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{1}_{\{\langle W_i(t), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(t), \widetilde{x}\rangle \ge 0\}} - \mathbf{1}_{\{\langle W_i(0), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(0), \widetilde{x}\rangle \ge 0\}} \right) \right| \\
&\le \frac{1}{m} \sum_{i=1}^{m} \left| \mathbf{1}_{\{\langle W_i(t), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(t), \widetilde{x}\rangle \ge 0\}} - \mathbf{1}_{\{\langle W_i(0), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(0), \widetilde{x}\rangle \ge 0\}} \right| \\
&\le \frac{1}{m} \sum_{i=1}^{m} \left| \mathbf{1}_{\{\langle W_i(t), \widetilde{x}\rangle \ge 0\}} - \mathbf{1}_{\{\langle W_i(0), \widetilde{x}\rangle \ge 0\}} \right| + \frac{1}{m} \sum_{i=1}^{m} \left| \mathbf{1}_{\{\langle W_i(t), x\rangle \ge 0\}} - \mathbf{1}_{\{\langle W_i(0), x\rangle \ge 0\}} \right| \\
&\le \frac{1}{m} \left(S_t(x) + S_t(\widetilde{x}) \right).
\end{aligned}$$

The conclusion follows by taking the supremum over x and $\widetilde{x}$ on both hand sides.

To see $\|H_0 - \Phi\|_\infty \le \frac{1}{m^{1/3}} + C_3\sqrt{\frac{d}{m}}$, note

$$\begin{aligned}
|H_0(x,\widetilde{x}) - \Phi(x,\widetilde{x})| &= \left| \langle x, \widetilde{x}\rangle \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{\langle W_i(0), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(0), \widetilde{x}\rangle \ge 0\}} - \mathbb{E}_{w\sim N(0,I_d)} \left[\mathbf{1}_{\{\langle w, x\rangle \ge 0\}} \mathbf{1}_{\{\langle w, \widetilde{x}\rangle \ge 0\}} \right] \right) \right| \\
&\le \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{\langle W_i(0), x\rangle \ge 0\}} \mathbf{1}_{\{\langle W_i(0), \widetilde{x}\rangle \ge 0\}} - \mathbb{E}_{w\sim N(0,I_d)} \left[\mathbf{1}_{\{\langle w, x\rangle \ge 0\}} \mathbf{1}_{\{\langle w, \widetilde{x}\rangle \ge 0\}} \right] \right|,
\end{aligned}$$

which completes the proof by taking the supremum of $(x, \widetilde{x})$ and invoking the definition of Ω_2. □

The next lemma further shows that when $\|W(t) - W(0)\|_{\mathrm{F}}$ is small and m is large, $\frac{1}{m} \|S_t\|_\infty$ is small under Ω_1.

Lemma A.5. *Under* Ω_1,

$$\frac{1}{m}\left\|S_t\right\|_\infty \leq \frac{1}{m^{1/3}} + C_2\sqrt{\frac{d}{m}} + \frac{2^{\frac{4}{3}}\|W(t)-W(0)\|_{\mathrm{F}}^{\frac{2}{3}}}{m^{1/3}\pi^{1/3}}.$$

Proof. Fix any R and input x. Denote $B_R(x) = \{i : |\langle W_i(0), x\rangle| \leq R\}$. Then $S_t(x) \leq |B_R(x)| + |O_t(x) \cap B_R^c(x)|$. If neuron $i \in O_t(x) \cap B_R^c(x)$, then $|\langle W_i(t), x\rangle - \langle W_i(0), x\rangle| > R$. Thus, $\|W(t)-W(0)\|_{\mathrm{F}}^2 \geq R^2\,|O_t(x) \cap B_R^c(x)|$. Under Ω_1, we have

$$\sup_x |B_R(x)| \leq m^{2/3} + C_2\sqrt{md} + m\mathbb{E}_{w\sim N(0,I_d)}\left[\mathbf{1}_{\{|\langle w,x\rangle|\leq R\}}\right] \leq m^{2/3} + C_2\sqrt{md} + \frac{2mR}{\sqrt{2\pi}}.$$

Thus, we get

$$\|S_t\|_\infty \leq m^{2/3} + C_2\sqrt{md} + \frac{2mR}{\sqrt{2\pi}} + \frac{\|W(t)-W(0)\|_{\mathrm{F}}^2}{R^2}.$$

Optimally choosing R to be $\left(\frac{\sqrt{2\pi}\|W(t)-W(0)\|_{\mathrm{F}}^2}{2m}\right)^{1/3}$, we get that

$$\begin{aligned}
\|S_t\|_\infty &\leq m^{2/3} + C_2\sqrt{md} + \frac{4m}{\sqrt{2\pi}}\left(\frac{\sqrt{2\pi}}{2m}\|W(t)-W(0)\|_{\mathrm{F}}^2\right)^{1/3} \\
&= m^{2/3} + C_2\sqrt{md} + \frac{2^{4/3}m^{2/3}\|W(t)-W(0)\|_{\mathrm{F}}^{2/3}}{\pi^{1/3}}.
\end{aligned}$$

The conclusion follows by dividing both hand sides by m.

□

Third term: Next we derive an upper bound of the third term of (19). Recall $\sigma_t^2 = \mathbb{E}\left[\|\Delta_t\|_2^2\right] + \tau^2$.

Lemma A.6. *Suppose* $0 \leq \eta_s \leq 2$ *for any* $s \geq 0$, *then,*

$$\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{i=s+1}^{t}\mathsf{Q}_i \circ v_s\right\|_2\right] \leq \sqrt{\sum_{s=0}^{t}\eta_s^2\sigma_s^2}.$$

Proof. Denote F_t as the filtration of $\{X_1, \cdots, X_t\}$. Let $q_t = \sum_{r=0}^t \prod_{i=r+1}^t \mathsf{Q}_i \circ v_r$ and $h_t = \mathsf{Q}_t \circ q_{t-1}$. Thus, $q_t = v_t + h_t$. Then

$$\mathbb{E}\left[\|q_t\|_2^2\right] = \mathbb{E}\left[\|v_t + h_t\|_2^2\right] \overset{(a)}{=} \mathbb{E}\left[\|v_t\|_2^2\right] + \mathbb{E}\left[\|h_t\|_2^2\right] \overset{(b)}{\leq} \mathbb{E}\left[\|v_t\|_2^2\right] + \mathbb{E}\left[\|q_{t-1}\|_2^2\right]$$

where (a) uses the fact that $\mathbb{E}\left[\langle v_t, h\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle v_t, h\rangle | F_{t-1}\right]\right] = \mathbb{E}\left[\langle \mathbb{E}\left[v_t | F_{t-1}\right], h\rangle\right] = 0$; (b) follows from (16). Recursively applying the last displayed equation yields that $\mathbb{E}\left[\|q_t\|_2^2\right] \leq \sum_{r=0}^t \mathbb{E}\left[\|v_r\|_2^2\right]$.

Furthermore, note that

$$\begin{aligned}
&\mathbb{E}\left[v_t^2(x, X_t; W_t)\right] \\
&= \eta_t^2\mathbb{E}\left[\left(H_t(x, X_t)\left(\Delta_t(X_t) + e_t\right) - \mathbb{E}_{X_t}\left[H_t(x, X_t)\Delta_t(X_t)\right]\right)^2\right] \\
&= \eta_t^2\mathbb{E}_{F_{t-1}}\left[\mathbb{E}_{X_t,e_t}\left[H_t^2\left(x, X_t\right)\left(\Delta_t\left(X_t\right) + e_t\right)^2 | F_{t-1}\right] - \eta_t^2\left\{\mathbb{E}_{X_t}\left[H_t\left(x, X_t\right)\Delta_t\left(X_t\right)|F_{t-1}\right]\right\}^2\right] \\
&\leq \eta_t^2\mathbb{E}_{F_{t-1}}\left[\mathbb{E}_{X_t,e_t}\left[H_t^2(x, X_t)\left(\Delta_t(X_t) + e_t\right)^2 | F_{t-1}\right]\right] \\
&\leq \eta_t^2\left(\mathbb{E}\left[\|\Delta_t\|_2^2\right] + \tau^2\right) \\
&= \eta_t^2\sigma_t^2,
\end{aligned} \tag{20}$$

where the last inequality holds from $\|H_t\|_\infty \leq 1$ and independence of e_t and F_t. Therefore, $\mathbb{E}\left[\|v_t\|_2^2\right] \leq \eta_t^2\sigma_t^2$ for any $t \geq 0$. The conclusion follows by applying Cauchy-Schwartz inequality. □

Remark A.1. *One key technical challenge is how to control the accumulation of the noise v_t due to the stochasticity of the gradients. Unlike the conventional SGD analysis such as [Nemirovski et al., 2009], there is no deterministic upper bound on $\|v_t\|_2$. In the existing neural networks literature on SGD such as [Allen-Zhu et al., 2019], a vanishing step size with order $\Theta(\frac{1}{\log m})$ is used to ensure a small accumulation of the noise v_t, which is particularly undesirable in the overparameterized regime when m is large. In contrast, we utilize the fact that v_t is a sequence of martingale difference and carefully bound the accumulation of v_t in expectation in Lemma A.6 when $\eta_t = O(1/t)$.*

Next, we show an recursive formula of σ_t^2.

Lemma A.7. *For any $t \geq 0$,*

$$\sigma_{t+1}^2 \leq \prod_{s=0}^{t} (1+2\eta_s)^2 \sigma_0^2.$$

Proof. Recall from (17), $\Delta_{t+1} = \mathsf{Q}_t \circ \Delta_t - v_t + \epsilon_t$. Therefore,

$$\begin{aligned}\|\Delta_{t+1}\|_2^2 &= \|\mathsf{Q}_t \circ \Delta_t - v_t + \epsilon_t\|_2^2 \\ &= \|\mathsf{Q}_t \circ \Delta_t\|_2^2 + \|v_t\|_2^2 + \|\epsilon_t\|_2^2 - 2\langle \mathsf{Q}_t \circ \Delta_t, v_t\rangle - 2\langle v_t, \epsilon_t\rangle + 2\langle \mathsf{Q}_t \circ \Delta_t, \epsilon_t\rangle \\ &\leq \|\Delta_t\|_2^2 + \|v_t\|_2^2 + \|\epsilon_t\|_2^2 + 2\|\Delta_t\|_2 \|v_t\|_2 + 2\|v_t\|_2 \|\epsilon_t\|_2 + 2\|\Delta_t\|_2 \|\epsilon_t\|_2 . \end{aligned} \tag{21}$$

where the last inequality holds by $\|\mathsf{Q}_t\|_2 \leq 1$ and Cauchy-Schwartz inequality.

Note $\|L_t\|_\infty \leq 1$ and $\|M_t\|_\infty \leq 1$ for any t. Thus, by (14), $\|\epsilon_t\|_2^2 \leq \eta_t^2 (\Delta_t(X_t) + e_t)$ and hence

$$\mathbb{E}\left[\|\epsilon_t\|_2^2\right] \leq \eta_t^2 \left(\mathbb{E}\left[\|\Delta_t\|_2^2\right] + \tau^2\right) = \eta_t^2 \sigma_t^2. \tag{22}$$

Conditioning on the initialization $W(0)$, taking expectation over both hand sides of (21), adding τ^2 on both hand sides, and applying the upper bound of $\mathbb{E}\left[\|\epsilon_t\|_2^2\right]$ in (22) and $\mathbb{E}\left[\|v_t\|_2^2\right]$ in (20), we get

$$\begin{aligned}\sigma_{t+1}^2 &\leq \sigma_t^2 + \eta_t^2\sigma_t^2 + \eta_t^2\sigma_t^2 + 2\mathbb{E}\left[\|\Delta_t\|_2 \|v_t\|_2\right] + 2\mathbb{E}\left[\|v_t\|_2 \|\epsilon_t\|_2\right] + 2\mathbb{E}\left[\|\Delta_t\|_2 \|\epsilon_t\|_2\right] \\ &\leq \left(2\eta_t^2+1\right)\sigma_t^2 + 2\sqrt{\mathbb{E}\left[\|\Delta_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\|v_t\|_2^2\right]} + 2\sqrt{\mathbb{E}\left[\|v_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\|\epsilon_t\|_2^2\right]} + 2\sqrt{\mathbb{E}\left[\|\Delta_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\|\epsilon_t\|_2^2\right]} \\ &\leq \left(2\eta_t^2+1\right)\sigma_t^2 + 2\eta_t\sigma_t^2 + 2\eta_t^2\sigma_t^2 + 2\eta_t\sigma_t^2 \\ &= (1+2\eta_t)^2\sigma_t^2 \end{aligned}$$

where the second inequality holds by Cauchy-Schwartz inequality. $\square$

By Lemma A.7, we get

$$\begin{aligned}\eta_r\sigma_r &\leq \frac{\theta}{r+1}\prod_{k=0}^{r-1}\left(1+\frac{2\theta}{(k+1)}\right)\sigma_0 \\ &\leq \frac{\theta}{r+1}\exp\left(2\theta\left(\log(r+1)+1\right)\right)\sigma_0 \\ &\leq \theta(r+1)^{2\theta-1}e^{2\theta}\sigma_0. \end{aligned} \tag{23}$$

Plugging (23) into Lemma A.6, we get

$$\begin{aligned}
\sqrt{\sum_{r=0}^{t}\mathbb{E}\left[\|v_r\|_2^2\right]} &\leq \sqrt{\sum_{r=0}^{t}\eta_r^2\sigma_r^2} \\
&\leq \sqrt{\sum_{r=0}^{t}\frac{e^{4\theta}\theta^2\sigma_0^2}{(r+1)^2}\exp(4\theta\log(r+1))} \\
&\leq \sqrt{\sum_{r=0}^{t}\sigma_0^2 e^{4\theta}\theta^2(r+1)^{4\theta-2}} \\
&\leq \sqrt{\theta^2 e^{4\theta}\left(\frac{1}{1-4\theta}+1\right)\sigma_0^2} = c_1
\end{aligned} \tag{24}$$

where the last inequality holds since $\sum_{r=0}^{t}(r+1)^{4\theta-2} \leq \int_1^{t+1} x^{4\theta-2}dx + 1 \leq \frac{1}{4\theta-1}x^{4\theta-1}\Big|_1^{t+1} + 1 \leq \frac{1}{1-4\theta}+1$.

Fourth term: For the fourth term of (19), taking the L_2 norm and the conditional expectation of (14), by Cauchy-Schwartz inequality, we have

$$\mathbb{E}\left[\|\epsilon_r\|_2\right] \leq \eta_r\sigma_r\sqrt{\mathbb{E}\left[\|L_r\|_\infty^2 + \|M_r\|_\infty^2\right]}. \tag{25}$$

It remains to bound $\mathbb{E}\left[\|L_r\|_\infty^2\right]$ and $\mathbb{E}\left[\|M_r\|_\infty^2\right]$. Note

$$\begin{aligned}
\mathbb{E}\left[\|L_r\|_\infty^2\right] &= \mathbb{E}\left[\|L_r\|_\infty^2\,\mathbf{1}_{\left\{\|W(r+1)-W(0)\|_\mathrm{F}\leq m^{1/3},\,\|W(r)-W(0)\|_\mathrm{F}\leq m^{1/3}\right\}}\right] \\
&+ \mathbb{E}\left[\|L_r\|_\infty^2\,\mathbf{1}_{\left\{\|W(r+1)-W(0)\|_\mathrm{F}> m^{1/3}\text{ or }\|W(r)-W(0)\|_\mathrm{F}> m^{1/3}\right\}}\right] \\
&\leq \mathbb{E}\left[\|L_r\|_\infty^2\,\mathbf{1}_{\left\{\|W(r+1)-W(0)\|_\mathrm{F}\leq m^{1/3},\,\|W(r)-W(0)\|_\mathrm{F}\leq m^{1/3}\right\}}\right] \\
&+ \mathbb{P}\left[\|W(r+1)-W(0)\|_\mathrm{F} > m^{1/3}\text{ or }\|W(r)-W(0)\|_\mathrm{F} > m^{1/3}\right],
\end{aligned} \tag{26}$$

where the inequality holds by $\|L_r\|_\infty \leq 1$.

Through Lemma A.5 and the following Lemma A.8, we can upper bound the first component of (26) as

$$\mathbb{E}\left[\|L_r\|_\infty^2\,\mathbf{1}_{\left\{\|W(r+1)-W(0)\|_\mathrm{F}\leq m^{1/3},\,\|W(r)-W(0)\|_\mathrm{F}\leq m^{1/3}\right\}}\right] \leq \left[\frac{2}{m^{1/3}} + 2C_2\sqrt{\frac{d}{m}} + \frac{2^{10/3}}{\pi^{1/3}m^{1/9}}\right]^2. \tag{27}$$

Lemma A.8.

$$\|L_t\|_\infty \leq \frac{1}{m}\|S_t\|_\infty + \frac{1}{m}\|S_{t+1}\|_\infty,$$

$$\|M_t\|_\infty \leq \frac{1}{m}\|S_t\|_\infty + \frac{1}{m}\|S_{t+1}\|_\infty.$$

Proof. Fix x and $\widetilde{x}$, we have

$$\begin{aligned}
|L_t(x,\widetilde{x})| &= \frac{1}{m}\left|\langle x,\widetilde{x}\rangle \sum_{j\in A} \mathbf{1}_{\{\langle W_j(t),\widetilde{x}\rangle\geq 0\}}\left(\mathbf{1}_{\{\langle W_j(t+1),x\rangle\geq 0\}}-\mathbf{1}_{\{\langle W_j(t),x\rangle\geq 0\}}\right)\right| \\
&\leq \frac{1}{m}\sum_{j\in A}\left|\mathbf{1}_{\{\langle W_j(t),\widetilde{x}\rangle\geq 0\}}\left(\mathbf{1}_{\{\langle W_j(t+1),x\rangle\geq 0\}}-\mathbf{1}_{\{\langle W_j(t),x\rangle\geq 0\}}\right)\right| \\
&\leq \frac{1}{m}\sum_{j\in A}\left|\mathbf{1}_{\{\langle W_j(t+1),x\rangle\geq 0\}}-\mathbf{1}_{\{\langle W_j(t),x\rangle\geq 0\}}\right| \\
&\leq \frac{1}{m}\sum_{j\in A}\left|\mathbf{1}_{\{\langle W_j(t+1),x\rangle\geq 0\}}-\mathbf{1}_{\{\langle W_j(0),x\rangle\geq 0\}}\right| + \frac{1}{m}\sum_{j\in A}\left|\mathbf{1}_{\{\langle W_j(t),x\rangle\geq 0\}}-\mathbf{1}_{\{\langle W_j(0),x\rangle\geq 0\}}\right| \\
&\leq \frac{1}{m}\left(S_{t+1}(x)+S_t(x)\right).
\end{aligned}$$

Thus, by taking the supremum on both hand sides, we get the desired bound on $\|L_t\|_\infty$. The conclusion for $\|M_t\|_\infty$ follows analogously. $\square$

For the second component of (26), note by (7) and Markov's inequality, we have for $s\in\{r,r+1\}$

$$\begin{aligned}
&\mathbb{P}\left[\|W(s)-W(0)\|_{\mathrm{F}} > m^{1/3}\right] \\
&\leq \frac{\left(\|\Delta_0\|_2+\tau+2c_1\right)\theta\left(\log(s)+1\right)}{m^{1/3}}.
\end{aligned} \tag{28}$$

Combining (26), (27) and (28), we have

$$\mathbb{E}\left[\|L_r\|_\infty^2\right] \leq \left[\frac{2}{m^{1/3}}+2C_2\sqrt{\frac{d}{m}}+\frac{2^{10/3}}{\pi^{1/3}m^{1/9}}\right]^2 + \frac{2\left[\|\Delta_0\|_2+\tau+2c_1\right]\theta\left(\log(t+1)+1\right)}{m^{1/3}} \tag{29}$$

Denote $\Omega_3 = \left\{\|\Delta_0\|_2 \leq \frac{\sqrt{\|f^*\|_2+1}}{\delta}\right\}$ where $0<\delta<1$. Under Ω_3, we can further bound the RHS of (29) in terms of δ.

The upper bound for $\mathbb{E}\left[\|M_t\|_\infty^2\right]$ can be obtained analogously.

Plugging (29) and (23) into (25), we get

$$\begin{aligned}
\sum_{r=0}^{t}\mathbb{E}\left[\|\epsilon_r\|_2\right] &\leq \frac{2\sqrt{14}\sigma_0}{m^{1/9}}\sum_{r=0}^{t}\frac{\theta}{r+1}\prod_{k=0}^{r-1}\left(1+\frac{2\theta}{k+1}\right) \\
&\leq \frac{\sqrt{14}e^{2\theta}\left(t+2\right)^{2\theta}\sigma_0}{m^{1/9}}
\end{aligned} \tag{30}$$

for $m \geq \max\left\{\left[\left(\frac{\sqrt{\|f^*\|_2^2+1}}{\delta}+\tau+2c_1\right)\theta\left(\log(T)+1\right)\right]^9, 2^{14}C_2^3d^2\right\}$.

Combining Lemma A.3, Lemma A.4, Lemma A.5, (24) and (30), we get that conditioning on $W(0)$ and the outer weight a such that $\Omega_1\cap\Omega_2\cap\Omega_3$ holds,

$$\begin{aligned}
\mathbb{E}\left[\|\Delta_{t+1}\|_2\right] &\leq \inf_{\ell}\left\{\prod_{k=0}^{t}(1-\eta_k\lambda_\ell)\|\Delta_0\|_2+\mathcal{R}(\Delta_0,\ell)\right\} + \frac{\theta}{m^{1/3}}\left(\log(t+1)+1\right)\|\Delta_0\|_2 + \frac{\sqrt{14}e^{2\theta}\sigma_0}{m^{1/9}}(t+2)^{2\theta}+c_1 \\
&\leq \inf_{\ell}\left\{\prod_{k=0}^{t}(1-\eta_k\lambda_\ell)\|\Delta_0\|_2+\mathcal{R}(\Delta_0,\ell)\right\} + 2c_1
\end{aligned}$$

for $m \geq \max \left\{ 27\left(2C_2+C_3\right)^3 d^2, \left\{ 24\theta\left(\log(T)+1\right)\left(\frac{\sqrt{\|f^*\|_2^2+1}}{\delta}+\tau+2c_1\right)\right\}^{9/2}, \left[10\theta\left(\log T+1\right)\right]^3, 14^5\left[\frac{(T+1)^{2\theta}}{\theta}\right]^9 \right\}$.

Ω_1, Ω_2 and Ω_3 occur with high probability :

Lemma A.9.

$$\mathbb{P}\left[\Omega_1\right] \geq 1-\exp(-2m^{1/3}),$$
$$\mathbb{P}\left[\Omega_2\right] \geq 1-\exp(-2m^{1/3}).$$

Proof. We show the conclusion for Ω_2; the conclusion for Ω_1 follows analogously. Denote

$$\phi(w_1,\cdots,w_m)=\sup_{x,x'}\left|\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{\langle w_i,x\rangle\geq 0\}}\mathbf{1}_{\{\langle w_i,x'\rangle\geq 0\}}-\mathbb{E}_w\left[\mathbf{1}_{\{\langle w,x\rangle\geq 0\}}\mathbf{1}_{\{\langle w,x'\rangle\geq 0\}}\right]\right|.$$

By the triangle inequality, we have

$$|\phi(w_1,\cdots,w_{i-1},w_i,w_{i+1},w_m)-\phi(w_1,\cdots,w_{i-1},w_i',w_{i+1},\cdots,w_m)|\leq \frac{1}{m}.$$

Let $W_1,\ldots,W_m$ denote m *i.i.d.* $\mathcal{N}(0,\mathbf{I}_d)$. Thus, by McDiarmid's inequality, we get

$$\mathbb{P}\left[\phi(W_1,\cdots,W_m)\geq m^{-1/3}+\mathbb{E}\left[\phi(W_1,\cdots,W_m\right]\right]\leq \exp\left(-2m^{1/3}\right).$$

The proof is then completed by invoking the following claim

$$\mathbb{E}\left[\phi(W_1,\cdots,W_m)\right]\leq C_3\sqrt{\frac{d}{m}}.$$

To prove the claim, by Proposition B.2, it suffices to show the VC dimension of $\mathcal{F}_1$ is upper bounded by $11d$, where $\mathcal{F}_1=\left\{g_{x,x'}: g_{x,x'}(w)=\mathbf{1}_{\{\langle w,x\rangle\geq 0\}}\mathbf{1}_{\{\langle w,x'\rangle\geq 0\}}\right\}$.

To see $\mathrm{VC}(\mathcal{F}_1)\leq 11d$, we first show $\mathrm{VC}(\mathcal{F}_1)\leq 11\mathrm{VC}(\mathcal{G})$ where $\mathcal{G}=\left\{g_x: g_x(w)=\mathbf{1}_{\{\langle w,x\rangle\geq 0\}}\right\}$ and then show $\mathrm{VC}(\mathcal{G})=d$.

Now we show $\mathrm{VC}(\mathcal{F}_1)\leq 11\mathrm{VC}(\mathcal{G})$. For any class of Boolean functions $\mathcal{F}$ on $\mathbb{R}^d$, we define $\mathcal{C}_{\mathcal{F}}=\{D_f, f\in\mathcal{F}\}$ where $D_f=\left\{x: x\in\mathbb{R}^d, f(x)=1\right\}$.

We claim $\mathcal{C}_{\mathcal{F}_1}=\mathcal{C}_{\mathcal{G}}\sqcap\mathcal{C}_{\mathcal{G}}$ where $\sqcap_{i=1}^N\mathcal{C}_i=\left\{\cap_{j=1}^N C_j: C_j\in\mathcal{C}_j, 1\leq j\leq N\right\}$. To see this, note that for any $f\in\mathcal{F}_1$, we can find g_1 and g_2 in $\mathcal{G}$ such that $D_f=D_{g_1}\cap D_{g_2}$. In particular, if $f=\mathbf{1}_{\{\langle w,x_1\rangle\geq 0\}}\mathbf{1}_{\{\langle w,x_2\rangle\geq 0\}}$, then we can take $g_1=\mathbf{1}_{\{\langle w,x_1\rangle\geq 0\}}$ and $g_2=\mathbf{1}_{\{\langle w,x_2\rangle\geq 0\}}$. Similarly, for any $g_1,g_2\in\mathcal{G}$, $D_{g_1}\cap D_{g_2}=D_f$ for some $f\in\mathcal{F}_1$. Then by Proposition B.1,

$$\mathrm{VC}(\mathcal{F}_1)\leq 5\log(8)\mathrm{VC}(\mathcal{G})\leq 11\mathrm{VC}(\mathcal{G}). \tag{31}$$

Next, we show $\mathrm{VC}(\mathcal{G})=d$ following the idea of [Hajek and Raginsky, 2019, Proposition 7.1].

Choose $\{w_1,w_2,\cdots,w_d\}$ to be linearly independent vectors in $\mathbb{R}^d$. Fix an arbitrary binary valued vector $b\in\{\pm1\}^d$.

Consider the linear system $w_i^T x=b_i$ for $1\leq i\leq d$. Since $\{w_1,w_2,\cdots,w_d\}$ are linearly independent, we can always find $x_b=W^{-1}b$ where $W=[w_1,w_2,\cdots,w_d]^T$. Thus, $g_{x_b}(w_i)=\mathbf{1}_{\{b_i=1\}}$ for all i. This shows $\mathrm{VC}(\mathcal{G})\geq d$.

Now we show $\mathrm{VC}(\mathcal{G})<d+1$. Fix arbitrary $\{w_1,w_2,\cdots,w_{d+1}\}$. Suppose for any binary valued vector $b=\{\pm1\}^{d+1}$, $\exists$ x_b such that $g_{x_b}(w_i)=\mathbf{1}_{\{b_i=1\}}$ for all i. Define $V=\{(\langle w_1,x\rangle,\langle w_2,x\rangle,\cdots,\langle w_{d+1},x\rangle): x\in\mathbb{R}^d\}$ which is a linear subspace in $\mathbb{R}^{d+1}$. Since $x\in\mathbb{R}^d$, $\dim(V)\leq d$. Therefore, $\exists v\neq 0\in V^\perp$ s.t. for any $x\in\mathbb{R}^d$,

$$\sum_{i=1}^{d+1} v_i\langle w_i,x\rangle=0$$

where v_i is the i-th coordinate of v.

WLOG we can assume that $v_j < 0$ for some j. To see this, since $v \neq 0$, there must exist some $v_k \neq 0$. If $v_k \geq 0$ for all k, then we consider $-v_k$ for any k. Thus, we can always assume $v_j < 0$ for some j.

Let $b_k = \mathbf{1}_{\{v_k \geq 0\}} - \mathbf{1}_{\{v_k < 0\}}$ for all k. Denote $x_0 \in \mathbb{R}^d$ which solves $g_{x_0}(w_k) = \mathbf{1}_{\{b_k = 1\}}$ for all k. This implies

$$\mathbf{1}_{\{\langle w_k, x_0 \rangle \geq 0\}} = \mathbf{1}_{\{v_k \geq 0\}}$$

for any k.

Thus, $v_k \langle w_k, x_0 \rangle \geq 0$ for any k. However, $\sum_{i=1}^{d+1} v_i \langle w_i, x_0 \rangle = 0$ which implies

$$v_k \langle w_k, x_0 \rangle = 0$$

for any k.

Since $v_j < 0$, $\langle w_j, x_0 \rangle < 0$. This contradicts the fact that $v_k \langle w_k, x_0 \rangle = 0$ for any k. Thus, we conclude that $\mathrm{VC}(\mathcal{G}) < d + 1$.

□

Lemma A.10. *For any* $0 < \delta < 1$,

$$\mathbb{P}[\Omega_3] \geq 1 - \delta.$$

Proof. Recall that a_i's are *i.i.d.* Rademacher random variables. Thus,

$$\begin{aligned}
\mathbb{E}_{a,W(0)}\left[\|\Delta_0\|_2^2\right] &= \|f^*\|_2^2 - 2\mathbb{E}_{a,W(0)}\{\langle f^*, f\rangle\} + \mathbb{E}_{a,W(0)}\left[\|f\|_2^2\right] \\
&\overset{(a)}{=} \|f^*\|_2^2 + \mathbb{E}_{a,W(0)}\left[\|f\|_2^2\right] \\
&\overset{(b)}{=} \|f^*\|_2^2 + \mathbb{E}_{W(0),X}\left[\frac{1}{m}\sum_{i=1}^m \sigma^2(\langle W_i(0), X\rangle)\right] \\
&\overset{(c)}{\leq} \|f^*\|_2^2 + \mathbb{E}_{W(0),X}\left[\langle W_1(0), X\rangle^2\right] = \|f^*\|_2^2 + 1,
\end{aligned}$$

where (a) holds since $\mathbb{E}_a[f] \equiv 0$; (b) holds by $\mathbb{E}[a_i a_j] = 0$ for $i \neq j$; (c) holds due to $\sigma^2(x) \leq x^2$; and the last equality holds because $\langle W_1(0), X\rangle \sim \mathcal{N}(0, 1)$. The conclusion then follows by Markov's inequality and Cauchy-Schwartz inequality. □

B Auxiliary Results

B.1 VC dimension

Let $\mathcal{C}$ be a collection of subsets of $\mathbb{R}^d$. For any set A consisting of finite points in $\mathbb{R}^d$, we denote $\mathcal{C}_A = \{C \cap A : C \in \mathcal{C}\}$. We say $\mathcal{C}$ shatters A if $|\mathcal{C}_A| = 2^{|A|}$. Let $\mathcal{M}_{\mathcal{C}}(n) = \max\left\{|\mathcal{C}_F| : F \subset \mathbb{R}^d, |F| = n\right\}$ and $\mathcal{S}(\mathcal{C}) = \sup\{n : \mathcal{M}_{\mathcal{C}}(n) = 2^n\}$ which is the largest cardinality of a set that can be shattered by $\mathcal{C}$.

Consider a class of Boolean functions $\mathcal{F}$ on $\mathbb{R}^d$. For each $f \in \mathcal{F}$, we denote $D_f = \left\{x : x \in \mathbb{R}^d, f(x) = 1\right\}$. As a result, the collection $\mathcal{C}_{\mathcal{F}} \triangleq \{D_f, f \in \mathcal{F}\}$ forms a collection of subsets of $\mathbb{R}^d$. The VC dimension of $\mathcal{F}$ is defined as $\mathrm{VC}(\mathcal{F}) \triangleq \mathcal{S}(\mathcal{C}_{\mathcal{F}})$.

We now present the propositions that are used in Lemma A.9.

Proposition B.1. *[Van Der Vaart and Wellner, 2009, Theorem 1.1]*

$$\mathcal{S}(\sqcap_{i=1}^N \mathcal{C}_i) \leq \frac{5}{2}\log(4N)\sum_{i=1}^N \mathcal{S}(\mathcal{C}_i),$$

where $\sqcap_{i=1}^N \mathcal{C}_i = \left\{\cap_{j=1}^N C_j : C_j \in \mathcal{C}_j, 1 \leq j \leq N\right\}$.

Proposition B.1 is used to bound the VC dimension of the function class of the product of two Boolean functions. Another application of VC dimension used in Lemma A.9 is the following proposition.

Proposition B.2. *[Vershynin, 2019, Theorem 8.3.23] Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension $VC(\mathcal{F}) \geq 1$. Let $X_1, X_2, \cdots, X_n$ be independent random points in Ω. Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_X\left[f(X)\right]\right|\right] \leq C\sqrt{\frac{VC(\mathcal{F})}{n}}$$

for some constant C.

B.2 Eigen-decomposition of Φ when data is uniform on $\mathbb{S}^{d-1}$

Here, we present a way to compute the eigenvalues λ_ℓ and the projection $\mathcal{R}(f^*, \ell)$ in Corollary 1 and Corollary 2. Both can be viewed as the applications of the following Theorem 2.

Define the space of homogeneous harmonic polynomials of order ℓ on the sphere as

$$H_\ell = \left\{P : \mathbb{S}^{d-1} \to \mathbb{R} : P(x) = \sum_{|\alpha|=\ell} c_\alpha x^\alpha,\ \Delta P = 0\right\}$$

where $x^\alpha = x_1^{\alpha_1}\cdots x_d^{\alpha_d}$, $|\alpha| = \sum_{i=1}^d \alpha_i$, $c_\alpha \in \mathbb{R}$ and $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator.

Denote for all $\ell \geq 0$, $\{Y_{\ell,i}\}_{i=1}^{N_\ell}$ as some orthonormal basis of H_ℓ where N_ℓ is the dimension of H_ℓ, i.e., $\langle Y_{\ell,i}, Y_{\ell,j}\rangle = 0$ for $i \neq j$. Moreover, from [Dai and Xu, 2013, Theorem 1.1.2] for $\ell \neq \ell'$, H_ℓ and $H_{\ell'}$ are orthogonal. Hence, $\{Y_{\ell,i}\}$ are orthogonal across different ℓ as well.

We now derive in Theorem 2 an expansion for functions with the form $\mathcal{K}(x, y) = h(\langle x, y\rangle),\ x, y \in \mathbb{S}^{d-1}, d \geq 3$ in terms of $\{Y_{\ell,i}\}, 1 \leq i \leq N_\ell, \ell \geq 0$. A similar result is obtained in [Su and Yang, 2019] without a full proof. We provide a proof here for completeness.

Theorem 2. *Suppose the function $\mathcal{K}$ has the form $\mathcal{K}(x, y) = h(\langle x, y\rangle)$ where h is analytic on $[-1, 1]$, $x, y \in \mathbb{S}^{d-1}$ and $d \geq 3$. Then*

$$\mathcal{K}(x, y) = \sum_{\ell\geq 0}\beta_\ell(h)\sum_{i=1}^{N_\ell} Y_{\ell,i}(x)Y_{\ell,i}(y)$$

where

$$\beta_\ell(h) = \frac{d-2}{2}\sum_{m=0}^{\infty}\frac{h_{\ell+2m}}{2^{\ell+2m}m!(\frac{d-2}{2})_{\ell+m+1}} \tag{32}$$

with $h_{\ell+2m}$ is the $(\ell + 2m)$-th derivative of h at 0 and $(\cdot)_n$ is the Pochhammer symbol recursively defined as $(a)_0 = 1$, $(a)_k = (a + k - 1)(a)_{k-1}$ for $k \geq 1$.

Remark B.1. *The case $d = 2$ can be analyzed using Fourier analysis. Since this is not of particular interest in our study, we do not provide the analysis here. One can refer to [Dai and Xu, 2013, Section 1.6] if interested.*

Before presenting the proof of Theorem 2, we first show a key result that will be used in the proof of Theorem 2.

Proposition B.3. *[Cantero and Iserles, 2012, Theorem 2, eq (2.1)] Let h be analytic in $[-1, 1]$. Letting $h_n = h^{(n)}(0)$ be n-th order derivative, then for any $\alpha > -1, \alpha \neq -\frac{1}{2}$,*

$$h(x) = \sum_{n=0}^{\infty}\widetilde{h}_n C_n^{\alpha+1/2}(x),\ x \in [-1, 1] \tag{33}$$

where

$$C_n^{\alpha+1/2}(x) = \frac{(2\alpha+1)_n}{n!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{(n+2\alpha+1)_k}{(\alpha+1)_k}\left(\frac{1-x}{2}\right)^k,$$

is the Gegenbauer polynomial, and

$$\widetilde{h}_n = (\alpha + n + 1/2)\sum_{m=0}^{\infty}\frac{h_{n+2m}}{2^{n+2m}m!(\alpha+1/2)_{n+m+1}}, \tag{34}$$

with $h_{n+2m} = h^{(n+2m)}(0)$*, the* $n+2m$*-th derivative of* h *at* 0*.*

Remark B.2. *Gegenbauer polynomials are orthogonal across different* n*, i.e., for* $m \neq n$*,* $d \geq 3$ *and any fixed* $y \in \mathbb{S}^{d-1}$*,* $\left\langle C_n^{\frac{d-2}{2}}(\langle \cdot, y\rangle), C_m^{\frac{d-2}{2}}(\langle \cdot, y\rangle)\right\rangle_{\mathbb{S}^{d-1}} = 0$*. The proof is based on the orthogonality of* H_ℓ*. One can check [Dai and Xu, 2013, Corollary 2.8] for a detailed proof.*

The form of $\beta_\ell(h)$ in (32) depends on the specific function h. Throughout this section, we abbreviate $\beta_\ell(h)$ as β_ℓ.

Now we proceed to the proof of Theorem 2.

Proof. From [Dai and Xu, 2013, eq(2.8)], we know for any $l \geq 0$,

$$\frac{\ell+\lambda}{\lambda} C_\ell^\lambda(\langle x, y\rangle) = \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y) \tag{35}$$

where $\lambda = \frac{d-2}{2}$, $x, y \in \mathbb{S}^{d-1}$.

Plug (35) in (33) and note that $\alpha + 1/2 = \lambda = \frac{d-2}{2}$, we get

$$h(\langle x, y\rangle) = \sum_{\ell \geq 0} \widetilde{h}_\ell \frac{\lambda}{\ell+\lambda} \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y) = \beta_\ell \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y)$$

where

$$\beta_\ell = \widetilde{h}_\ell \frac{\lambda}{\ell+\lambda} = \frac{d-2}{2} \sum_{m=0}^{\infty} \frac{h_{\ell+2m}}{2^{\ell+2m} m! (\frac{d-2}{2})_{\ell+m+1}}.$$

□

Theorem 2 directly implies the following corollary. Recall that the eigenvalues of Φ are denoted as $\{\lambda_i\}_{i=1}^\infty$ with $\lambda_1 \geq \lambda_2 \geq \cdots$.

Corollary 1. *Let* $\Phi(x, x') = h(\langle x, x'\rangle)$ *with* $h(u) = \frac{u}{2\pi}(\pi - \arccos(u)), u \in [-1, 1]$*. Then the eigenfunctions of* Φ *is* $\{Y_{\ell,i}\}, 1 \leq i \leq N_\ell, \ell \geq 0$ *with corresponding eigenvalues* β_ℓ *with the same form as (32) and multiplicity* N_ℓ *for each* ℓ*. More specifically,* $\lambda_1 = \beta_1$ *and* $\lambda_k = \beta_{2(k-2)}, k \geq 2$*.*

Proof. From the orthonormality of $\{Y_{\ell,i}\}$, it remains to show $\beta_{2k+1} = 0$ for any $k \geq 1$, $\beta_\ell \leq \beta_{\ell-2}$ for any $l \geq 2$, and $\beta_1 \geq \beta_0$.

Firstly, we derive a common form of h_{l+2m}. Note $h(0) = 0$. By induction, we can get

$$h^{(k)}(u) = \frac{1}{2}\mathbf{1}_{\{k=1\}} - \frac{1}{2\pi}\left[k \arccos^{(k-1)}(u) + u \arccos^{(k)}(u)\right] \tag{36}$$

for any $k \geq 1$.

Thus, $h_k = \frac{1}{2}\mathbf{1}_{\{k=1\}} - \frac{1}{2\pi} k \arccos^{(k-1)}(0)$.

Note $\arccos^{(2i-1)}(0) = -\left[(2i-3)!!\right]^2$ and $\arccos^{(2i)}(0) = 0$ for $i \geq 1$. Thus, we get $h_1 = \frac{1}{4}$, $h_{2i} = \frac{i}{\pi}\left[(2i-3)!!\right]^2$ and $h_{2i+1} = 0$ for all $i \geq 1$.

Plugging h_{2k+1} into (32), we get $\beta_{2k+1} = 0$ for any $k \geq 1$.

Now we show $\beta_k \geq \beta_{k+2}$ for any k. Fix any $d \geq 3$, from (32), we get

$$\begin{aligned}\beta_k &= \frac{d-2}{2} \sum_{m=0}^{\infty} \frac{h_{k+2m}}{2^{k+2m} m! (\frac{d-2}{2})_{k+m+1}} \\ &= \frac{d-2}{2} \frac{h_k}{2^k (\frac{d-2}{2})_{k+1}} + \frac{d-2}{2} \sum_{m=0}^{\infty} \frac{1}{m+1} \frac{h_{k+2+2m}}{2^{k+2+2m} (m)! (\frac{d-2}{2})_{k+2+m}}.\end{aligned}$$

Similarly,

$$\begin{aligned}\beta_{k+2} &= \frac{d-2}{2}\sum_{m=0}^{\infty}\frac{h_{k+2+2m}}{2^{k+2+2m}m!(\frac{d-2}{2})_{k+2+m+1}} \\ &= \frac{d-2}{2}\sum_{m=0}^{\infty}\frac{1}{\frac{d-2}{2}+k+m+2}\frac{h_{k+2+2m}}{2^{k+2+2m}m!(\frac{d-2}{2})_{k+2+m}}.\end{aligned}$$

For any term involving h_{k+2+2m}, the coefficient in β_k is large than the coefficient in β_{k+2}. Since h_{k+2+2m} are non-negative for any $m \geq 0$ and $h_k \geq 0$, we get $\beta_k \geq \beta_{k+2}$.

Lastly, we show $\beta_0 \leq \beta_1$. By (32) and (36), we get

$$\beta_1 = \frac{d-2}{2}\frac{h_1}{2(\frac{d-2}{2})_2} = \frac{1}{4d}, \tag{37}$$

and

$$\begin{aligned}\beta_0 &= \frac{d-2}{2}\sum_{m=0}^{\infty}\frac{h_{2m}}{4^m m!\left(\frac{d-2}{2}\right)_{m+1}} \\ &= \frac{d-2}{2\pi}\left[\frac{1}{4\left(\frac{d-2}{2}\right)_2}+\sum_{m\geq 2}\frac{\left((2m-3)!!\right)^2}{4^m(m-1)!\left(\frac{d-2}{2}\right)_{m+1}}\right] \\ &= \frac{1}{2\pi d}+\sum_{m\geq 2}a_m, \end{aligned}\tag{38}$$

where $a_m = \frac{d-2}{2\pi}\frac{[(2m-3)!!]^2}{4^m(m-1)!\left(\frac{d-2}{2}\right)_{m+1}}$ for $m \geq 2$.

Note for any $d \geq 3$ and $m \geq 2$,

$$\frac{a_{m+1}}{a_m} = \frac{(2m-1)^2}{4m(m+1+\frac{d-2}{2})} \leq \frac{m^2}{(m+1)^2}.$$

Thus,

$$\sum_{m\geq 2}a_m \leq 4a_2\left(\sum_{m\geq 2}\frac{1}{m^2}\right) \overset{(a)}{\leq} \frac{1}{\pi d(d+2)}\left(\frac{\pi^2}{6}-1\right) \tag{39}$$

where (a) holds by $a_2 = \frac{1}{4\pi d(d+2)}$.

Combining (37), (38) and (39), we get

$$\beta_1 - \beta_0 \geq \frac{1}{4d} - \left[\frac{1}{2\pi d}+\frac{1}{\pi d(d+2)}\left(\frac{\pi^2}{6}-1\right)\right] > 0.$$

□

With the eigendecomposition of Φ, we now compute the projection $\mathcal{R}(f, r)$.

Corollary 2. *Suppose the function f has the form $f(x) = h(\langle w, x\rangle)$ where $w \in \mathbb{S}^{d-1}$ is the parameter, then*

$$\mathcal{R}(f, r) = \sqrt{\sum_{k=r-1}^{\infty}\beta_{2k}^2\frac{2k+\lambda}{\lambda}C_{2k}^{\lambda}(1)}$$

where β_ℓ has the same form as (32) and $\lambda = \frac{d-2}{2}$.

Proof. Since $\{Y_{\ell,i}, 1 \leq i \leq N_\ell\}$ forms an orthonormal basis of H_ℓ, it follows from Theorem 2 that $\langle f, Y_{\ell,i}\rangle = \beta_\ell Y_{\ell,i}(w)$ which gives the orthogonal projection of $f(x)$ on H_ℓ as $\sum_{i=1}^{N_\ell} \beta_\ell Y_{\ell,i}(w) Y_{\ell,i}(x)$. Then by the definition of $\mathcal{R}(f,\ell)$ and the fact that $\beta_\ell = 0$ for $\ell = 2j+1, j \geq 1$, we have

$$\mathcal{R}(f,r) = \sqrt{\sum_{k=r-1}^{\infty} \beta_{2k}^2 \sum_{i=1}^{N_{2k}} Y_{2k,i}^2(w)}. \tag{40}$$

By (35), we get

$$\sum_{i=1}^{N_\ell} Y_{\ell,i}^2(w) = \frac{\ell+\lambda}{\lambda} C_\ell^\lambda(1).$$

Plug it back into (40), we get the desired conclusion.

□

C Additional numerical experiments

C.1 Simulations

We focus on two specific settings:

- Linear: $f^*(x) = \langle b, x\rangle$ with $b \sim N(0, I_d)$.
- Teacher neural network: $f^*(x) = \sum_{i=1}^{3} b_i \psi(\langle v_i, x\rangle)$, where $\psi(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, b_i's are $i.i.d.$ Rademacher random variables, and $v_i \sim N(0, I_d)$.

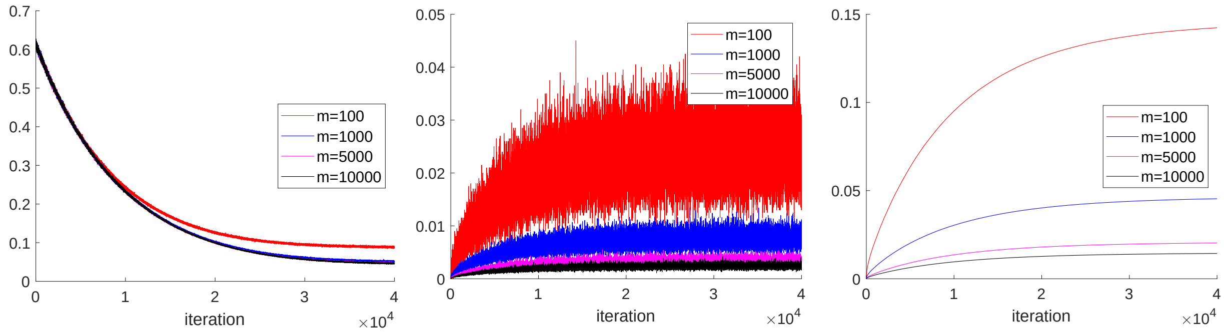
We run SGD on the streaming data with constant step size $\eta = 0.2$. We assume the symmetric initialization to ensure the initial prediction error $\Delta_0 = f^*$. At each iteration, we randomly draw data X uniformly from $\mathbb{S}^{d-1}$ to obtain (X, y) where $y = f^*(X)$. The average prediction error is estimated using freshly drawn 400 data points, and the resulting error is further averaged over 20 independent runs.

Figure 1 considers the setting with a varying number of hidden neurons m, when f^* is the teacher neural network and $d = 500$. Similar to the case with $d = 5$, Figure 1a shows that the averaged prediction error convergences faster when m increases from 100 to 1000, but there is not much difference when m is increased further. Again, this is consistent with our theory, because when m is large enough, the random kernel H_t is already well approximated by the Neural Tangent Kernel Φ. We also observe a small proportion of sign changes from figure 1b when m is above 1000, which leads to a small approximation error ϵ_t in view of Lemma A.8 and Lemma A.5. Figure 1c shows the relative deviation of the weight matrix at each iteration from the initialization. Following Lemma A.1, we see $\|W(t) - W(0)\|_{\mathrm{F}} = O(t)$ while $\|W(0)\|_{\mathrm{F}} = O(\sqrt{md})$. As a result, we see $\frac{\|W(t)-W(0)\|_{\mathrm{F}}}{\|W(0)\|_{\mathrm{F}}}$ decreases as m increases for fixed t and $\frac{\|W(t)-W(0)\|_{\mathrm{F}}}{\|W(0)\|_{\mathrm{F}}}$ increases as t grows for fixed m.

The same experiment is performed on the linear f^* and the results are shown in Figure 2 for $d = 5$ and Figure 3 for $d = 500$. We again see an increase in the convergence rate, a decrease in the number of sign changes, and a decrease in the relative deviation of the weight matrix from the initialization as m increases. In addition, we observe a smaller convergence rate when $d = 500$ compared to $d = 5$. This is due to the following reason. Compared to $d = 5$, when $d = 500$, λ_r is smaller and thus the contraction factor $\prod_{s=0}^{t}(1 - \eta_s \lambda_r)$ is larger, resulting in a slower convergence rate, as is shown in Corollary 1.

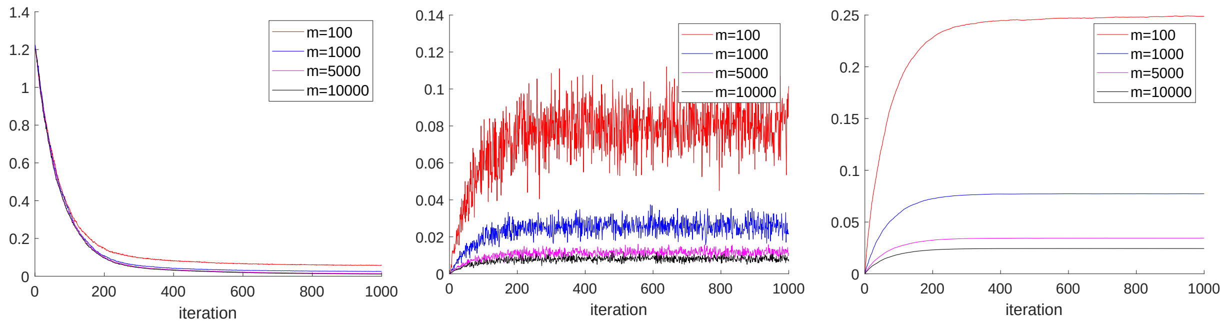
C.2 Real Data

We also run a numerical experiment on the MNIST dataset. We only use the classes of images 0 and 1 for simplicity. We treat the empirical distribution of 14780 images with 28×28 pixels as the underlying true data distribution. We reshape the data to have each $x_i \in \mathbb{R}^{784}$. For each $x_i \in \mathbb{R}^{784}$ in the dataset, we assign $y_i = 1$ if the corresponding image is 1 and $y_i = -1$ if the image is 0. We then normalize x_i to have $\|x_i\|_2 = 1$. We run the SGD on streaming data with step size $\eta = 0.02$ to learn the model. At each iteration, we randomly draw one x_i



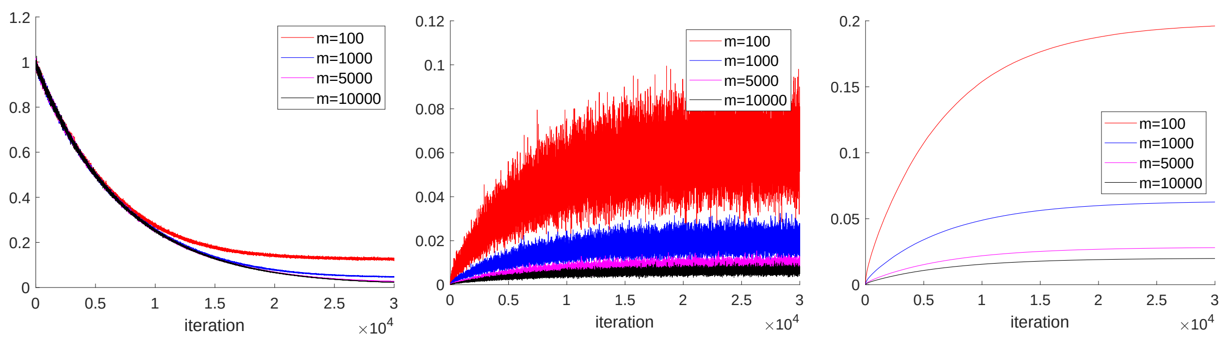
(a) averaged prediction error (b) $S_t(X_t)/m$ (c) $\|W(t) - W(0)\|_{\mathrm{F}}/\|W(0)\|_{\mathrm{F}}$

Figure 1: comparison of different number of neurons with teacher neural network f^* with $d = 500$



(a) averaged prediction error (b) $S_t(X_t)/m$ (c) $\|W(t) - W(0)\|_{\mathrm{F}}/\|W(0)\|_{\mathrm{F}}$

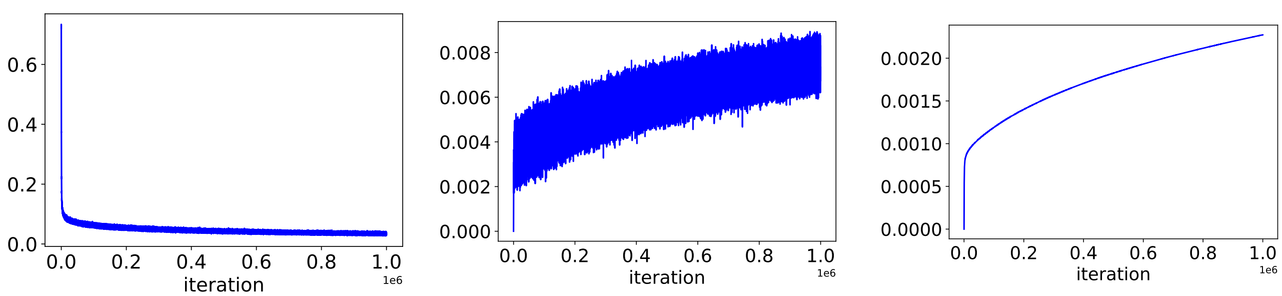
Figure 2: comparison of different number of neurons with linear f^* with $d = 5$



(a) averaged prediction error (b) $S_t(X_t)/m$ (c) $\|W(t) - W(0)\|_{\mathrm{F}}/\|W(0)\|_{\mathrm{F}}$

Figure 3: comparison of different number of neurons with linear f^* with $d = 500$

from the dataset to obtain (x_i, y_i). The averge prediction error is estimated using freshly drawn 200 data points, and the resulting error is further averaged over 20 independent runs. Figure 4 shows the result with $m = 10000$. Figure 4a shows that the overparametrized two-layer ReLU neural network under the one-pass SGD can learn f^* in the handwritten digit recognition scenario. Figure 4b and Figure 4c show a small proportion of sign changes and a small relative deviation of the weight matrix from the initialization.



(a) averaged prediction error

(b) $S_t(X_t)/m$

(c) $\|W(t) - W(0)\|_{\mathrm{F}}/\|W(0)\|_{\mathrm{F}}$

Figure 4: Results on the MNIST dataset with $m = 10000$

References

[Allen-Zhu et al., 2019] Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. 7

[Cantero and Iserles, 2012] Cantero, M. J. and Iserles, A. (2012). On rapid computation of expansions in ultra-spherical polynomials. *SIAM Journal on Numerical Analysis*, 50(1):307–327. 12

[Dai and Xu, 2013] Dai, F. and Xu, Y. (2013). *Approximation theory and harmonic analysis on spheres and balls*, volume 23. Springer. 12, 13

[Hajek and Raginsky, 2019] Hajek, B. and Raginsky, M. (2019). Statistical learning theory. *Lecture Notes*, 387. 10

[Nemirovski et al., 2009] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609. 7

[Su and Yang, 2019] Su, L. and Yang, P. (2019). On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in Neural Information Processing Systems*, pages 2641–2650. 2, 12

[Van Der Vaart and Wellner, 2009] Van Der Vaart, A. and Wellner, J. A. (2009). A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5:103. 11

[Vershynin, 2019] Vershynin, R. (2019). *High-dimensional probability.* Cambridge, UK: Cambridge University Press. 12