Generalization error bounds for two-layer neural networks with Lipschitz loss function

Jiang Yu Nguwi* Nicolas Privault†
Division of Mathematical Sciences
School of Physical and Mathematical Sciences
Nanyang Technological University
21 Nanyang Link, Singapore 637371

March 14, 2026

Abstract

We derive generalization error bounds for the training of two-layer neural networks without assuming boundedness of the loss function, using Wasserstein distance estimates on the discrepancy between a probability distribution and its associated empirical measure, together with moment bounds for the associated stochastic gradient method. In the case of independent test data, we obtain a dimension-free rate of order $O\big(n^{-1/2}\big)$ on the n-sample generalization error, whereas without independence assumption, we derive a bound of order $O\big(n^{-1/(d_{\rm in}+d_{\rm out})}\big)$, where $d_{\rm in}$, $d_{\rm out}$ denote input and output dimensions. Our bounds and their coefficients can be explicitly computed prior to the training of the model, and are confirmed by numerical simulations.

Keywords: Generalization error, neural networks, stochastic gradient method, Lipschitz bounds, concentration inequalities.
Mathematics Subject Classification (2020): 62M45, 68T07.

1 Introduction

The study of two-layer neural networks using stochastic gradient descent and their approximation guarantees has attracted considerable attention in recent years, see e.g. [MMM19] and references therein for a review using mean-field theory, [NDHR21] for the use of information-theoretic generalization bounds, or [PSE22] for covering arguments.

*nguw0003@e.ntu.edu.sg
†nprivault@ntu.edu.sg

The aim of the present paper is to propose generalization bounds for two-layer neural networks without assuming the boundedness of loss and activation functions, using bounds established in [FG15] on the Wasserstein between a probability distribution and its associated empirical measure.

Let

$$Z := (Z_i)_{1\leq i\leq n} = (X_i, Y_i)_{1\leq i\leq n} \overset{\text{i.i.d.}}{\sim} \rho$$

be a set of $n \geq 1$ independent data samples $(X_i, Y_i) \in \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}$, identically distributed according to a probability distribution ρ on $\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}$ which is not accessible in practice. Consider

$$l : \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{out}}} \to \mathbb{R}$$

a loss function, and a two-layer neural network function

$$f(\cdot, v, w) : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$$

parameterized by matrices $v \in \mathbb{R}^{d_{\text{in}}\times d}$, $w \in \mathbb{R}^{d\times d_{\text{out}}}$, $d, d_{\text{in}}, d_{\text{out}} \geq 1$.

In this context, given the output $(V(t), W(t)) \in \mathbb{R}^{d_{\text{in}}\times d} \times \mathbb{R}^{d\times d_{\text{out}}}$ of a Stochastic Gradient Method (SGM) trained on data sets $Z^{(t)} = \left(X_i^{(t)}, Y_i^{(t)}\right)_{1\leq i\leq n}$ of independent samples identically distributed according to ρ at epochs $t = 0, 1, \ldots, T$, we consider the generalization error defined as the difference

$$\begin{aligned}&\varepsilon_{\text{gen}}(n, V(T), W(T)) \\ &\quad := \int_{\mathbb{R}^{d_{\text{in}}}\times\mathbb{R}^{d_{\text{out}}}} l(f(x, V(T), W(T)), y)\rho(dx, dy) - \frac{1}{n}\sum_{i=1}^{n} l(f(X_i, V(T), W(T)), Y_i)\end{aligned} \tag{1.1}$$

between the expected loss of the neural network under the true data distribution $\rho(dx, dy)$ and its average loss on the training dataset $(X_i, Y_i)_{1\leq i\leq n}$, which measures how well the model generalizes to the unknown underlying distribution ρ.

In the case of a uniformly bounded loss function, a 0-1 error bound of order $O\left(n^{-1/2}\right)$ on the expected generalization error has been obtained in [CG19]. Related bounds have been derived in [HRS16], [RRT17], [MWZZ18], [PJL18], and in [ZZB+22] using a stability approach, by further assuming the boundedness of the gradients $\nabla_v l(f(x, v, w), y)$ and $\nabla_w l(f(x, v, w), y)$, or the β-smoothness property, or under a uniform boundedness assumption on the loss function [WLW+25]. See also [AZLL19] for the case of three-layer networks,

and [ACS23] for the derivation of $O\big(n^{-1}\big)$ error bounds using calculus on the space of measures.

In this paper, we aim at bounding $\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big|$ in a context where the loss function $l(y_1, y_2)$ may not be bounded, by relaxing the boundedness of $l(f(x, v, w), y)$ or its gradients using a Lipschitz condition which is satisfied by loss functions such as the mean absolute error or the Huber loss function. We also require a $\mathcal{C}^1$ Lipschitz condition on the activation function of the neural network, which is satisfied by e.g. the softplus, tanh, and sigmoid functions.

In Proposition 3.1, we start by deriving moment bounds of the SGM output $(V(T), W(T))$ for two-layer neural networks. Next, using a testing data set

$$Z = (Z_i)_{1\le i\le n} = (X_i, Y_i)_{1\le i\le n} \overset{\text{i.i.d.}}{\sim} \rho$$

independent of $\big(Z^{(t)}\big)_{0\le t\le T}$ and Proposition 3.1, we derive an error bound of dimension-free order $O\big(n^{-1/2}\big)$ on the L^1 norm $\mathbb{E}\big[\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big|\big]$, see Proposition 4.1, and related deviation inequalities in Proposition 4.2.

In Proposition 5.1, without the above independence assumption, we apply the Wasserstein distance bounds of [FG15] and Proposition 3.1 to derive generalization error bounds of order $O\big(n^{-1/(d_{\text{in}}+d_{\text{out}})}\big)$ on the expectation and deviation probability of $\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big|$. Related dimension-dependent phenomena have been observed in e.g. [FCAO18] when no boundedness is assumed on the loss function and its gradient. In Proposition 5.3 we also derive L^p bounds on the Lipschitz constant of regularized loss functions, with concentration inequalities presented in Corollary 5.4.

Unlike in e.g. [XR17], [LJ18], [ADH+19], [WDFC19], where the bounds rely on quantities that may not be available in practice, all constants appearing in our bounds can be explicitly computed without actually training the network. In contrast, the bounds derived by [NTS15], [NBS17], [DR17], [NLB+18], [CG19] rely on some properties of a trained network that are unknown before the training.

This paper is organized as follows. In Section 2, we introduce the SGM dynamics, its generalization error, and the Wasserstein distance bounds of [FG15]. In Section 3, we derive moment bounds for the SGM dynamics, see Proposition 3.1. In Section 4 we obtain L^1 error bounds in the case where the testing set is independent of the training data sequence used

for SGM updates, see Proposition 4.1 and the deviation inequalities of Proposition 4.2. Generalization error bounds without independence assumption are obtained in Proposition 5.1. This is followed by bounds and concentration inequalities for the Lipschitz constant of the loss function in Proposition 5.3 and Corollary 5.4. Numerical confirmations are presented in Section 6.

2 Preliminaries and notation

For $x = (x_1, \ldots, x_d)^\top$ a vector in $\mathbb{R}^d$ we use the Euclidean norm defined by $|x| = \sqrt{x_1^2 + \cdots + x_d^2}$. For $v \in \mathbb{R}^{m \times n}$ a matrix, we let

$$\|v\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |v_{i,j}|^2}$$

denote the matrix Frobenius norm of v, and we let $\mathrm{Supp}(\mu)$ denote the support of any probability measure μ.

Wasserstein distance

Let $p \geq 1$. For μ, ν in the space $\mathcal{P}_p\big(\mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}}\big)$ of probability measures on $\mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}}$ with finite p-moment, the Wasserstein-p distance between μ and ν is defined as

$$\mathcal{W}_p(\mu, \nu) := \inf_{\substack{\pi \text{ coupling} \\ \text{of } \mu \text{ and } \nu}} \left(\int_{(\mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}})^2} |z_1 - z_2|^p \pi(dz_1, dz_2) \right)^{1/p}.$$

Recall that by [KR58], for any $\mu, \nu \in \mathcal{P}_1\left(\mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}}\right)$ we have

$$\mathcal{W}_1(\mu, \nu) = \sup_{\substack{h \text{ is 1-Lipschitz} \\ \text{on } \mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}}}} \left(\int_{\mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}}} h d\mu - \int_{\mathbb{R}^{d_{\mathrm{in}}+d_{\mathrm{out}}}} h d\nu \right), \tag{2.1}$$

and from [FG15] we have the following proposition, where δ_x represents the Dirac delta at the point x.

Proposition 2.1 *[FG15, Theorems 1 and 2]. Suppose that $d_{\mathrm{in}} + d_{\mathrm{out}} \geq 5$, and let*

$$\tilde{\rho}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}(dz)$$

denote the empirical measure associated to the sequence $(Z_i)_{1 \leq i \leq n} = (X_i, Y_i)_{1 \leq i \leq n}$.

a) We have the Wasserstein bound

$$\mathbb{E}\left[\mathcal{W}_F^2(\rho, \tilde{\rho}_n)\right] \leq Cn^{-2/(d_{\text{in}}+d_{\text{out}})}. \tag{2.2}$$

b) For any $\zeta \in (0,1)$, we have the concentration inequality

$$\mathbb{P}\left(\mathcal{W}_1(\rho, \tilde{\rho}_n) \leq \left(\frac{Cn}{\log(C/\zeta)}\right)^{-1/(d_{\text{in}}+d_{\text{out}})}\right) \geq 1-\zeta, \tag{2.3}$$

where $C > 0$ is a constant independent of $\zeta \in (0,1)$.

SGM dynamics

For $\lambda > 0$, let ℓ_λ denote the loss function

$$\ell_\lambda(x, y, v, w) := l(x, y) + \frac{\lambda}{2}\left(\|v\|_F^2 + \|w\|_F^2\right), \tag{2.4}$$

where $\|v\|_F$, $\|w\|_F$ are the Frobenius norms of $v \in \mathbb{R}^{d_{\text{in}} \times d}$, $w \in \mathbb{R}^{d \times d_{\text{out}}}$, and $\lambda > 0$ is a regularization parameter. Given a neural network function $f(\cdot, v, w) : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ parameterized by $v \in \mathbb{R}^{d_{\text{in}} \times d}$ and $w \in \mathbb{R}^{d \times d_{\text{out}}}$, we aim at finding the infimum

$$\inf_{(v,w) \in \mathbb{R}^{d_{\text{in}} \times d} \times \mathbb{R}^{d \times d_{\text{out}}}} \mathcal{L}(v, w), \tag{2.5}$$

where

$$\mathcal{L}(v, w) := \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} \ell_\lambda(f(x, v, w), y, v, w)\rho(dx, dy). \tag{2.6}$$

As we do not have the access to the actual data distribution ρ, given

$$Z^{(t)} = \left(X_i^{(t)}, Y_i^{(t)}\right)_{i=1,\ldots,k}$$

a set of independent data samples $\left(X_i^{(t)}, Y_i^{(t)}\right) \in \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}$ of batch size $k \geq 1$, identically distributed according to ρ at times $t \geq 0$, we approximate (2.5) by minimization of

$$\inf_{(v,w) \in \mathbb{R}^{d_{\text{in}} \times d} \times \mathbb{R}^{d \times d_{\text{out}}}} \mathcal{L}_k\left(Z^{(t)}, v, w\right)$$

where

$$\mathcal{L}_k\left(Z^{(t)}, v, w\right) := \frac{1}{k}\sum_{i=1}^{k} \ell_\lambda\left(f\left(X_i^{(t)}, v, w\right), Y_i^{(t)}, v, w\right). \tag{2.7}$$

For this, we use the sequence $(V(t), W(t))_{0\leq t\leq T}$ defined by the Stochastic Gradient Method (SGM), through the dynamics

$$\begin{cases} V(t+1) = V(t) - \eta_V(t)\nabla_v \mathcal{L}_k\big(Z^{(t)}, V(t), W(t)\big), \\ W(t+1) = W(t) - \eta_W(t)\nabla_w \mathcal{L}_k\big(Z^{(t)}, V(t), W(t)\big), \quad t = 0, 1, \dots, T-1, \end{cases} \tag{2.8}$$

where $(\eta_V(t))_{0\leq t<T}$ and $(\eta_W(t))_{0\leq t<T}$ denote (positive) learning rate sequences.

We will work under the following conditions.

Assumption 1 *We assume that*

- $\mathrm{Supp}(\rho) \subset \{z = (x, y) \in \mathbb{R}^{d_{\mathrm{in}}} \times \mathbb{R}^{d_{\mathrm{out}}} \ : \ \max(|x|, |y|) \leq 1\}$,
- *the function* $l : \mathbb{R}^{d_{\mathrm{out}}} \times \mathbb{R}^{d_{\mathrm{out}}} \to \mathbb{R}$ *is* $\mathcal{C}^1$, 1*-Lipschitz, and satisfies*

$$l(y, y) = 0, \quad y \in \mathbb{R}^{d_{\mathrm{out}}}, \ |y| \leq 1,$$

- $f(x, v, w)$ *is a two-layer neural network of the form*

$$f(x, v, w) = w^\top \sigma\big(v^\top x\big), \quad x \in \mathbb{R}^{d_{\mathrm{in}}}, \tag{2.9}$$

  where $\sigma : \mathbb{R} \to \mathbb{R}$ *is a* $\mathcal{C}^1$ *and* 1*-Lipschitz activation function such that* $\sigma(0) = 0$*, which is applied componentwise to* $v^\top x \in \mathbb{R}^d$*,*

- *the SGD dynamics satisfies the learning rate conditions*

$$0 \leq \eta_W(t) \leq \eta_V(t) \leq \frac{1}{\lambda}, \quad 0 \leq t < T.$$

- *The entries of the matrices* $V(0) \in \mathbb{R}^{d_{\mathrm{in}} \times d}$ *and* $W(0) \in \mathbb{R}^{d \times d_{\mathrm{out}}}$ *are initialized via He initialization, using independent centered Gaussian samples with variance* κ/d *(resp.* κ/d_{out}*), with* $\kappa = 2$*, see [HZRS15].*

We note that by taking $K > 0$, Assumption 1 can be relaxed by only assuming that $\max(|x|, |y|) \leq K$ for all $(x, y) \in \mathrm{Supp}(\rho)$, and that the function l and activation function σ are K-Lipschitz.

3 SGM moment bounds

In this section, we control the spectral norms of $V(T)$ and $W(T)$ in the SGM dynamics (2.8). We let $p!!$ denote the double factorial of $p \geq 0$.

Proposition 3.1 *Moment bounds. Suppose that Assumption 1 holds.*

a) If $W(t)$ remains frozen at $W(t) := w \in \mathbb{R}^{d\times d_{\text{out}}}$, i.e. $\eta_W(t) = 0$, $0 \leq t < T$, then for all $p \geq 1$ we have

$$\mathbb{E}\left[\|V(T)\|_F^p\right] \leq (p-1)!!2^{p-1}(\kappa d_{\text{in}})^{p/2}\left(\prod_{t=0}^{T-1}(1-\eta_V(t)\lambda)\right)^p + 2^{p-1}\frac{\|w\|_F^p}{\lambda^p}\left(1-\prod_{t=0}^{T-1}(1-\eta_V(t)\lambda)\right)^p. \tag{3.1}$$

b) If $\eta_W(t) = \eta_V(t) := \eta(t)$, $0 \leq t < T$, then for any $p \geq 1$ we have

$$\mathbb{E}\big[\|V(T)\|_F^p\|W(T)\|_F^p\big] \leq \frac{\kappa^d}{2}(2p-1)!!\big(d_{\text{in}}^p + d_{\text{out}}^p\big)\prod_{t=0}^{T-1}(1-\eta(t)\lambda+\eta(t))^{2p}. \tag{3.2}$$

Proof. From (2.9), the regularized loss function (2.4) satisfies

$$\ell_\lambda(f(x,v,w),y,v,w) = l\big(w^\top\sigma\big(v^\top x\big),y\big) + \frac{\lambda}{2}\left(\|v\|_F^2 + \|w\|_F^2\right),$$

hence

$$\nabla_v\ell_\lambda(f(x,v,w),y,v,w) = x\big((\nabla_{y_1}l)\big(w^\top\sigma\big(v^\top x\big),y\big)\big)^\top w^\top\Sigma + \lambda v,$$

where

$$\Sigma := \text{Diag}\big(\sigma'\big(v^\top x\big)\big)$$

is the square diagonal matrix with $\sigma'\big(v^\top x\big)$ as diagonal entries, and the derivative σ' of the activation function σ is applied componentwise to $v^\top x \in \mathbb{R}^d$. Therefore, from (2.7) we have

$$\begin{aligned}\nabla_v\mathcal{L}_k\big(Z^{(t)},v,w\big) &= \frac{1}{k}\sum_{i=1}^k \nabla_v\ell_\lambda\big(f\big(X_i^{(t)},v,w\big),Y_i^{(t)},v,w\big)\\ &= \frac{1}{k}\sum_{i=1}^k\big(X_i^{(t)}\big(\nabla_{y_1}l\big(w^\top\sigma\big(v^\top X_i^{(t)}\big),Y_i^{(t)}\big)\big)^\top w^\top\Sigma + \lambda v\big),\end{aligned}$$

and (2.8) yields

$$V(t+1) = V(t) - \eta_V(t)\nabla_v\mathcal{L}_k\big(Z^{(t)},V(t),W(t)\big)$$

$$= \big(1 - \eta_V(t)\lambda\big)V(t) - \frac{\eta_V(t)}{k}\sum_{i=1}^{k} X_i^{(t)}\big(\nabla_{y_1} l\big(W(t)^\top \sigma\big(v^\top X_i^{(t)}\big), Y_i^{(t)}\big)\big)^\top W(t)^\top \Sigma,$$

hence from the bound

$$\max(|\nabla_{y_1} l(y_1, y_2)|, |\nabla_{y_2} l(y_1, y_2)|) \le 1, \quad y_1, y_2 \in \mathbb{R}^{d_{\text{out}}}, \tag{3.3}$$

we have

$$\begin{aligned}
\|V(t+1)\|_F &= \|V(t) - \eta_V(t)\nabla_v \mathcal{L}_k\big(Z^{(t)}, V(t), W(t)\big)\|_F \\
&\le (1 - \eta_V(t)\lambda)\|V(t)\|_F + \frac{\eta_V(t)}{k}\sum_{i=1}^{k}\big\|X_i^{(t)}\big(\nabla_{y_1} l\big(W(t)^\top \sigma\big(v^\top X_i^{(t)}\big), Y_i^{(t)}\big)\big)^\top W(t)^\top \Sigma\big\|_F \\
&\le (1 - \eta_V(t)\lambda)\|V(t)\|_F + \eta_V(t)\|W(t)\|_F.
\end{aligned} \tag{3.4}$$

Next, from the relation

$$\nabla_w \ell_\lambda(f(x, v, w), y, v, w) = \sigma\big(v^\top x\big)\big((\nabla_{y_1} l)\big(w^\top \sigma\big(v^\top x\big), y\big)\big)^\top + \lambda w$$

and (2.7), we have

$$\begin{aligned}
\nabla_w \mathcal{L}_k\big(Z^{(t)}, V(t), W(t)\big) &= \frac{1}{k}\sum_{i=1}^{k}\nabla_w\big(\ell_\lambda\big(f\big(X_i^{(t)}, V(t), W(t)\big), Y_i^{(t)}, V(t), W(t)\big)\big) \\
&= \frac{1}{k}\sum_{i=1}^{k}\big(\sigma\big(V(t)^\top X_i^{(t)}\big)\big((\nabla_{y_1} l)\big(W(t)^\top \sigma\big(V(t)^\top X_i^{(t)}\big), Y_i^{(t)}\big)\big)^\top + \lambda W(t)\big),
\end{aligned}$$

and (2.8) yields

$$\begin{aligned}
W(t+1) &= W(t) - \eta_W(t)\nabla_w \mathcal{L}_k\big(Z^{(t)}, V(t), W(t)\big) \\
&= \big(1 - \eta_W(t)\lambda\big)W(t) - \frac{\eta_W(t)}{k}\sum_{i=1}^{k}\sigma\big(V(t)^\top X_i^{(t)}\big)\big(\nabla_{y_1} l\big(W(t)^\top \sigma\big(V(t)^\top X_i^{(t)}\big), Y_i^{(t)}\big)\big)^\top,
\end{aligned}$$

hence from (3.3) we have

$$\begin{aligned}
\|W(t+1)\|_F &= \|W(t) - \eta_W(t)\nabla_w \mathcal{L}_k(Z^{(t)}, V(t), W(t))\|_F \\
&\le (1 - \eta_W(t)\lambda)\|W(t)\|_F + \frac{\eta_W(t)}{k}\sum_{i=1}^{k}\big\|\sigma\big(V(t)^\top X_i^{(t)}\big)\big(\nabla_{y_1} l\big(W(t)^\top \sigma\big(V(t)^\top X_i^{(t)}\big), Y_i^{(t)}\big)\big)^\top\big\|_F \\
&\le (1 - \eta_W(t)\lambda)\|W(t)\|_F + \eta_W(t)\|V(t)\|_F \\
&\le (1 - \eta_W(t)\lambda)\|W(t)\|_F + \eta_V(t)\|V(t)\|_F.
\end{aligned} \tag{3.5}$$

a) If $\eta_W(t) = 0$, $0 \le t < T$, then from (3.4) and (3.5) we have

$$\begin{aligned}\|V(T)\|_F &\le \|V(0)\|_F \left(\prod_{t=0}^{T-1}(1-\eta_V(t)\lambda)\right) + \|W(0)\|_F \sum_{t=0}^{T-1}\eta_V(t)\prod_{i=t+1}^{T-1}(1-\eta_V(i)\lambda)\\ &= \|V(0)\|_F \left(\prod_{t=0}^{T-1}(1-\eta_V(t)\lambda)\right) + \frac{\|w\|_F}{\lambda}\left(1-\prod_{t=0}^{T-1}(1-\eta_V(t)\lambda)\right). \end{aligned} \tag{3.6}$$

We conclude by taking expectations on both sides and noting that since the matrix $V(0) = (v_{i,j})_{(i,j)\in d_{\text{in}}\times d} \in \mathbb{R}^{d_{\text{in}}\times d}$ has centered independent Gaussian entries with variance κ/d, we have

$$\begin{aligned}\mathbb{E}\left[\|V(0)\|_F^p\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{d_{\text{in}}}\sum_{j=1}^{d}|v_{i,j}|^2\right)^{p/2}\right]\\ &\le (d_{\text{in}}d)^{p/2-1}\sum_{(i,j)\in d_{\text{in}}\times d}\mathbb{E}\left[|v_{i,j}|^p\right]\\ &= (p-1)!!(\kappa d_{\text{in}})^{p/2}, \quad p \ge 1. \end{aligned} \tag{3.7}$$

b) If $\eta_V(t) = \eta_W(t) := \eta(t)$, $0 \le t < T$, then from (3.4) and (3.5) we have

$$\|V(T)\|_F + \|W(T)\|_F \le (\|V(0)\|_F + \|W(0)\|_F)\prod_{t=0}^{T-1}(1-\eta(t)\lambda+\eta(t)), \tag{3.8}$$

hence

$$\begin{aligned}\mathbb{E}\big[\|V(T)\|_F^p\|W(T)\|_F^p\big] &\le 2^{-2p}\mathbb{E}\big[(\|V(T)\|_F + \|W(T)\|_F)^{2p}\big]\\ &\le 2^{-2p}\mathbb{E}\big[(\|V(0)\|_F + \|W(0)\|_F)^{2p}\big]\prod_{t=0}^{T-1}(1-\eta(t)\lambda+\eta(t))^{2p}\\ &\le \frac{1}{2}\mathbb{E}\big[\|V(0)\|_F^{2p} + \|W(0)\|_F^{2p}\big]\prod_{t=0}^{T-1}(1-\eta(t)\lambda+\eta(t))^{2p}\\ &\le \kappa^d\frac{(2p-1)!!}{2}\big(d_{\text{in}}^p + d_{\text{out}}^p\big)\prod_{t=0}^{T-1}(1-\eta(t)\lambda+\eta(t))^{2p}, \end{aligned}$$

since, as in (3.7), we have

$$\mathbb{E}\left[\|V(0)\|_F^{2p}\right] \le (2p-1)!!(\kappa d_{\text{in}})^p \quad \text{and} \quad \mathbb{E}\left[\|W(0)\|_F^{2p}\right] \le (2p-1)!!(\kappa d_{\text{out}})^p, \quad p \ge 1,$$

see also [RV10]. □

We note that the upper bounding constants in Proposition 3.1 and in subsequent results remain bounded as T tends to infinity, provided that the sequence $(\eta_V(t))_{t\geq 0}$ is summable, i.e.

$$\sum_{t\geq 0} |\eta_V(t)| < \infty,$$

which is the case in particular for schedules with polynomial time decay of the form $1/a^t$, $a > 1$.

In the case of constant schedules $\eta_W(t) = \eta_V(t) = \eta$, $0 \leq t < T$, the bounds (3.1) and (3.2) become

$$\mathbb{E}\left[\|V(T)\|_F^p\right] \leq (p-1)!!2^{p-1}(\kappa d_{\text{in}})^{p/2}(1-\lambda\eta)^{pT} + 2^{p-1}\frac{\|w\|_F^p}{\lambda^p}\big(1-(1-\eta\lambda)^T\big)^p \tag{3.9}$$

and

$$\mathbb{E}\big[\|V(T)\|_F^p\|W(T)\|_F^p\big] \leq \frac{\kappa^d}{2}(2p-1)!!\big(d_{\text{in}}^p + d_{\text{out}}^p\big)(1+(1-\lambda)\eta)^{2pT}. \tag{3.10}$$

In this case, (3.9) tends to $2^{p-1}\|w\|_F^p/\lambda^p$ while (3.10) explodes as T tends to infinity.

4 Independent samples

In Proposition 4.1 we derive L^1 error bounds of dimension-free order for the absolute generalization error $\left|\varepsilon_{\text{gen}}(n, V(T), W(T))\right|$ by assuming as in [KKB17] that the testing set is independent of the data sequence used for SGM updates in (2.8).

Proposition 4.1 *Suppose that Assumption 1 holds and that the testing set*

$$Z := (Z_i)_{1\leq i\leq n} = (X_i, Y_i)_{1\leq i\leq n} \overset{i.i.d.}{\sim} \rho$$

is independent of the training sequence $(Z^{(t)})_{0\leq t\leq T}$.

a) If $W(t)$ *remains frozen at* $W(t) := w \in \mathbb{R}^{d\times d_{\text{out}}}$, *i.e.* $\eta_W(t) = 0$, $0 \leq t < T$, *then we have*

$$\mathbb{E}\left[\left|\varepsilon_{\text{gen}}(n, V(T), w)\right|\right] \leq (1 + C_1(w,T))\frac{2}{\sqrt{n}}, \tag{4.1}$$

where

$$C_1(w,T) := \|w\|_F\sqrt{\kappa d_{\text{in}}}\prod_{t=0}^{T-1}(1-\eta(t)\lambda) + \frac{\|w\|_F^2}{\lambda}\left(1 - \prod_{t=0}^{T-1}(1-\eta(t)\lambda)\right).$$

b) If $\eta_W(t) = \eta_V(t) := \eta(t)$, $0 \le t < T$, then we have

$$\mathbb{E}\big[\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big|\big] \le (1 + C_2(1,T))\frac{2}{\sqrt{n}}, \tag{4.2}$$

where $C_2(1,T)$ is defined from

$$C_2(p,T) := (2p-1)!!2^{p-1}\big(d_{\text{in}}^p + d^p\big)\kappa^p \prod_{t=0}^{T-1}(1 + (1-\lambda)\eta(t))^{2p}, \quad p \ge 1. \tag{4.3}$$

Proof. *a)* Due to the relations

$$\begin{cases} \nabla_x l(f(x,v,w),y) = v\Sigma w(\nabla_{y_1} l)(f(x,v,w),y), \\ \nabla_y l(f(x,v,w),y) = (\nabla_{y_2} l)(f(x,v,w),y), \end{cases}$$

the function $(x,y) \mapsto l(f(x,v,w),y)$ is $(\|v\|_F\|w\|_F)$-Lipschitz in x and 1-Lipschitz in y, with

$$\big|l(f(x',v,w),y') - l(f(x,v,w),y)\big| \le \|v\|_F\|w\|_F|x-x'| + |y'-y|, \tag{4.4}$$

$x, x' \in \mathbb{R}^{d_{\text{in}}}$, $y, y' \in \mathbb{R}^{d_{\text{out}}}$, $v \in \mathbb{R}^{d_{\text{in}} \times d}$, $w \in \mathbb{R}^{d \times d_{\text{out}}}$. Hence, using Hölder's inequality, the independence of $(X_i, Y_i)_{1 \le i \le n}$ and $V(T)$, and the facts that for all $z, z' \in \text{Supp}(\rho)$, $|z - z'| \le 2$ and (3.3), we have

$$\begin{aligned}
\big|\varepsilon_{\text{gen}}(n, V(T), w)\big| &= \left|\frac{1}{n}\sum_{i=1}^n l(f(X_i, V(T), w), Y_i) - \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} l(f(x, V(T), w), y)\rho(dx, dy)\right| \\
&= \left|\int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} l(f(x, V(T), w), y)(\tilde{\rho}(dx,dy) - \rho(dx,dy))\right| \\
&\le \sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n l(f(X_i, V(T), w), Y_i) - \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} l(f(x, V(T), w), y)\rho(dx,dy)\right)^2 \,\middle|\, V(T)\right]} \\
&= \sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} (l(f(X_i, V(T), w), Y_i) - l(f(x, V(T), w), y))\rho(dx,dy)\right)^2 \,\middle|\, V(T)\right]} \\
&\le \sqrt{\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} (l(f(X_i, V(T), w), Y_i) - l(f(x, V(T), w), y))^2\rho(dx,dy) \,\middle|\, V(T)\right]} \\
&\le \sqrt{\frac{4}{n^2}\sum_{i=1}^n \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} (1 + \|V(T)\|_F\|w\|_F)^2\rho(dx,dy)} \\
&= \frac{2}{\sqrt{n}}(1 + \|V(T)\|_F\|w\|_F),
\end{aligned}$$

where we applied (4.4), hence

$$\mathbb{E}\big[\big|\varepsilon_{\text{gen}}(n, V(T), w)\big|\big] \leq \frac{2}{\sqrt{n}}(1 + \|w\|_F \mathbb{E}\left[\|V(T)\|_F\right]),$$

and we conclude by the application of Proposition 3.1-(a) with $p = 1$.
b) By the same argument as in part (a) we have

$$\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big| \leq \frac{2}{\sqrt{n}}(1 + \mathbb{E}\left[\|W(T)\|_F \|V(T)\|_F\right]),$$

and we conclude by the application of Proposition 3.1-(b) with $p = 1$. □

In Proposition 4.2, we present related deviation inequalities.

Proposition 4.2 *Suppose that Assumption 1 holds and that the testing set*

$$Z := (Z_i)_{1 \leq i \leq n} = (X_i, Y_i)_{1 \leq i \leq n} \overset{i.i.d.}{\sim} \rho$$

is independent of the training sequence $(Z^{(t)})_{0 \leq t \leq T}$.

a) If $W(t)$ *remains frozen at* $W(t) := w \in \mathbb{R}^{d \times d_{\text{out}}}$, *i.e.* $\eta_W(t) = 0$, $0 \leq t < T$, *then for any* $\zeta \in (0, 1)$ *we have*

$$\mathbb{P}\left(\big|\varepsilon_{\text{gen}}(n, V(T), w)\big| \leq (1 + C_3(w, \zeta, T))\sqrt{\frac{2}{n}\log\frac{4}{\zeta}}\right) \geq 1 - \zeta,$$

where

$$C_3(w, \zeta, T) := \|w\|_F \left(\sqrt{d_{\text{in}}} + \sqrt{d} + \sqrt{2\log\frac{4}{\zeta}}\right)\sqrt{\frac{\kappa}{d}\prod_{t=0}^{T-1}(1-\eta(t)\lambda)} + \frac{\|w\|_F^2}{\lambda}\left(1 - \prod_{t=0}^{T-1}(1 - \eta(t)\lambda)\right).$$

b) If $\eta_W(t) = \eta_V(t) := \eta(t)$, $0 \leq t < T$, *then for any* $\zeta \in (0, 1)$ *we have*

$$\mathbb{P}\left(\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big| \leq (1 + C_5(\zeta, T))\sqrt{\frac{2}{n}\log\frac{4}{\zeta}}\right) \geq 1 - \zeta,$$

where

$$C_5(\zeta, T) := 3\kappa\left(2 + \frac{d_{\text{in}}}{d} + \frac{d}{d_{\text{out}}} + \left(\frac{2}{d} + \frac{2}{d_{\text{out}}}\right)\log\frac{8}{\zeta}\right)\prod_{t=0}^{T-1}(1 + (1-\lambda)\eta(t))^2.$$

Proof. a) From (4.4), we have the bound

$$\begin{aligned}|l(f(x, v, w), y)| &\leq |l(f(x, v, w), y) - l(y, y)| + l(y, y)\\ &\leq |f(x, v, w) - y|\end{aligned}$$

$$\leq 1 + \|v\|_F \|w\|_F,$$

$x \in \mathbb{R}^{d_{\text{in}}}$, $y \in \mathbb{R}^{d_{\text{out}}}$, $v \in \mathbb{R}^{d_{\text{in}} \times d}$, $w \in \mathbb{R}^{d \times d_{\text{out}}}$, hence by Hoeffding's inequality, see Theorem 1 in [Hoe63], the generalization error

$$\varepsilon_{\text{gen}}(n, V(T), w) = \frac{1}{n} \sum_{i=1}^{n} l(f(X_i, V(T), w), Y_i) - \int_{\mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{out}}}} l(f(x, V(T), w), y) \rho(dx, dy)$$

satisfies

$$\mathbb{P}\left(\left| \varepsilon_{\text{gen}}(n, V(T), w) \right| \leq (1 + \|V(T)\|_F \|w\|_F) \sqrt{\frac{2}{n} \log \frac{2}{\zeta}} \;\middle|\; V(T) \right) \geq 1 - \zeta,$$

which yields

$$\mathbb{P}\left(\left| \varepsilon_{\text{gen}}(n, V(T), w) \right| \leq (1 + \|V(T)\|_F \|w\|_F) \sqrt{\frac{2}{n} \log \frac{2}{\zeta}} \right) \geq 1 - \zeta. \tag{4.5}$$

Next, by (3.6) and the bound (2.3) in [RV10], which implies

$$\mathbb{P}\left(\|V(0)\|_F \leq \sqrt{\frac{\kappa}{d}} \left(\sqrt{d_{\text{in}}} + \sqrt{d} + \sqrt{2 \log \frac{4}{\zeta}} \right) \right) \geq 1 - \frac{\zeta}{2}, \quad \zeta \in (0, 1), \tag{4.6}$$

we get

$$\mathbb{P}(\|V(T)\|_F \|w\|_F \leq C_3(w, \zeta, T)) \geq 1 - \frac{\zeta}{2}. \tag{4.7}$$

Hence, from (4.5)-(4.7) and the inequality $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$, we have

$$\begin{aligned}
&1 - \zeta \\
&\leq \mathbb{P}\left(\left| \varepsilon_{\text{gen}}(n, V(T), w) \right| \leq \sqrt{\frac{2}{n} \log \frac{4}{\zeta}} (1 + \|V(T)\|_F \|w\|_F) \right) && (4.8) \\
&\quad - 1 + \mathbb{P}\left(\|V(T)\|_F \|w\|_F \leq C_3(w, \zeta, T) \right) && (4.9) \\
&\leq \mathbb{P}\left(\left| \varepsilon_{\text{gen}}(n, V(T), w) \right| \leq (1 + \|V(T)\|_F \|w\|_F) \sqrt{\frac{2}{n} \log \frac{4}{\zeta}} \;\text{ and }\; \|V(T)\|_F \|w\|_F \leq C_3(w, \zeta, T) \right) \\
&\leq \mathbb{P}\left(\left| \varepsilon_{\text{gen}}(n, V(T), w) \right| \leq (1 + C_3(w, \zeta, T)) \sqrt{\frac{2}{n} \log \frac{4}{\zeta}} \right) && (4.10)
\end{aligned}$$

which completes the proof.

b) From (4.6), we have

$$\begin{aligned}
1 - \zeta &\leq \left(1 - \frac{\zeta}{2} \right)^2 \\
&\leq \mathbb{P}\left(\|V(0)\|_F \leq \sqrt{\frac{\kappa}{d}} \left(\sqrt{d_{\text{in}}} + \sqrt{d} + \sqrt{2 \log \frac{4}{\zeta}} \right) \right)
\end{aligned}$$

$$\times \mathbb{P}\left(\|W(0)\|_F \leq \sqrt{\frac{\kappa}{d_{\text{out}}}}\left(\sqrt{d}+\sqrt{d_{\text{out}}}+\sqrt{2\log\frac{4}{\zeta}}\right)\right)$$
$$\leq \mathbb{P}\left(\|V(0)\|_F^2 \leq 3\kappa\left(1+\frac{d_{\text{in}}}{d}+\frac{2}{d}\log\frac{4}{\zeta}\right) \text{ and } \|W(0)\|_F^2 \leq 3\kappa\left(1+\frac{d}{d_{\text{out}}}+\frac{2}{d_{\text{out}}}\log\frac{4}{\zeta}\right)\right)$$
$$\leq \mathbb{P}\left(\|V(0)\|_F^2+\|W(0)\|_F^2 \leq 3\kappa\left(2+\frac{d_{\text{in}}}{d}+\frac{d}{d_{\text{out}}}+\left(\frac{2}{d}+\frac{2}{d_{\text{out}}}\right)\log\frac{4}{\zeta}\right)\right), \tag{4.11}$$

where the third inequality uses Hölder's inequality and the independence of $V(0), W(0)$. We conclude from (3.8) using the same argument as in (4.10). □

5 Random subset

In this section, we make no independence assumption between the testing set

$$Z := (Z_i)_{1\leq i\leq n} = (X_i, Y_i)_{1\leq i\leq n} \overset{\text{i.i.d.}}{\sim} \rho$$

and the training sequence $(Z^{(t)})_{0\leq t\leq T}$ used for SGM updates in (2.8). In Proposition 5.1, using Propositions 2.1 and 3.1 we derive bounds on the generalization error (1.1) of (2.8) under the technical condition $d_{\text{in}} + d_{\text{out}} \geq 5$ which originates in Proposition 2.1, and can be removed at the expense of additional analysis.

Proposition 5.1 *Suppose that $d_{\text{in}} + d_{\text{out}} \geq 5$ and that Assumption 1 holds.*

a) Assume that $\eta_W(t) = 0$, $0 \leq t < T$, i.e. $W(t)$ remains frozen at $W(t) := w \in \mathbb{R}^{d\times d_{\text{out}}}$, $0 \leq t \leq T$. Then, we have

$$\mathbb{E}\big[\big|\varepsilon_{\text{gen}}(n, V(T), w)\big|\big] \leq \frac{\sqrt{(1+C_4(w,T))C}}{n^{1/(d_{\text{in}}+d_{\text{out}})}}, \quad n \geq 1,$$

where

$$C_4(w,T) := 2\|w\|_F\kappa d_{\text{in}}\left(\prod_{t=0}^{T-1}(1-\eta(t)\lambda)\right)^2 + 2\frac{\|w\|_F^3}{\lambda^2}\left(1-\prod_{t=0}^{T-1}(1-\eta(t)\lambda)\right)^2,$$

and $C > 0$ is the constant given in (2.2).

b) If $\eta_W(t) = \eta_V(t) := \eta(t)$, $0 \leq t < T$, then we have

$$\mathbb{E}\big[\big|\varepsilon_{\text{gen}}(n, V(T), W(T))\big|\big] \leq \frac{\sqrt{(1+C_2(2,T))C}}{n^{1/(d_{\text{in}}+d_{\text{out}})}}, \quad n \geq 1,$$

where $C_2(2,T)$ is defined in (4.3) and $C > 0$ is the constant given in (2.2).

Proof. *a)* Since from (4.4) the function $(x,y) \mapsto l(f(x,V(T),w),y)$ is $(\|w\|_F\|V(T)\|_F)$-Lipschitz in x and 1-Lipschitz in y, using Hölder's inequality, (2.1) and (3.3), we have

$$\begin{aligned}\left|\varepsilon_{\text{gen}}(n,V(T),w)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} l(f(X_i,V(T),w),Y_i) - \int_{\mathbb{R}^{d_{\text{in}}}\times\mathbb{R}^{d_{\text{out}}}} l(f(x,V(T),w),y)\rho(dx,dy)\right| \\ &= \left|\int_{\mathbb{R}^{d_{\text{in}}}\times\mathbb{R}^{d_{\text{out}}}} l(f(x,V(T),w),y)(\tilde{\rho}(dx,dy)-\rho(dx,dy))\right| \\ &\le \mathcal{W}_1(\rho,\tilde{\rho}_n)\sqrt{1+\|V(T)\|_F^2\|w\|_F^2}, \end{aligned} \tag{5.1}$$

hence

$$\begin{aligned}\mathbb{E}\left[\left|\varepsilon_{\text{gen}}(n,V(T),w)\right|\right] &\le \mathbb{E}\left[\mathcal{W}_1(\rho,\tilde{\rho}_n)\sqrt{1+\|V(T)\|_F^2\|w\|_F^2}\right] \\ &\le \sqrt{\left(1+\|w\|_F^2\mathbb{E}\left[\|V(T)\|_F^2\right]\right)\mathbb{E}\left[\mathcal{W}_F^2(\rho,\tilde{\rho}_n)\right]},\end{aligned}$$

which completes the proof by (2.2) and Proposition 3.1-(a).
b) The argument is the same as in part (a), replacing the use of Proposition 3.1-(a) with that of Proposition 3.1-(b). □

Similarly, using Proposition 2.1 we obtain the following concentration inequality.

Proposition 5.2 *Suppose that $d_{\text{in}}+d_{\text{out}} \ge 5$ and that Assumption 1 holds.*

a) Assume that $\eta_W(t)=0$, $0\le t<T$, i.e. $W(t)$ remains frozen at $W(t):=w\in\mathbb{R}^{d\times d_{\text{out}}}$, $0\le t\le T$. Then, for any $\zeta\in(0,1)$ we have

$$\mathbb{P}\left(\left|\varepsilon_{\text{gen}}(n,V(T),w)\right| \le \left(\frac{Cn}{\log(2C/\zeta)}\right)^{-1/(d_{\text{in}}+d_{\text{out}})}(1+C_3(w,\zeta,T))\right) \ge 1-\zeta, \quad n\ge 1,$$

where $C_3(\zeta,T)$ is defined in Proposition 4.2-(a) and C is given in (2.3).

b) If $\eta_W(t)=\eta_V(t):=\eta(t)$, $0\le t<T$, then for any $\zeta\in(0,1)$ we have

$$\mathbb{P}\left(\left|\varepsilon_{\text{gen}}(n,V(T),W(T))\right| \le \left(\frac{Cn}{\log(2C/\zeta)}\right)^{-1/(d_{\text{in}}+d_{\text{out}})}(1+C_5(\zeta,T))\right) \ge 1-\zeta, \quad n\ge 1,$$

where $C_5(\zeta,T)$ is defined in Proposition 4.2-(b).

Proof. *a)* By (3.6) and (4.6) we have

$$\mathbb{P}\big(\|V(T)\|_F\|w\|_F \le C_3(w,\zeta,T)\big) \ge 1-\frac{\zeta}{2},$$

hence, using the inequality $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$, the bounds (2.3) and (5.1), we have

$$
\begin{aligned}
1 - \zeta &\leq \mathbb{P}\left(\mathcal{W}_1(\rho, \tilde{\rho}_n) \leq \left(\frac{Cn}{\log(2C/\zeta)} \right)^{-1/(d_{\mathrm{in}}+d_{\mathrm{out}})} \right) + \mathbb{P}\big(\|V(T)\|_F \|w\|_F \leq C_3(w, \zeta, T) \big) - 1 \\
&\leq \mathbb{P}\left(\mathcal{W}_1(\rho, \tilde{\rho}_n) \leq \left(\frac{Cn}{\log(2C/\zeta)} \right)^{-1/(d_{\mathrm{in}}+d_{\mathrm{out}})} \text{ and } \ \|V(T)\|_F \|w\|_F \leq C_3(w, \zeta, T) \right) \\
&\leq \mathbb{P}\left(\big| \varepsilon_{\mathrm{gen}}(n, V(T), w) \big| \leq \mathcal{W}_1(\rho, \tilde{\rho}_n) \sqrt{1 + \|w\|_F^2 \|V(T)\|_F^2} \leq \left(\frac{Cn}{\log(2C/\zeta)} \right)^{-1/(d_{\mathrm{in}}+d_{\mathrm{out}})} (1 + C_3(w, \zeta, T)) \right).
\end{aligned}
$$

b) The proof proceeds as in part (*a*), by replacing the uses of (3.6) and (4.6) with those of (3.8) and (4.11). □

Proposition 5.3 presents L^p bounds on the Lipschitz constant of the loss function $\ell_\lambda(f(x, v, w), y, v, w)$.

Proposition 5.3 *Lipschitz bound. Suppose that Assumption 1 holds.*

a) Assume that $\eta_W(t) = 0$, $0 \leq t < T$, i.e. $W(t)$ remains frozen at $W(t) := w \in \mathbb{R}^{d \times d_{\mathrm{out}}}$, $0 \leq t \leq T$. For $p = 1$ we have the gradient bound

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{(x,y) \,\in\, \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x, V(T), w), y, V(T), w)| \right] \\
&\qquad \leq \|w\|_F \big(\sqrt{d_{\mathrm{in}}} + \sqrt{d} \big) \sqrt{\frac{\kappa}{d}} \left(\prod_{t=0}^{T-1} (1 - \eta(t)\lambda) \right) + \frac{\|w\|_F^2}{\lambda} \left(1 - \prod_{t=0}^{T-1} (1 - \eta(t)\lambda) \right),
\end{aligned}
$$

and for $p \geq 2$ we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{(x,y) \,\in\, \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x, V(T), w), y, V(T), w)|^p \right] \\
&\qquad \leq (p-1)!! 2^{p-1} \|w\|_F^p (\kappa d_{\mathrm{in}})^{p/2} \left(\prod_{t=0}^{T-1} (1 - \eta(t)\lambda) \right)^p + 2^{p-1} \frac{\|w\|_F^{2p}}{\lambda^p} \left(1 - \prod_{t=0}^{T-1} (1 - \eta(t)\lambda) \right)^p.
\end{aligned}
$$

b) If $\eta_W(t) = \eta_V(t) := \eta(t)$, $0 \leq t < T$, then for any $p \geq 1$ we have

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{(x,y) \,\in\, \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x, V(T), W(T)), y, V(T), W(T))|^p \right] \\
&\qquad\qquad \leq \frac{(2p-1)!!}{2^{1-p}} \big(d_{\mathrm{in}}^p + d_{\mathrm{out}}^p \big) \kappa^p \prod_{t=0}^{T-1} (1 + (1-\lambda)\eta(t))^{2p}.
\end{aligned}
$$

Proof. *a*) From the relation

$$\nabla_x \ell_\lambda(f(x,v,w),y,v,w) = v\Sigma w(\nabla_{y_1} l)(f(x,v,w),y) \tag{5.2}$$

combined with (3.3) and (3.6), we have

$$\begin{aligned}\sup_{(x,y)\ \in\ \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x,V(T),w),y,V(T),w)|^p &\le \|w\|_F^p \|V(T)\|_F^p \qquad (5.3)\\ &\le 2^{p-1}\|w\|_F^p\|V(0)\|_F^p \left(\prod_{t=0}^{T-1}(1-\eta(t)\lambda)\right)^p + 2^{p-1}\|w\|_F^{2p}\lambda^{-p}\left(1-\prod_{t=0}^{T-1}(1-\eta(t)\lambda)\right)^p,\end{aligned}$$

and we conclude by the bound (3.7).

b) Using Hölder's inequality, the bound (3.3) and (5.2), as in (3.8) we have

$$\begin{aligned}\sup_{(x,y)\ \in\ \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x,V(T),W(T)),y,V(T),W(T))|^p &\le \|V(T)\|_F^p \|W(T)\|_F^p \qquad (5.4)\\ &\le 2^{p-1}\left(\|V(0)\|_F^{2p} + \|W(0)\|_F^{2p}\right)\prod_{t=0}^{T-1}(1+(1-\lambda)\eta(t))^{2p},\end{aligned}$$

and we conclude similarly to part (*a*). □

As a consequence of the above arguments, we have the following concentration inequality.

Corollary 5.4 *Suppose that Assumption 1 holds.*

a) Assume that $\eta_W(t) = 0$, $0 \le t < T$, *i.e.* $W(t)$ *remains frozen at* $W(t) := w \in \mathbb{R}^{d\times d_{\mathrm{out}}}$, $0 \le t \le T$. *For any* $\zeta \in (0,1)$, *with probability at least* $1-\zeta$ *we have the concentration inequality*

$$\begin{aligned}&\sup_{(x,y)\ \in\ \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x,V(T),w),y,V(T),w)|\\ &\quad\le \|w\|_F\left(\sqrt{d_{\mathrm{in}}}+\sqrt{d}+\sqrt{2\log\frac{2}{\zeta}}\right)\sqrt{\frac{\kappa}{d}}\prod_{t=0}^{T-1}(1-\eta(t)\lambda) + \frac{\|w\|_F^2}{\lambda}\left(1-\prod_{t=0}^{T-1}(1-\eta(t)\lambda)\right).\end{aligned}$$

b) If $\eta_W(t) = \eta_V(t) := \eta(t)$, $0 \le t < T$, *then for any* $\zeta \in (0,1)$, *with probability at least* $1-\zeta$ *we have the concentration inequality*

$$\begin{aligned}&\sup_{(x,y)\ \in\ \mathrm{Supp}\,(\rho)} |\nabla_x \ell_\lambda(f(x,V(T),W(T)),y,V(T),W(T))|\\ &\quad\le 3\kappa\left(2+\frac{d_{\mathrm{in}}}{d}+\frac{d}{d_{\mathrm{out}}}+\left(\frac{2}{d}+\frac{2}{d_{\mathrm{out}}}\right)\log\frac{4}{\zeta}\right)\prod_{t=0}^{T-1}(1+(1-\lambda)\eta(t))^2.\end{aligned}$$

Proof. *a*) This inequality follows from the bounds (3.6), (4.6) and (5.3).

b) This inequality follows from the bounds (3.8), (4.11) and (5.4). □

6 Numerical results

In this section we present the numerical simulations for the bounds derived in Propositions 4.1. For this, we consider

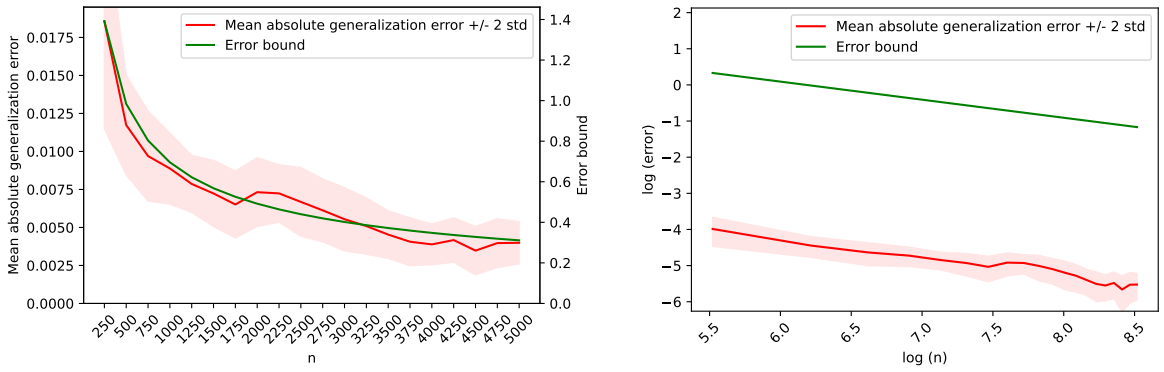
$$Y = \max\big(\min\big(\beta^\top X + \epsilon, 1\big), -1\big),$$

where X follows a uniform distribution on the 100-dimensional unit sphere S^{99}, β is a fixed point on S^{99}, and ϵ follows a standard normal distribution. Here, ρ is the corresponding distribution of (X, Y).

In the He initialization, we use a centered normal distribution with variance $1/500$ for every entry of the matrix $V(0)$. Subsequently, $V(t) \in \mathbb{R}^{100\times 1000}$ is updated according to (2.8), and $W(t) := w$ is frozen on the 1000-dimensional unit sphere.

We use the ReLU activation function $\sigma(x) := \max(0, x)$, the L^1 loss function $l(x, y) := |x - y|$, the regularization parameter $\lambda := 0.1$, the learning rate $\eta_V(t) = \eta_W(t) := 0.01$, $T := 300$ epochs, $n := 250, 500, \ldots, 5000$ samples with the batch size $k := n/10$. The simulation is repeated 20 times, and at each time we record the samples of the absolute generalization error $\big|\varepsilon_{gen}(n, V(T), w)\big|$.

The mean absolute generalization error in Proposition 4.1 is approximated using the mean absolute value of the recorded samples. Since the true loss $\mathcal{L}(V(T), w)$ in (2.6) is not available in closed form, it is approximated by Monte Carlo simulations with sample size 10^5.



(a) Dual scale. (b) Log scale.

Figure 1: Mean absolute value of $\varepsilon_{gen}(n, V(T), w)$ *vs.* error bound (4.1).

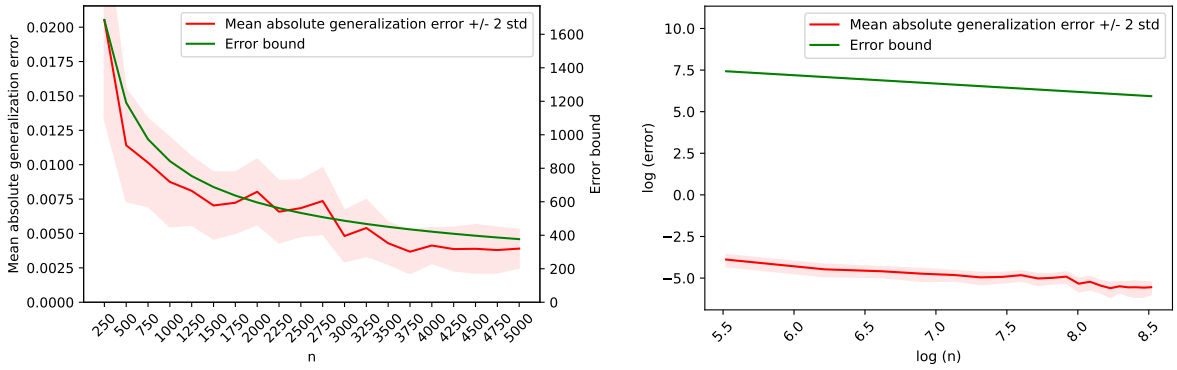
In Figure 1 we compare the mean absolute value of the generalization error to the uniform error bound (4.1) derived in Proposition 4.1-a). Figure 1-a) is plotted on a dual scale by matching the maximum (initial) values of the two curves to a same level on the graph.

In Table 1 we present the log-log linear regression displayed in Figure 1-b), which confirms the rate of $O(n^{-1/2})$ obtained in Proposition 4.1.

	Mean	Stdev	t-statistics	p-value	95% conf. interval
intercept	-1.1588	0.228	-5.094	0.000	(-1.637, -0.681)
slope	-0.5139	0.030	-17.345	0.000	(-0.576, -0.452)

Table 1: Log-log regression of the mean absolute generalization error.

Next, we run simulations without freezing $W(t)$, i.e. each element in $W(0)$ follows a centered normal distribution with variance 2 and subsequently updated according to (2.8). In Figure 2 we compare the mean absolute value of the generalization error and the error bound derived in Proposition 4.1$-b$) in the same setting as Figure 1.



(a) Dual scale. (b) Log scale.

Figure 2: Mean absolute value of $\varepsilon_{gen}(n, V(T), W(T))$ *vs.* error bound (4.2).

In Table 2 we present the log-log linear regression displayed in Figure 2-b), which confirms the rate of $O(n^{-1/2})$ obtained in Proposition 4.1.

	Mean	Stdev	t-statistics	p-value	95% conf. interval
intercept	-0.9745	0.292	-3.333	0.004	(-1.589, -0.360)
slope	-0.5366	0.038	-14.095	0.000	(-0.617, -0.457)

Table 2: Log-log regression of the mean absolute generalization error.

We note that although the constants $C_1(w, T)$, $C_2(2, T)$ in Figures 1 and 2 can be quite large, the $O(n^{-1/2})$ rate of decrease has been correctly identified. As the dimension-dependent

bounds in Proposition 5.1 are not sharp, the numerical simulations based on them are not presented.

Acknowledgement

We thank the anonymous referees for useful suggestions and corrections.

References

[ACS23] G. Aminian, S.N. Cohen, and Ł. Szpruch. Mean-field analysis of generalization errors, 2023. Preprint arXiv:2306.11623.

[ADH+19] S. Arora, S.S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[AZLL19] Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[CG19] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in Neural Information Processing Systems*, 32:10836–10846, 2019.

[DR17] G.K. Dziugaite and D.M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Preprint arXiv:1703.11008*, 2017.

[FCAO18] C. Finlay, J. Calder, B. Abbasi, and A. Oberman. Lipschitz regularized deep neural networks generalize and are adversarially robust, 2018. Preprint arXiv:1808.09540.

[FG15] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963.

[HRS16] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

[HZRS15] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[KKB17] K. Kawaguchi, L.P. Kaelbling, and Y. Bengio. Generalization in deep learning. Preprint arXiv:1710.05468, 28 pages, 2017.

[KR58] L.V. Kantorovič and G.S. Rubinšteĭn. On a space of completely additive functions. *Vestnik Leningrad. Univ.*, 13(7):52–59, 1958.

[LJ18] A.T. Lopez and V. Jog. Generalization error bounds using Wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2018.

[MMM19] S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Proceedings of the 32nd Annual Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1–77, 2019.

[MWZZ18] W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on Learning Theory*, pages 605–638. PMLR, 2018.

[NBS17] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *Preprint arXiv:1707.09564*, 2017.

[NDHR21] G. Neu, G.K. Dziugaite, M. Haghifam, and D.M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

[NLB+18] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2018.

[NTS15] B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

[PJL18] A. Pensia, V. Jog, and P.L. Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

[PSE22] S. Park, U. Simsekli, and M.A. Erdogdu. Generalization bounds for stochastic gradient descent via localized ϵ-covers. In *Advances in Neural Information Processing Systems*, volume 35, pages 2790–2802, 2022.

[RRT17] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

[RV10] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

[WDFC19] H. Wang, M. Diaz, J.C.S. Santos Filho, and F.P. Calmon. An information-theoretic view of generalization via Wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 577–581. IEEE, 2019.

[WLW+25] P. Wang, Y. Lei, D. Wang, Y. Ying, and D.-X. Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(1):1–45, 2025.

[XR17] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Preprint arXiv:1705.07809*, 2017.

[ZZB+22] Y. Zhang, W. Zhang, S. Bald, V. Pingali, and C. Chenand M. Goswami. Stability of SGD: Tightness analysis and improved bounds. In *Uncertainty in Artificial Intelligence*, pages 2364–2373. PMLR, 2022.