Optimal Risk Bounds of Stochastic Gradient Descent for Shallow ReLU Networks

Yuqing Liu[1,2] and Yunwen Lei[2]

[1]RPTU Kaiserslautern-Landau [2]The University of Hong Kong

NeurIPS 2026

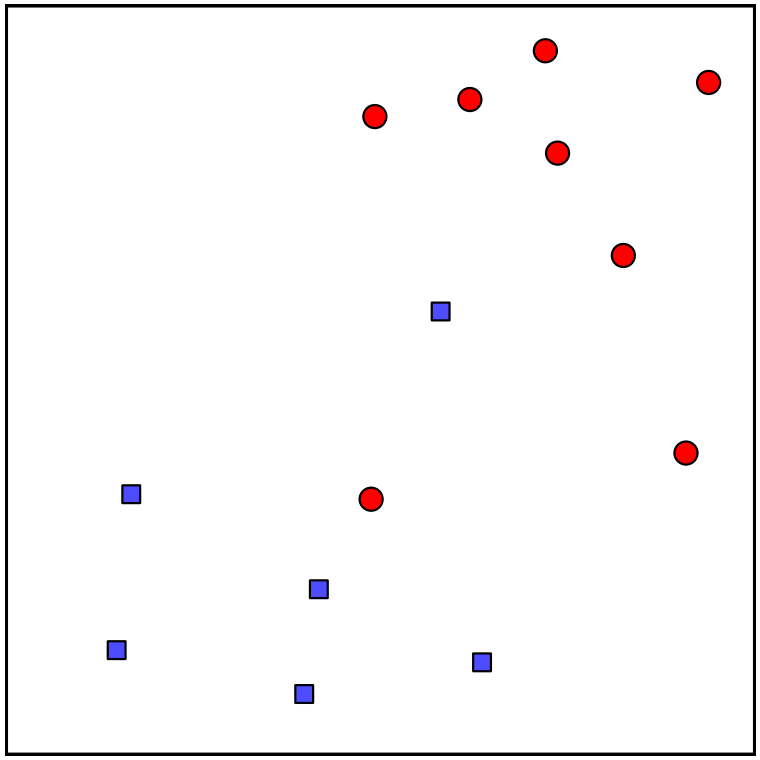
Outline

1. Background
2. Optimization and Generalization Analysis
3. Application to NTK Separable Data
4. Conclusion

Background

Binary Classification

Data: $S = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n \sim \rho$ defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{Y} = \{-1, +1\}$

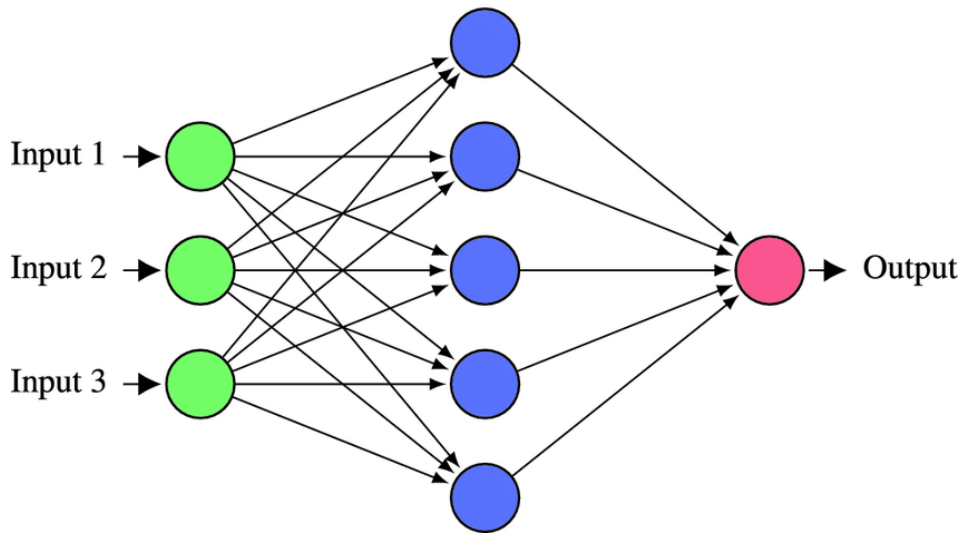


Learning with Shallow ReLU Network

- **Shallow Neural Network**:

$$\Phi(\mathbf{W}; \mathbf{x}) = \sum_{j=1}^{m} a_j \sigma(\mathbf{w}_j^\top \mathbf{x})$$

  - we fix $a_j \in \{\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\}$
  - we train the weight matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_m)$
  - $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is the ReLU activation, i.e., $\sigma(a) = \max\{a, 0\}$



Population and Empirical Risk

- Loss function: $f(\mathbf{W}; z) = \ell(y\Phi(\mathbf{W}; \mathbf{x}))$, where $\ell : \mathbb{R} \mapsto \mathbb{R}_+$.
  - Prediction is correct if $y\Phi(\mathbf{W}; \mathbf{x}) > 0$. We choose a decreasing function ℓ
- Population risk and Empirical risk

$$F_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{W}; \mathbf{z}_i), \quad F(\mathbf{W}) = \mathbb{E}_{\mathbf{z}\sim\rho}[f(\mathbf{W}; \mathbf{z})]$$

- We want to get a model $\mathbf{W}$ with a small population risk!

This is achieved by an optimization algorithm A, which outputs $A(S)$!

- empirical risk minimization: $A(S) = \arg\min_{\mathbf{W}} F_S(\mathbf{W})$
- regularized risk minimization: $A(S) = \arg\min_{\mathbf{W}} \left\{F_S(\mathbf{W}) + \lambda\|\mathbf{W}\|_2^2\right\}$
- gradient descent: $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \eta\nabla F_S(\mathbf{W}_t)$

GD requires n gradient computations and is not efficient!

Stochastic Gradient Descent

Symmetric Initialization: $\Phi(\mathbf{W}_1; \mathbf{x}) = 0, \forall \mathbf{x}$ For $j = 1, \ldots, m/2$, we sample

$$\mathbf{w}_{1,j} \sim N(0, I_d) \quad \text{and set} \quad \mathbf{w}_{1,j+m/2} = \mathbf{w}_{1,j},$$

and assign output weights in opposite pairs:

$$a_{j+m/2} = -a_j, \qquad |a_j| = \frac{1}{\sqrt{m}}.$$

Stochastic Gradient Descent (SGD)

for $t = 1, 2, \ldots$ **to** $T - 1$ **do**
$\quad \mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \eta \nabla f(\mathbf{W}_t; \mathbf{z}_{i_t}), \quad i_t \sim \text{Unif}(\{1, \ldots, n\})$

Subgradient

$$\mathbf{w}_{t+1,j} = \mathbf{w}_{t,j} - \eta a_j y_{i_t} \ell'(y_{i_t} \Phi(\mathbf{W}_t; \mathbf{x}_{i_t})) \mathbb{I}_{[\mathbf{w}_{t,j}^\top \mathbf{x}_{i_t} \geq 0]} \mathbf{x}_{i_t}.$$

Error Decomposition

We decompose $F(\mathbf{W}_T)$ into

$$F(\mathbf{W}_T) = \underbrace{\big(F(\mathbf{W}_T) - F_S(\mathbf{W}_T)\big)}_{\text{generalization gap}} + \underbrace{F_S(\mathbf{W}_T)}_{\text{optimization error}} .$$

- We leverage Rademacher complexity to study generalization gap
- We use optimization theory to study optimization error

Optimization and Generalization Analysis

Weak Convexity of SNNs

Assumption on ℓ

- ℓ is convex and L-smooth, i.e., $|\ell'(a) - \ell'(a')| \leq L|a - a'|$ for all $a, a' \in \mathbb{R}$.
- Self-bounding property:$|\ell'(a)| \leq \widetilde{L}\ell(a)$ for some $\widetilde{L} > 0$
- Lipschitzness: $|\ell'(a)| \leq G$ for some $G > 0$.

Examples include the logistic loss $\ell(a) = \log(1 + \exp(-a))$, polynomially-tailed loss and sub-exponential loss (Schliserman and Koren, 2022).

Weak Convexity of SNNs (Lei et al., 2026)

For any $\mathbf{W}, \mathbf{W}'$ close to $\mathbf{W}_1$, we have

$$F_S(\mathbf{W}) - F_S(\mathbf{W}') \geq \langle \mathbf{W} - \mathbf{W}', \nabla F_S(\mathbf{W}') \rangle - \frac{\|\mathbf{W} - \mathbf{W}'\|_2 F_S(\mathbf{W}')}{\sqrt{m}}. \tag{1}$$

Optimization Error

Consider a reference $\mathbf{W}^*$ and introduce

$$\mathfrak{C}_S(\mathbf{W}^*) := \eta T F_S(\mathbf{W}^*) + \|\mathbf{W}_1 - \mathbf{W}^*\|_2^2. \tag{2}$$

Convergence rate

If $m \gtrsim \mathfrak{C}_S^2(\mathbf{W}^*)$ and $\eta \lesssim 1/L$, then with high probability

$$\|\mathbf{W}_t - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}_S(\mathbf{W}^*) \quad \text{and} \quad \eta \sum_{k=1}^{t-1} F_S(\mathbf{W}_k) \leq \mathfrak{C}_S(\mathbf{W}^*). \tag{3}$$

Furthermore, there holds

$$\|\mathbf{W}_t - \mathbf{W}_1\|_{2,\infty} \leq \mathfrak{C}_S(\mathbf{W}^*)/\sqrt{m}, \quad \forall t \in [T]. \tag{4}$$

Generalization Analysis

Rademacher Complexity

- Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher variables (taking only values ± 1, with equal probability)
- The Rademacher complexity of $\mathcal{F}$ is defined as ($S = \{z_i\}$)

$$\mathfrak{R}_S(\mathcal{F}) := \mathbb{E}_\epsilon \Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(\mathbf{z}_i) \Big]. \tag{5}$$

- Convergence analyses show that $\|\mathbf{W}_t - \mathbf{W}^*\|_2^2 \leq \mathfrak{C}_S(\mathbf{W}^*)$
- It suffices to consider the localized function class

$$\mathcal{G} := \{\mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \|\mathbf{W} - \mathbf{W}_1\|_2 \leq R\}. \tag{6}$$

Rademacher Complexity Bound

(Lei and Xie, 2026)

Let $R > 0$ and $\mathcal{G} := \{\mathbf{x} \mapsto \Phi(\mathbf{W}; \mathbf{x}) : \|\mathbf{W} - \mathbf{W}_1\|_2 \leq R\}$. Then,

$$\mathfrak{R}_S(\mathcal{G}) \leq R/\sqrt{n}.$$

Initialization-dependent complexity bound!

Risk Bounds

Risk Bounds

Assume $m \gtrsim \mathfrak{C}_S^2(\mathbf{W}^*)$ and $\mathfrak{C}_S(\mathbf{W}^*)\|\mathbf{W}^* - \mathbf{W}_1\|_{2,\infty} \lesssim 1$. If $\eta \lesssim 1/L$, then with probability at least $1-\delta$ we have

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{W}_t) = \widetilde{O}\Big(\frac{\mathfrak{C}_S(\mathbf{W}^*) + \log(1/\delta)}{n}\Big).$$

- If $\mathfrak{C}_S(\mathbf{W}^*) = \widetilde{O}(1)$, then we get fast rates of order $\widetilde{O}(1/n)$.
- If $\mathfrak{C}_S(\mathbf{W}^*)$ grows as logarithmic function of n, then we only require logartihmic width!

Application to NTK Separable Data

Application: NTK Separability

The idea of NTK is to use first-order approximation

$$\Phi(\mathbf{W};\mathbf{x}) \approx \Phi(\mathbf{W}_1;\mathbf{z}) + \langle \mathbf{W}-\mathbf{W}_1, \nabla\Phi(\mathbf{W}_1;\mathbf{x})\rangle = \langle \mathbf{W}-\mathbf{W}_1, \nabla\Phi(\mathbf{W}_1;\mathbf{x})\rangle.$$

NTK Separability

(Taheri and Thrampoulidis, 2023; Ji and Telgarsky, 2019)

Assume there exists $\mathbf{W}_* \in \mathbb{R}^{nd}$ with $\|\mathbf{W}_*\|_2 = 1$ and $\|\mathbf{W}_*\|_{2,\infty} \lesssim \frac{1}{\sqrt{m}}$ such that the following inequality holds

$$y_i\big\langle \nabla\Phi(\mathbf{W}_1;\mathbf{x}_i), \mathbf{W}_*\big\rangle \geq \gamma, \quad \forall i \in [n].$$

Risk Bounds

Let ℓ be the logistic loss and the NTK separability assumption holds. If $m = \widetilde{\Omega}(1/\gamma^8)$ and $\eta T \asymp n$, then with probability at least $1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[F(\mathbf{W}_t)] = \widetilde{O}\Big(\frac{1}{n\gamma^2}\Big).$$

Comparison

Work	Method	Bound	Activation	Width	GC
Nitanda et al. (2019)	GD	$\widetilde{O}\big(1/(n^{\frac{1}{4}}\gamma^{\frac{1}{2}})\big)$	Smooth	poly	$n^{\frac{3}{2}}/\gamma$
Ji and Telgarsky (2019)	GD	$\widetilde{O}\big(1/(\sqrt{n}\gamma^2)\big)$	ReLU	log	n^2
Taheri and Thrampoulidis (2024)	GD	$\widetilde{O}(1/(n\gamma^2))$	Smooth	log	n^2
Lei et al. (2026)	GD	$\widetilde{O}(1/(n\gamma^2))$	ReLU	log	n^2
Ours	SGD	$\widetilde{O}(1/(n\gamma^2))$	ReLU	log	n

Table: Comparison of the risk bounds under NTK separability condition with margin γ. "GC" means the gradient complexity.

Conclusion

Conclusion

Conclusion

- We study optimization and generalization for SGD on shallow ReLU networks
- We show SGD can achieve optimal rates with $O(n)$ gradient computations for SNN

Future Work

- How to consider NNs with other structures?
- How to consider deep neural networks?
- How to incorporate the approximation error analysis?

References I

Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2019.

Y. Lei and Y. Xie. Towards initialization-dependent and non-vacuous generalization bounds for overparameterized shallow neural networks. *arXiv preprint arXiv:2604.00505*, 2026.

Y. Lei, P. Wang, Y. Ying, and D.-X. Zhou. Optimization and generalization of gradient descent for shallow relu networks with minimal width. *Journal of Machine Learning Research*, 27(34):1–35, 2026.

A. Nitanda, G. Chinot, and T. Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

M. Schliserman and T. Koren. Stability vs implicit bias of gradient methods on separable data and beyond. *arXiv preprint arXiv:2202.13441*, 2022.

H. Taheri and C. Thrampoulidis. On generalization of decentralized learning with separable data. In *International Conference on Artificial Intelligence and Statistics*, pages 4917–4945. PMLR, 2023.

H. Taheri and C. Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *Journal of Machine Learning Research*, 25(156):1–41, 2024.

Thank you!