Learning a Single Neuron for Non-monotonic Activation Functions

Lei Wu
School of Mathematical Sciences, Peking University
leiwu@math.pku.edu.cn

Abstract

We study the problem of learning a single neuron $\mathbf{x} \mapsto \sigma(\mathbf{w}^T\mathbf{x})$ with gradient descent (GD). All the existing positive results are limited to the case where σ is monotonic. However, it is recently observed that non-monotonic activation functions outperform the traditional monotonic ones in many applications. To fill this gap, we establish learnability without assuming monotonicity. Specifically, when the input distribution is the standard Gaussian, we show that mild conditions on σ (e.g., σ has a dominating linear part) are sufficient to guarantee the learnability in polynomial time and polynomial samples. Moreover, with a stronger assumption on the activation function, the condition of input distribution can be relaxed to a non-degeneracy of the marginal distribution. We remark that our conditions on σ are satisfied by practical non-monotonic activation functions, such as SiLU/Swish and GELU. We also discuss how our positive results are related to existing negative results on training two-layer neural networks.

1 Introduction

Neural networks play a fundamental role in deep learning, which has achieved unprecedented successes in many applications, such as computer vision, natural language processing, and scientific computing. Despite tremendous efforts devoted, theoretical understandings of learning neural networks are still rather unsatisfactory because of the inherent non-convexity.

In this paper, we consider the simplest setting: learning a single neuron $\mathbf{x} \mapsto \sigma(\mathbf{w}^T\mathbf{x})$, where $\mathbf{w}$ is the parameter to be learned and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a fixed activation function. This problem has been widely studied previously (see the related work section below for more details) and plays an important role in understanding general neural networks, e.g., the superiority of neural networks over kernel methods (Yehudai and Shamir, 2019) and the hardness of training neural networks (Shamir, 2018; Livni et al., 2014b).

We assume that inputs are drawn from an underlying distribution $\mathcal{D}$ and the labels are generated by some unknown neuron $\mathbf{x} \mapsto \sigma(\mathbf{w}^{*T}\mathbf{x})$, i.e., the realizable case. As such, the population risk is given by

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}_{\mathbf{x}}[\frac{1}{2}(\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))^2].$$

In practice, only finite training samples $\{\mathbf{x}_i\}_{i=1}^n$ are available, and we instead minimize the empirical risk

$$\hat{\mathcal{R}}_n(\mathbf{w}) := \frac{1}{2n}\sum_{i=1}^n(\sigma(\mathbf{w}^T\mathbf{x}_i) - \sigma(\mathbf{w}^{*T}\mathbf{x}_i))^2.$$

Despite the simplicity, this problem is still highly non-trivial due to the non-convexity.

To attack this problem, existing works (Frei et al., 2020; Mei et al., 2018; Yehudai and Shamir, 2020; Tian, 2017) all assume σ to be monotonic, for which

$$\begin{aligned} G(\mathbf{w}) &= \langle \nabla\mathcal{R}(\mathbf{w}), \mathbf{w} - \mathbf{w}^*\rangle \\ &= \mathbb{E}_{\mathbf{x}}[(\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\mathbf{w}^T\mathbf{x})(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})] \\ &\geq 0. \end{aligned} \tag{1}$$

The above inequality implies two critical facts:

- All local minima are global minima if σ is monotonic. Note that (1) implies that $\mathrm{d}\mathcal{R}(\mathbf{w}^* + \beta(\mathbf{w} - \mathbf{w}^*))/\mathrm{d}\beta \geq 0$ for any $\mathbf{w} \in \mathbb{R}^d$. This suggests that starting from any $\mathbf{w}$, we can find a loss-decreasing curve connecting $\mathbf{w}$ to $\mathbf{w}^*$. Therefore, there is no bad local minima; see also (Auer et al., 1996, Theorem 5.1).
- The gradient at every point points to a direction of decreasing $\|\mathbf{w}_t - \mathbf{w}^*\|$ since $\mathrm{d}\|\mathbf{w}_t - \mathbf{w}^*\|^2/\mathrm{d}t = -G(\mathbf{w}_t) \leq 0$.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Moreover, Frei et al. (2020); Mei et al. (2018); Yehudai and Shamir (2020); Tian (2017) impose stronger assumptions on the activation function and input distribution to ensure the lower boundedness of $G(\mathbf{w})$, thereby guaranteeing the convergence.

However, recent practical evidence (Devlin et al., 2018; Radford et al., 2018, 2019; Sitzmann et al., 2020) shows that in many applications, non-monotonic activation functions, e.g., SiLU and GELU, are superior to the traditional monotonic ones (See the related work section below for more details). This motivates us to analyze the case where σ is non-monotonic. Note that for general activation functions, the risk landscape may have a large number of bad local minima (Brady et al., 1989; Ros et al., 2019). Moreover, Shamir (2018); Livni et al. (2014a) show that if σ is highly oscillated, gradient-based methods suffer from the curse of dimensionality in learning a single neuron. These suggest that some conditions (beyond the monotonicity) on σ must be imposed to ensure learnability.

Our **main contributions** are be summarized as follows.

- We first consider activation functions that are increasing in $[0, \infty)$ and satisfy $\inf_{0<z<\alpha} \sigma'(z) \geq \gamma, \inf_{z_1 \geq 0, z_2 \leq 0} \sigma'(z_1)\sigma'(z_2) \geq -\zeta^2$ for some constants $\alpha, \gamma, \zeta > 0$. We prove in Theorem 3.3 that if the input distribution $\mathcal{D}$ is sufficiently "spread", GD converges to a global minimum exponentially fast as long as γ is relatively larger than ζ. This condition essentially means that the monotonic component of the activation function dominates.

- Then fine-grained analyses of the GD dynamics are provided for the case where the input distribution is the standard Gaussian. In this case, the condition on σ can be further relaxed. Specifically, we consider two settings: GD with zero initialization and Riemannian GD with a random initialization.

  The analysis of zero initialization relies on the observation that the gradient at zero points to the ground truth $\mathbf{w}^*$ and therefore, the original problem can be reduced to minimizing a one-dimensional risk. The same observation has been exploited in Tian (2017); Soltanolkotabi (2017); Kalan et al. (2019) for the specific ReLU activation function, whereas we show that it holds for general activation functions. In addition, we identify further conditions on σ to ensure that this one-dimensional risk has a benign landscape, thereby guaranteeing the convergence of GD.

  For random initialization, we consider the Riemannian GD with $\mathbf{w}_t \in \mathbb{S}^{d-1}$. For this case, we show that the population risk has a simple closed-form analytic expression (see Lemma 4.6), which depends on σ only through the Hermite coefficients $\{\hat{\sigma}_k\}_k$. Here $\hat{\sigma}_k = \mathbb{E}_{z\sim N(0,1)}[h_k(z)\sigma(z)]$, where h_k is the k-th probabilistic Hermite polynomial. By using this analytic expression, we provide a thorough study of how the decay of Hermite coefficients affects the property of risk landscape and the convergence of Riemannian GD. In particular, we establish in Proposition 4.9 a high-probability convergence to the global minimum by assuming that the linear component of the activation function, i.e., $\hat{\sigma}_1 = \mathbb{E}[z\sigma(z)]$, is sufficiently large. On the other hand, if $\hat{\sigma}_1 = 0$, we construct a counterexample in Lemma 4.8 , for which the Riemannian GD converges to a bad local minimum with a probability close to $1/2$. These together partially explain the wide use of ReLU and its variants in practice since they all have dominating linear components.

- Lastly, we consider the finite sample case. In Proposition 5.1, we establish the closeness between the empirical landscape and the population landscape using the theory of empirical process. With these closeness results, we can convert our positive results of the population GD to the empirical GD (see Proposition 5.3 and Proposition 5.4). In particular, in all the settings, we show that GD can learn the ground truth using only polynomial samples and polynomial time.

Note that, for all the settings we considered, the conditions are satisfied by all the popular activation functions used in practice, including the non-monotonic ones.

1.1 Related work

Non-monotonic activation functions Ramachandran et al. (2017) uses the neural architecture search (NAS) method to search the best activation function for classifying the CIFAR-10 data. It is discovered that the non-monotonic Swish function, $\sigma_{\text{swish}}(z) = z\sigma_{\text{sigmoid}}(\beta z)$ with $\beta > 0$, performs the best. In particular, when $\beta = 1$, it becomes the sigmoid-weighted linear unit (SiLU) (Elfwing et al., 2018). Recently, SiLU/Swish also show extraordinary performances on many other applications, such as adversarial training (Xie et al., 2020), model compression (Tessera et al., 2021), etc. Gaussian error linear unit (GELU) (Hendrycks and Gimpel, 2016) is another popular non-monotonic activation function and has the similar properties to SiLU/Swish. GELU has been widely applied in large-scaled pre-trained language models, such as GPT/GPT-2 (Radford

et al., 2018, 2019), BERT (Devlin et al., 2018), and most other Transformer-based (Vaswani et al., 2017) models (Liu et al., 2019). In addition, non-monotonic activations also see lots of applications in solving scientific computing problems. For these problems, one may need to restrict the activation function to be periodic or compactly supported, where activation functions are always non-monotonic, e.g., Sitzmann et al. (2020); Li et al. (2020); Liang et al. (2021); Chen et al. (2020) to name a few. Therefore, understanding the learning of neural networks with non-monotonic activation functions becomes crucially important.

Learning a single neuron under the realizable setting A single neuron is essentially the same as the traditional generalized linear models (GLMs) and single-index models (SIMs). For GLMs, σ is usually a nondecreasing function, such as the sigmoid function for the logistic binary classification. Except for the monotonicity, SIMs further assume that σ is unknown, to be learned from data. When σ is nondecreasing, $\sigma^{-1}(\cdot)$ can be defined. Hence, this problem can be efficiently solved by fitting the linear function: $\mathbf{x} \mapsto \sigma^{-1}(y)$. Indeed, the algorithms for GLMs and SIMs are based on this observation (Kalai and Sastry, 2009; Kakade et al., 2011), which obviously does not hold if σ is non-monotonic.

In contrast, the GD method is applicable irrespective of the monotonicity of σ. However, the theoretical understanding of GD is non-trivial because of the non-convexity of the risk landscape. When σ is strictly monotonic and $\mathcal{D}$ is non-degenerate, there is only one critical point: $\mathbf{w} = \mathbf{w}^*$ demonstrated previously. However, for general activation functions and general input distribution, there may exist many bad local minima and saddle points (Brady et al., 1989; Ros et al., 2019). Moreover, the empirical landscape can be much more complex. For instance, even when σ is strictly monotonic, there may exist many bad critical points when $n/d \le c_\sigma$ for some constant $c_\sigma > 0$. Using Kac-Rice replicated method (Ros et al., 2019) from theoretical physics, Maillard et al. (2020) provides an explicit characterization of the critical points in the thermodynamics limit: $n, d \to \infty$ with $n/d \to \alpha > 1$. Lastly we mention that for the non-realizable case, there may exist bad local minima even if σ is strictly monotonic (Auer et al., 1996).

Apart from the above landscape analyses, the hardness of learning can be substantiated for the case where σ is periodic. Specifically, Kearns (1998); Blum et al. (1994); Diakonikolas et al. (2020); Malach and Shalev-Shwartz (2020) shows that learning the parity function: $\{0,1\}^d \mapsto \{-1,1\} : f_{\mathbf{v}}(x) = (-1)^{\mathbf{v}^T x}$ suffers from the curse of dimensionality. The parity function is essentially a single neuron with $\sigma(z) = (-1)^z$ and $\mathcal{D} = \mathrm{Unif}(\{0,1\}^d)$. Shamir (2018) later extends the above understanding to general periodic activation functions and $\mathcal{D} = \mathcal{N}(0, I_d)$. Our results are consistent with these negative results since the constants in our bounds are exponentially large for these periodic activations. Moreover, our results imply that GD can learn a single neuron efficiently as long as the activation function does not oscillate too much.

The previous positive results are summarized as follows. Mei et al. (2018); Oymak and Soltanolkotabi (2019) show that the empirical GD can return a good approximation of $\mathbf{w}^*$. However, the analysis requires σ to be *strictly* monotonic. Yehudai and Shamir (2020) later shows that as long as the input distribution is sufficiently "spread", a weak monotonicity condition on σ is sufficient to guarantee a constant-probability convergence for a random initialization. A similar analysis for the agnostic setting is provided in Frei et al. (2020). For the specific ReLU activation function and standard Gaussian input distribution, Tian (2017) proves the exponential convergence of the population GD. Soltanolkotabi (2017); Kalan et al. (2019) considered a similar setting but for the empirical GD. Our work differentiates from these works by removing the requirement of monotonicity.

Another line of related research is phase retrieval (Sun et al., 2018; Tan and Vershynin, 2019; Chen et al., 2019), which fits our setting with $\sigma(z) = z^2$ or $\sigma(z) = |z|$. In phase retrieval, the activation function is indeed non-monotonic, but the analysis is specific to those activation functions. By contrast, our analysis holds for more general non-monotonic activation functions, including the popular SiLU/Swish and GELU.

2 Preliminaries

Notation Let I_d denote the $d \times d$ identity matrix. We use bold-faced letters to denote vectors. For a vector $\mathbf{w}$, let w_i denote the i-th coordinate, $\|\mathbf{w}\|^2 = \sum_i w_i^2$. For $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, we use $\theta(\mathbf{w}, \mathbf{v}) = \arccos(\frac{\mathbf{w}^T\mathbf{v}}{\|\mathbf{w}\|\|\mathbf{v}\|})$ to denote the angle between $\mathbf{w}$ and $\mathbf{v}$. Let $\mathbb{S}^{d-1} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| = 1\}$. We use $X \lesssim Y$ to denote $X \le CY$ for some absolute constant $C > 0$. We will occasionally use $\tilde{O}(\cdot)$ to hide logarithmic factors.

For simplicity, we assume that $\|\mathbf{w}^*\| = 1$ and $\sigma(0) = 0$, otherwise, we can replace $\sigma(z)$ with $\sigma(z/\|\mathbf{w}^*\|) - \sigma(0)$ without changing the risk landscape. The gradient of population risk can be written as

$$\nabla \mathcal{R}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}}\left[(\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{x}\right]. \quad (2)$$

When $\mathbf{w} \neq 0$, as long as the marginal distribution $\mathbf{w}^T\mathbf{x}$ is not singular, (2) holds if σ is differentiable almost

everywhere, since changing the value of $\sigma'(z)$ at a set of measure zero does not affect the expectation. When $\mathbf{w} = 0$ and $\sigma(\cdot)$ is not differentiable at the origin, we will explicitly specify the value of $\sigma'(0)$, e.g., $\sigma'(0) = 1$ for ReLU.

For the training method, we focus on the GD flow $\dot{\mathbf{w}}_t = -\nabla\mathcal{R}(\mathbf{w}_t)$, which is GD with an infinitesimal learning rate. Extending the results of GD flow to standard GD and stochastic gradient descent for learning a single neuron is straightforward; we refer to Yehudai and Shamir (2020) for some examples. Throughout this paper, we will use GD to denote GD flow for simplicity.

For non-monotonic activation functions, we are particularly interested in the *self-gated family*:

$$\sigma_\beta(z) = z\phi(\beta z), \tag{3}$$

where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is nondecreasing and satisfies that $\phi(-\infty) = 0, \phi(+\infty) = 1$. As $\beta \to \infty$, σ_β converges to ReLU. SiLU/Swish corresponds to the case that σ is the sigmoid function. GELU corresponds to the case where ϕ is the cumulative density function of $\mathcal{N}(0, 1)$

3 A General Result

In this section, we make the following assumption.

Assumption 1. The following holds for some fixed $\alpha, \beta, \gamma, \zeta, \tau > 0$:

- **Input distribution:** (1) $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^T] \leq \tau I_d$. (2) For any $\mathbf{w} \neq \mathbf{v} \in \mathbb{S}^{d-1}$, let $\mathcal{D}_{\mathbf{w},\mathbf{v}}$ denote the marginal distribution of $\mathbf{x}$ on $\mathrm{span}\{\mathbf{w}, \mathbf{v}\}$ (as a distribution over $\mathbb{R}^2$). Let $p_{\mathbf{w},\mathbf{v}}$ denote the density function of $\mathcal{D}_{\mathbf{w},\mathbf{v}}$. Assume $\inf_{\boldsymbol{z}\in\mathbb{R}^2:\|\boldsymbol{z}\|\leq\alpha} p_{\mathbf{w},\mathbf{v}}(\boldsymbol{z}) \geq \beta$.
- **Activation:** σ is increasing in $[0, \infty)$ and $\inf_{z_1\geq 0, z_2\leq 0} \sigma'(z_1)\sigma'(z_2) \geq -\zeta^2, \sup_{0<z<\alpha} \sigma'(z) \geq \gamma$.

This assumption is a modification of (Yehudai and Shamir, 2020, Assumption 4.1). The difference is that (1) σ is allowed to be non-monotonic in $(-\infty, 0]$ and we further assume the second-order moment of $\mathcal{D}$ to be bounded. The assumption on activation functions covers the popular self-gated family and excludes the hard examples where the activation function is periodic. The assumption on $\mathcal{D}$ is quite general and covers, for instance, log-concave distributions like Gaussian and uniform distributions with $\alpha, \beta, \tau = O(1)$.

Proposition 3.1. *Let $\theta(\mathbf{w}, \mathbf{w}^*)$ be the angle between $\mathbf{w}$ and $\mathbf{w}^*$. For any $\delta \in (0, \pi)$, let $c_\delta = \sin^3(\delta/4)/(8\sqrt{2})$. Under Assumption 1, for any $\mathbf{w} \in \mathbb{R}^d$ that satisfies $\theta(\mathbf{w}, \mathbf{w}^*) \leq \pi - \delta$, it holds that $\langle\nabla\mathcal{R}(\mathbf{w}), \mathbf{w} - \mathbf{w}^*\rangle \geq \lambda\|\mathbf{w} - \mathbf{w}^*\|^2$, where*

$$\lambda = (\gamma^2 + \zeta^2)\beta\alpha^4 c_\delta - \tau\zeta^2.$$

This proposition implies that the gradient $\nabla\mathcal{R}(\mathbf{w})$ provides a good direction for convergence as long as $\theta(\mathbf{w}, \mathbf{w}^*)$ is relatively small. In particular, $\lambda > 0$ for any $\delta > 0$ if $\zeta = 0$, and this corresponds to the monotonic case. In general, if $\gamma^2\beta\alpha^4 c_\delta \geq \tau\zeta^2$, we have $\lambda \geq \zeta^2\beta\alpha^4 c_\delta$. This condition means that the monotonic part of σ dominates the non-monotonic part in the sense that $\frac{\gamma^2}{\zeta^2} \geq \frac{\tau}{c_\delta\beta\alpha^4}$. When $\mathcal{D} = \mathcal{N}(0, I_d)$, it is easy to verify that this condition is satisfied by the popular SiLU/Swish and GELU activations. The proof of Proposition 3.1 is presented in Appendix A, which is modified from the proof of (Yehudai and Shamir, 2020, Theorem 4.2).

3.1 Convergence

In this section, let $\delta_t = \pi - \theta(\mathbf{w}_t, \mathbf{w}^*)$. We explicitly write $\lambda(\delta_t) = \lambda$ to emphasize the dependence on the angle $\theta(\mathbf{w}_t, \mathbf{w}^*)$. Then, Proposition 3.1 implies that $d\|\mathbf{w}_t - \mathbf{w}^*\|^2/dt \leq -\lambda(\delta_t)\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq 0$. By the definition in Proposition 3.1, we have $\lambda(\delta_t) \leq 0$ when $\delta_t = 0$. Therefore, for guaranteeing the convergence, we need to ensure that $\mathbf{w}_t$ always stay in a region where $\delta_t = \pi - \theta(\mathbf{w}_t, \mathbf{w}^*)$ is significantly large.

Intuition. The decreasing of $\|\mathbf{w}_t - \mathbf{w}^*\|$ does not alway imply the decreasing of $\theta(\mathbf{w}_t, \mathbf{w}^*)$. Yehudai and Shamir (2020) shows that $\theta(\mathbf{w}_t, \mathbf{w}^*)$ may increase and consequently $\lambda(\delta_t)$ decreases during the training. Let $H_+ = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^T\mathbf{w}^* \geq 0\}$. Obviously, $\theta(\mathbf{w}, \mathbf{w}^*) \leq \pi/2$ for any $\mathbf{w} \in H_+$. The following lemma formalizes the preceding intuition.

Lemma 3.2. *If $\|\mathbf{w} - \mathbf{w}^*\| < 1$, then $\theta(\mathbf{w}, \mathbf{w}^*) < \frac{\pi}{2}$.*

Proof. $\|\mathbf{w} - \mathbf{w}^*\|^2 = 1 - 2\mathbf{w}^T\mathbf{w}^* + \|\mathbf{w}\|^2 < 1$ implies that $\mathbf{w}^T\mathbf{w}^* \geq \|\mathbf{w}\|^2 > 0$. Hence, $\theta(\mathbf{w}, \mathbf{w}^*) < \frac{\pi}{2}$. □

Hence, if $\|\mathbf{w}_0 - \mathbf{w}^*\| < 1$, the decreasing of $\|\mathbf{w}_t - \mathbf{w}^*\|$ can ensure that $\|\mathbf{w}_t - \mathbf{w} * \| < 1$ for all $t \geq 0$. Consequently, $\delta_t = \pi - \theta(\mathbf{w}_t, \mathbf{w}^*) > \frac{\pi}{2}$ and $\lambda(\delta_t) > \lambda(\frac{\pi}{2})$ for any $t \geq 0$.

Theorem 3.3. *Suppose that Assumption 1 holds and $\lambda(\frac{\pi}{2}) > 0$. consider the random initialization $\mathbf{w}_0 \sim \mathcal{N}(0, \eta^2 I_d)$ with $\eta \leq \frac{1}{\sqrt{2d}}$. Then, with probability at least $\frac{1}{2} - \frac{1}{4}\eta d - 1.2^{-d}$ we have $\|\mathbf{w}_0 - \mathbf{w}^*\| \leq 1 - 2\eta^2 d$ and*

$$\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq e^{-\lambda(\frac{\pi}{2})t}.$$

This theorem provides a constant probability convergence. Note that $\lambda(\frac{\pi}{2}) = (\gamma^2 + \zeta^2)\beta\alpha^4 c_{\frac{\pi}{2}} - \tau\zeta^2 > 0$

means that σ has a dominated monotonic part. The specific choice of the variance of the random initialization can guarantee that $\|\mathbf{w}_0 - \mathbf{w}^*\| < 1$ holds with a constant probability (close to 1/2). The proof is presented in Appendix A.

4 Fine-Grained Analysis for Gaussian Inputs

In this section, we provide a fine-grained analysis of the risk landscape and the convergence of GD for the case of $\mathcal{D} = \mathcal{N}(0, I_d)$. The main message is that the conditions on $\sigma(\cdot)$ can be further relaxed. Similar results can be straightforward extended to other spherically symmetric distribution, e.g., Unif($\mathbb{S}^{d-1}$).

4.1 Zero Initialization

We first study GD with zero initialization. The analysis mainly relies on the following observation.

Lemma 4.1. *$\nabla\mathcal{R}(\beta\mathbf{w}^*) = -r'_\sigma(\beta)\mathbf{w}^*$, where r'_σ is the derivative of $r_\sigma : \mathbb{R} \to \mathbb{R}$ given by*

$$r_\sigma(\beta) = \frac{1}{2}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[(\sigma(\beta z) - \sigma(z))^2].$$

Proof. Let $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)^T \in \mathbb{R}^{d\times d}$ be an orthonormal matrix with $\mathbf{v}_1 = \mathbf{w}^*$. Let $\tilde{\mathbf{x}} = V\mathbf{x}$. $\tilde{\mathbf{x}} \sim \mathcal{N}(0, I_d)$ and $\mathbf{x} = V^T\tilde{\mathbf{x}} = \sum_{j=1}^d \tilde{x}_j\mathbf{v}_j$. Then,

$$\begin{aligned}
\nabla\mathcal{R}(\beta\mathbf{w}^*) &= \mathbb{E}_{\mathbf{x}}[(\sigma(\beta\mathbf{w}^{*T}\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\beta\mathbf{w}^{*T}\mathbf{x})\mathbf{x}] \\
&= \mathbb{E}_{\tilde{\mathbf{x}}}[(\sigma(\beta\tilde{x}_1) - \sigma(\tilde{x}_1))\sigma'(\beta\tilde{x}_1)\sum_{j=1}^d \mathbf{v}_j\tilde{x}_j] \\
&= \mathbb{E}_{\tilde{x}_1}[(\sigma(\beta\tilde{x}_1) - \sigma(\tilde{x}_1))\sigma'(\beta\tilde{x}_1)\tilde{x}_1]\mathbf{v}_1 \\
&:= -r'_\sigma(\beta)\mathbf{w}^*,
\end{aligned}$$

where the third equality is due to $\mathbb{E}[h(\tilde{x}_1)\tilde{x}_j] = 0$ for any $j \neq 1$. □

This lemma implies that $\nabla\mathcal{R}(\mathbf{w})$ at the line $\{\mathbf{w} = \beta\mathbf{w}^* : \beta \in \mathbb{R}\}$ exactly points to $\mathbf{w}^*$ (maybe up to a sign). Therefore, GD starting zero will always stay on this line. Note that (Tian, 2017; Soltanolkotabi, 2017; Kalan et al., 2019) have made the same observation but only for the specific ReLU activation.

Proposition 4.2. *Denote by $\mathbf{w}_t$ the GD solution that starts from $\mathbf{w}_0 = 0$. Then, $\mathbf{w}_t = \beta_t\mathbf{w}^*$ and β_t is the GD solution that minimizes $r_\sigma(\cdot)$, i.e., $\dot{\beta}_t = -r'_\sigma(\beta_t)$ with $\beta_0 = 0$.*

The proof is a straightforward application of Lemma 4.1. It is implied that the GD starting from 0 is equivalent to an one-dimensional GD that minimizes $r_\sigma(\cdot)$. In particular, $\beta = 1$ corresponds to the true solution. As a result, to ensure the convergence of GD, we only need $r_\sigma(\cdot)$ to have a nice landscape in $[0, 1+\delta]$ for some $\delta > 0$. Shown in Figure 1 are the landscapes of $r_\sigma(\cdot)$ for various commonly-used activation functions. On can see that for all the cases, $r_\sigma(\cdot)$ is monotonically decreasing in $[0, 1]$, which implies that GD can converge to the global minimum $\beta = 1$. Taking ReLU as a concrete example, we have

$$\begin{aligned}
r_\sigma(\beta) &= \frac{1}{2}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[|\sigma(\beta z) - \sigma(z)|^2] \\
&= \frac{(\beta-1)^2}{2}\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)^2] = \frac{(\beta-1)^2}{4}.
\end{aligned}$$

This implies that β_t converges exponentially fast. The following theorem generalizes it to general activation functions.

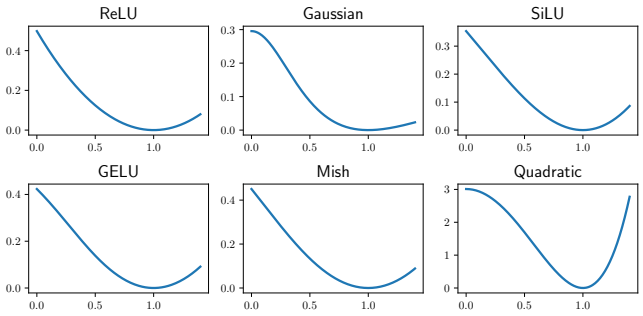

Figure 1: The landscape of $r_\sigma(\cdot)$ for various activation functions.

Theorem 4.3. *Suppose that $\sigma(\cdot)$ satisfies $r'_\sigma(\beta) \leq -C(1-\beta)$ for any $\beta \in [0, 1]$ and some constant $C > 0$. We have $\|\mathbf{w}_t - \mathbf{w}^*\| \leq e^{-Ct}$.*

Proof. It is obvious that $\|\mathbf{w}_t - \mathbf{w}^*\| = 1 - \beta_t$. $\dot{\beta}_t = -r'(\beta_t) \geq C(1-\beta_t)$, which leads to $1 - \beta_t \leq e^{-Ct}$. Hence, we complete the proof. □

The assumption of the activation function in Theorem 4.3 is quite general but abstract. In the following, we substantiate it with some explicit assumptions.

4.1.1 Monotonic activations

Lemma 4.4. *If σ is monotonic, $r_\sigma(\cdot)$ is also monotonic in $[0, 1]$. Furthermore, if there exists an interval $I = [z_0, z_1]$ such that $0 \in I$ and $\sigma'(z) \geq C_1 > 0$ for $z \in I$. Then, there exists $C_2 > 0$ such that $r'_\sigma(\beta) \leq -C_2(1-\beta)$ for any $\beta \in [0, 1]$.*

Proof. If σ is monotonically increasing, then $\sigma'(z) \geq 0$ a.e., thereby $(\sigma(z) - \sigma(\beta z))\sigma'(\beta)z \geq 0$ for $\beta \in [0, 1]$. Hence, $r'_\sigma(\beta) = -\mathbb{E}[(\sigma(z) - \sigma(\beta z))\sigma'(\beta z)z] \leq 0$, for

any $\beta \in [0,1]$, i.e., $r_\sigma(\cdot)$ is monotonically decreasing in $[0,1]$. If $\sigma'(z) \geq C_1$ for $z \in [z_0, z_1]$,

$$r'_\sigma(\beta) \geq \frac{1}{\sqrt{2\pi}} \int_{z_0}^{z_1} (\sigma(z) - \sigma(\beta z))\sigma'(\beta z) z e^{-z^2/2} \mathrm{d}z$$
$$\geq \frac{1}{\sqrt{2\pi}} \int_{z_0}^{z_1} C_1 (z - \beta z) z e^{-z^2/2} \mathrm{d}z = C_2(1-\beta),$$

where $C_2 = \frac{C_1}{\sqrt{2\pi}} \int_{z_0}^{z_1} z^2 e^{-z^2/2} \mathrm{d}z$. □

The condition that $\sigma'(\cdot)$ is bounded away from zero in a neighbor of the origin is satisfied by all the monotonic activations used in practice. We remark that this condition is also necessary, otherwise $r_\sigma(\cdot)$ could be flat in some place of $[0,1]$. Consider the activation function $\sigma(z) = \max(1, \max(z-1, 0))$, for which $\sigma'(z) = 0$ for $z \in (-\infty, 1)$. Figure 2 shows the landscapes of $\sigma(\cdot)$ and $r_\sigma(\cdot)$. One can see that $r'_\sigma(\beta) = 0$ when β is close to 0, which causes that GD starting from $\beta = 0$ gets trapped, thereby failing to converge.

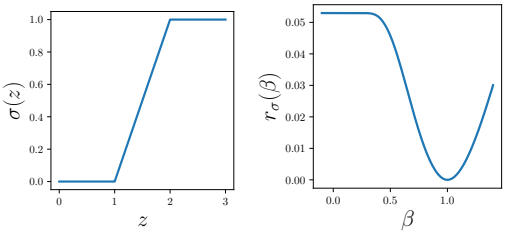

Figure 2: $\sigma(z) = \max(1, \max(z-1, 0))$ (left) and $r_\sigma(\cdot)$ (right).

4.1.2 Non-monotonic activations

We now consider non-monotonic activation functions.

Assumption 2. There exists $z_0 > 0$ such that $\sigma(\cdot)$ is monotonically decreasing in $[-\infty, -z_0]$ and monotonically increasing in $[z_0, \infty]$. Moreover, we assume that there exist a $C > 0$ such that $\sigma'(z) \geq C$ for $z \in [0, z_0]$, and $q(z) = \sigma(z) - \sigma(-z)$, $p(z) = \sigma(z) + \sigma(-z)$ are both monotonically increasing for $z \geq 0$.

The monotonicity of $q(\cdot)$ and $p(\cdot)$ ensure that the increasing part dominates the decreasing part. The above assumption is satisfied by the self-gated family $\sigma(z) = z\phi(z)$ with $\phi(z) + \phi(-z) = 1$. In particular, SiLU/Swish and GELU belongs to this family. This can be seen as follows. For any $z \geq 0$, $q'(z) = \phi(z) - \phi(-z) + z(\phi'(z) + \phi'(-z)) \geq 0$, and $p(z) = z(\phi(z) + \phi(-z)) = z$.

Lemma 4.5. *Under Assumption 2, there exists a constant $C > 0$ such that $r'_\sigma(\beta) \leq -C(1-\beta)$ for any $\beta \in [0,1]$.*

The proof is deferred to Appendix A.2, which is similar to the proof of Lemma 4.4 but more dedicated.

Relationship with existing negative results As a complement to these positive results, here we provide an analysis of the negative example used in Shamir (2018), where $\sigma(z) = \sin(dz)$. Figure 3 shows the landscape of $r_\sigma(\cdot)$ for various d's. When $d = 1$, the landscape is nice. However, when $d = 2$, a bad local minimum appears in $[0,1]$. The situation becomes severer as increasing d. Hence, GD with zero initialization fails to converge when d is relatively large.

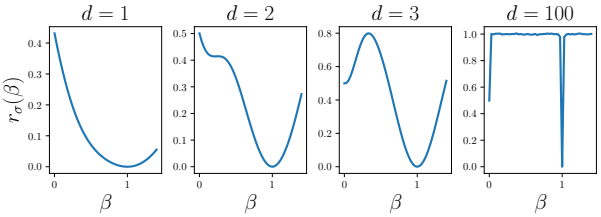

Figure 3: The landscape of $r_\sigma(\cdot)$ for $\sigma(z) = \sin(dz)$.

4.2 Random Initialization

In this section, we assume $\mathbf{w} \in \mathbb{S}^{d-1}, \sigma \in L^2(\mu_0)$ where $\mu_0 = \mathcal{N}(0,1)$ and consider the random initialization $\mathbf{w}_0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Let $\{h_i\}_{i=1}^\infty$ denote the probabilistic Hermite polynomials, which form a set of orthonormal basis of $L^2(\mu_0)$. In particular,

$$h_0(z) = 1,\ h_1(z) = z,\ h_2(z) = \frac{z^2-1}{\sqrt{2}}, h_3(z) = \frac{z^3 - 3z}{\sqrt{6}}.$$

We expand σ as $\sigma(z) = \sum_{i=0}^\infty \hat{\sigma}_i h_i(z)$, where $\hat{\sigma}_i = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(z)h_i(z)]$ is the *Hermite coefficient* of σ. We will study how the decay of $\hat{\sigma}_i$ affects the property of the risk landscape and converge of GD.

Lemma 4.6. *Assume that $\mathbf{w} \in \mathbb{S}^{d-1}$ and let $f(z) = \sum_{i=0}^\infty \hat{\sigma}_i^2 z^i$. The population risk can be written as*

$$\mathcal{R}(\mathbf{w}) = f(1) - f(\mathbf{w}^T\mathbf{w}^*). \tag{4}$$

Proof. Notice that $\mathcal{R}(\mathbf{w}) = \frac{1}{2}\mathbb{E}[\sigma(\mathbf{w}^T\mathbf{x})^2] - \mathbb{E}[\sigma(\mathbf{w}^T\mathbf{x})\sigma(\mathbf{w}^{*T}\mathbf{x})] + \frac{1}{2}\mathbb{E}[\sigma(\mathbf{w}^{*T}\mathbf{x})^2]$ and for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{S}^d$,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\sigma(\mathbf{w}_1^T\mathbf{x})\sigma(\mathbf{w}_2^T\mathbf{x})\right] \\
&= \mathbb{E}\Big[\sum_{i=0}^\infty \hat{\sigma}_i h_i(\mathbf{w}_1^T\mathbf{x}) \sum_{j=0}^\infty \hat{\sigma}_j h_j(\mathbf{w}_2^T\mathbf{x})\Big] \\
&= \sum_{i,j=0}^\infty \hat{\sigma}_i\hat{\sigma}_j\, \mathbb{E}\left[h_i(\mathbf{w}_1^T\mathbf{x})h_j(\mathbf{w}_2^T\mathbf{x})\right] \\
&= \sum_{i=0}^\infty \hat{\sigma}_i^2 (\mathbf{w}_1^T\mathbf{w}_2)^i,
\end{aligned} \tag{5}$$

where the last equality follows from (O'Donnell, 2014, Proposition 11.31). □

Denote by grad the Riemannian gradient on $\mathbb{S}^{d-1}$. Then, $\operatorname{grad}\mathcal{R}(\mathbf{w}) = -(1-\mathbf{w}\mathbf{w}^T)f'(\mathbf{w}^T\mathbf{w}^*)\mathbf{w}^*$ and the GD flow on the sphere is given by

$$\dot{\mathbf{w}}_t = (1-\mathbf{w}_t\mathbf{w}_t^T)f'(\mathbf{w}_t^T\mathbf{w}^*)\mathbf{w}^*. \tag{6}$$

Let $a_t = \langle \mathbf{w}_t, \mathbf{w}^*\rangle$. Then, we have

$$\dot{a}_t = f'(a_t)(1-a_t^2), \tag{7}$$

which is an one-dimensional ODE, completely determined by $f'(a) = \sum_{i=1}^{\infty}\hat{\sigma}_i^2 i a^{i-1}$. By (7), the set of critical points of $\mathcal{R}(\cdot)$ is given by

$$\mathcal{C} := \{\mathbf{w}\in\mathbb{S}^{d-1} : f'(\mathbf{w}^T\mathbf{w}^*) = 0 \text{ or } |\mathbf{w}^T\mathbf{w}^*|^2 = 1\}. \tag{8}$$

Remark. Here we only consider the Riemannian GD flow; otherwise, the $\mathbf{w}_t$ will leave away from $\mathbb{S}^{d-1}$, for which the risk landscape has a simple analytic expression. If we do not impose this constraint, the population landscape still has an analytic expression:

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2}H(1,1,1)+\frac{1}{2}H(1,\|\mathbf{w}\|,\|\mathbf{w}\|)-H(\hat{\mathbf{w}}^T\mathbf{w}^*,\|\mathbf{w}\|,1),$$

where $H : \mathbb{R}^3 \mapsto \mathbb{R}$ is given by $H(z,s_1,s_2) = H(z,s_2,s_1) = \sum_{k=0}^{\infty}\hat{\sigma}_k(s_1)\hat{\sigma}_k(s_2)z^k$ and $\hat{\sigma}_k(s) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(sz)h_k(z)]$. In such a case, the analysis is much more involved since we need to characterize how the Hermite coefficients are affected by the dilation of σ. We leave this to future work.

4.2.1 Convergence with Constant Probability

When $\sigma(\cdot)$ is nonzero, there must exist $i\in\mathbb{N}_+$ such that $\hat{\sigma}_i^2 > 0$. Hence, $f'(a) \geq \hat{\sigma}_i^2 i a^{i-1} > 0$ for $a > 0$. Consequently, the global minima $\mathbf{w} = \mathbf{w}^*$ is unique critical point in the positive halfspace: $\{\mathbf{w}\in\mathbb{S}^{d-1} : \mathbf{w}^T\mathbf{w}^* > 0\}$. Moreover, it is obvious that the whole positive halfspace is the basin of attraction. Using this observation, we have the following convergence result.

Proposition 4.7. *Assume that $\sigma(\cdot)$ is nonzero. Let $k = \min\{i : \sigma_i \neq 0\}$. Then, there exists a constant $C > 0$ such that for any $\delta\in(0,\frac{1}{2})$, with probability $\frac{1}{2} - \frac{C\delta}{\sqrt{d}}$, we have $1-\mathbf{w}_t^T\mathbf{w}^* \leq e^{-c_k t}$ with $c_k = k\hat{\sigma}_k^2(\frac{\delta}{d})^{k-1}$.*

Proof. Since $\mathbf{w}_0\sim\operatorname{Unif}(\mathbb{S}^{d-1})$, $a_0 = \mathbf{w}_0^T\mathbf{w}^*$ follows the distribution: $g(z) = \frac{1}{\sqrt{\pi}}\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})}(1-z^2)^{\frac{d-3}{2}}$. It is easy to verify that there exists a constant $C_1 > 0$ such that $(1-t)^q \leq 1 - C_1 qt$ for $t\in[0,\frac{1}{d}]$. Then, for $\delta\leq 1$,

$$\begin{aligned}\mathbb{P}\{a_0\geq\frac{\delta}{d}\} &= \frac{1}{2} - \int_0^{\frac{\delta}{d}} g(z)\mathrm{d}z \\ &\geq \frac{1}{2} - \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\int_0^{\frac{\delta}{d}}(1-c_1\frac{d-3}{2}z^2)\mathrm{d}z \\ &\geq \frac{1}{2} - C_2\frac{\delta}{\sqrt{d}}. \end{aligned}\tag{9}$$

Therefore, with probability $\frac{1}{2} - \frac{C_2\delta}{\sqrt{d}}$, $f'(a_0) \geq k\hat{\sigma}_k^2 a_0^{k-1} > 0$. With this initialization, a_t keep increasing for $t\geq 0$. Then, we have $\dot{a}_t = f'(a_t)(1-a_t^2) \geq f'(a_0)(1-a_t)$. This yields that $1-a_t \leq e^{-f'(a_0)t}$. We thus complete the proof since $f'(a_0)\geq k\hat{\sigma}_k^2 a_0^{k-1}$. □

Proposition 4.7 provides a constant-probability (close to 1/2) guarantee for the GD convergence, and it only require σ to be nonzero. Moreover, the more the Hermite coefficients concentrate at small k's, the faster is the convergence. In particular, if $\hat{\sigma}_1 \neq 0$, we have $c_k = \hat{\sigma}_1^2$ and as such, the convergence rate is independent of d.

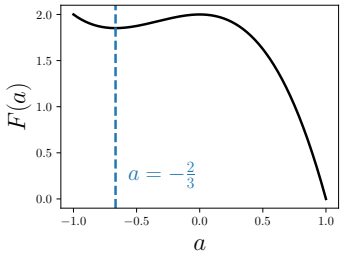

Figure 4: The landscape $F(a) = 1 - f(a)$ for $\sigma = h_2 + h_3$. Here $a = -2/3$ is a bad local minima.

Optimality. The following lemma shows that the success probability cannot be further improved without imposing stronger conditions on $\sigma(\cdot)$.

Lemma 4.8. *Assume $\sigma = h_2 + h_3$, where h_2 and h_3 are the 2-th and 3-th Hermite polynomial, respectively. Then, $\mathcal{R}(\cdot)$ has bad local minima: $\mathcal{Q} = \{\mathbf{w}\in\mathbb{S}^{d-1} : \mathbf{w}^T\mathbf{w}^* = -2/3\}$ and moreover, w.p. $\frac{1}{2} - O(\frac{1}{\sqrt{d}})$ over the random initialization, GD converges to $\mathcal{Q}$.*

Proof. By the assumption, $f(a) = a^2 + a^3$, $f'(a) = 2a + 3a^2$. Then, $\mathcal{R}(\mathbf{w}) = F(\mathbf{w}^T\mathbf{w}^*)$ with $F(a) = 2 - a^2 - a^3$. F has a bad local minimum at $a = -2/3$, where $F(-2/3) = 50/27 > F(1) = 0$ (See Figure 4 for an illustration). Hence, $\{\mathbf{w}\in\mathbb{S}^{d-1} : \mathbf{w}^T\mathbf{w}^* = -2/3\}$ is a set of bad local minima of $\mathcal{R}(\cdot)$. Substituting $f'(a) = 2a + 3a^2$ into (7) gives us

$$\dot{a}_t = a_t(2+3a_t)(1-a_t^2).$$

Following the estimate (9) and symmetry, we have w.p. $1/2 - C/(4\sqrt{d})$ that $-1/4 \leq a_0 < 0$. This will cause that a_t decreases to $a = -2/3$. Therefore, when $d \gg 1$, with a probability close to 1/2, GD fails to converge to global minima. □

4.2.2 A High-Probability Convergence

In this section, we show that the probability of GD convergence can be boosted (to 1) by making stronger assumptions on the activation function.

Let us first take a closer look at the risk landscape. Define

$$q_\sigma(\delta) = \hat{\sigma}_1^2 - \sum_{i=1}^{\infty}(2i)\hat{\sigma}_{2i}^2\delta^{2i-1}. \tag{10}$$

According to (7), when $f'(a) > 0$ for any $a \in [-1, 0]$, there are only two critical points $\mathbf{w} = \mathbf{w}^*$ (minimum) and $\mathbf{w} = -\mathbf{w}^*$ (maximum). One condition to ensure $f'(a) > 0$, $\forall a \in [-1, 0]$ is $q_\sigma(1) > 0$ since

$$f'(a) \geq \hat{\sigma}_1^2 - \sum_{i=1}^{\infty}(2i)\hat{\sigma}_{2i}^2 = q_\sigma(1) > 0. \tag{11}$$

Since $\hat{\sigma}_1 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z\sigma(z)]$, this condition implies that the linear component of σ dominates the high-order components. We numerically verify that, $q_\sigma(1) > 0$ for all the ReLU variants, including the non-monotonic SiLU/Swish and GELU.

The above landscape analysis implies that when $q_\sigma(1) > 0$, the success probability of convergence for random initialization is exactly 1. The proposition given below further shows that as long as $q_\sigma(\delta) > 0$ for some small constant $\delta > 0$ is sufficient to establish a high-probability convergence when $d \gg 1$. Note that under this condition, there may exist bad local minima and saddle points. The high-probability convergence is made possible by two facts: (1) The near-origin region lies in the basin of attraction of the global minimum; (2) The random initialization can avoid the pathologic region with a high probability.

Proposition 4.9. *Suppose that $q_\sigma(\delta) > 0$ for some constant $\delta \in (0, 1/2]$. Then, with probability at least $1 - 0.5e^{-d\delta^2}$, $1 - \mathbf{w}_t^T\mathbf{w}^* \leq e^{-q_\sigma(\delta)t/2}$.*

Proof. Notice that for $a \in [-\delta, 1]$,

$$f'(a) = \hat{\sigma}_1^2 + 2\hat{\sigma}_2^2 a + 3\hat{\sigma}_3^2 a^2 + \cdots \geq q_\sigma(\delta) > 0. \tag{12}$$

Since $\mathbf{w}_0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, with probability $1 - 0.5e^{-d\delta^2}$, $a_0 = \mathbf{w}_0^T\mathbf{w}^* \geq -\delta$. Thus, $f'(a_0) \geq q_\sigma(\delta) > 0$. Therefore, a_t is increasing for $t \in [0, \infty)$, and by using (12),

$$\dot{a}_t \geq q_\sigma(\delta)(1-a_t^2) = q_\sigma(\delta)(1-a_t)(1+a_t) \geq \frac{q_\sigma(\delta)}{2}(1-a_t)$$

This leads to that $1 - a_t \leq e^{-q_\sigma(\delta)t/2}$. □

Remark. Combined with Lemma 4.8, it is revealed that the dominance of linear component for the activation function is crucial for achieving high-probability convergence. This provides an explanation of the wide use of ReLU and its variants.

Relationship with existing negative results Consider the setting used in Shamir (2018), where $\mathcal{D} = \mathcal{N}(0, I_d)$ and $\sigma(z) = \sin(dz)$. A detailed calculation (provided in Appendix B.2) tells us

$$\hat{\sigma}_1 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z\sin(dz)] = de^{-d^2/2}. \tag{13}$$

Hence, $\hat{\sigma}_1$ and $q_\sigma(\delta)$ are exponentially small. Consequently, the convergence of GD is exponentially slow. Note that it is not surprising that $\hat{\sigma}_1$ and $q_\sigma(\delta)$ are exponentially small since the activation function is highly oscillated in this case.

5 Learning with Finite Samples

We now proceed to the finite sample case. Specifically, we focus on the case that the input distribution is standard Gaussian. The extension to the setting used in Section 3 is straightforward. We make the following assumption for technical simplicity, which is satisfied by SiLU/Swish and GELU.

Assumption 3. Assume that σ', σ'' exist and $\max(|\sigma'(z)|, |\sigma''(z)|) \lesssim 1$ for any $z \in \mathbb{R}$.

Let $E_Q = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}^*\| \leq Q\}$. The following proposition bounds the difference between the empirical and population landscape for $\mathbf{w} \in E_Q$.

Proposition 5.1. *Assume that $n \geq 10$. For any $\delta \in (0, 1)$, w.p. $1 - \delta$ over the sampling of training set,*

$$\sup_{\mathbf{w}\in E_Q} |\hat{\mathcal{R}}_n(\mathbf{w}) - \mathcal{R}(\mathbf{w})| \lesssim \frac{\log(n/\delta))d}{\sqrt{n}}(Q+1)^2$$

$$\sup_{\mathbf{w}\in E_Q} \|\nabla\hat{\mathcal{R}}_n(\mathbf{w}) - \nabla\mathcal{R}(\mathbf{w})\| \lesssim \frac{\log^{3/2}(n/\delta)\sqrt{d}}{\sqrt{n}}(Q+1)^2.$$

This proposition is proved by using the techniques of empirical processes. However, the empirical processes in our case are not sub-gaussian due to the squared loss and the unboundedness of the input distribution. To handle this issue, we adopt a truncation method to capture the tail behavior. We refer to Appendix E for more details.

The following lemma shows that the population risk and its gradient are Lipschitz continuous and the Lipschitz constants are independent of d. The proof is deferred to Appendix D.

Lemma 5.2. *For any $\mathbf{w}_1, \mathbf{w}_2 \in E_Q$, we have $|\mathcal{R}(\mathbf{w}_1) - \mathcal{R}(\mathbf{w}_2)| \lesssim Q\|\mathbf{w}_1 - \mathbf{w}_2\|$ and $\|\nabla\mathcal{R}(\mathbf{w}_1) - \nabla\mathcal{R}(\mathbf{w}_2)\| \lesssim (1+Q)\|\mathbf{w}_1 - \mathbf{w}_2\|$.*

Using Proposition 5.1 and Lemma 5.2, we can convert the preceding convergence results of population GD to the empirical GD as shown below. The proofs are deferred to Appendix C.

Proposition 5.3 (Zero initialization). *Suppose that the activation function satisfies Assumption 3 and the condition in Theorem 4.3. Let $\hat{\mathbf{w}}_t$ be the GD solution starting from zero. There exists $C_1, C_2, C_3 > 0$ and let $\epsilon_n = \frac{C_3\sqrt{d}\log^{3/2}(n/\delta)}{\sqrt{n}}$. There exists $T = \frac{\log(1/\epsilon_n)}{C_1+C_2}$ such that*

$$\|\hat{\mathbf{w}}_T - \mathbf{w}^*\| \le \epsilon_n^{\frac{C_1}{C_1+C_2}} \tag{14}$$

Proposition 5.4 (Random initialization). *Let $\delta_1 \in (0, 1/2], \delta_2 \in (0,1)$. Suppose that Assumption 3 holds and $q_\sigma(\delta_1) > 0$. Let $\hat{\mathbf{w}}_t$ be the solution of the Riemanian GD (6) initialized from $\hat{\mathbf{w}}_0 \sim Unif(\mathbb{S}^{d-1})$. Then, w.p. at least $1 - 0.5e^{-d\delta_1^2}$ over the initialization and $1-\delta_2$ over the sampling of training set, we have*

$$\|\hat{\mathbf{w}}_t - \mathbf{w}^*\|^2 \lesssim e^{-\frac{q_\sigma(\delta_1)}{2}t} + \frac{1}{q_\sigma(\delta_1)}\sqrt{\frac{d\log^3(n/\delta_2)}{n}}.$$

The above two propositions show that learning a single neuron via GD only requires polynomial samples and polynomial time. For instance, in Proposition 5.4, the sample and time complexities are $\tilde{O}(d/\varepsilon^2)$ and $O(\log(1/\varepsilon))$, respectively. It should be stressed that our upper bounds are not necessarily optimal and the logarithmic terms can be removed by assuming the input distribution to be bounded.

6 Conclusion

In this work, the problem of learning a single neuron with GD is studied under the realizable setting. We show that a single neuron can be learned efficiently (i.e., the sample complexity and time complexity are polynomial in the input dimension and target accuracy) as long as the activation function has a dominating linear or monotonic component. In contrast to existing work, our conditions remove the restriction of monotonicity and are satisfied by all the commonly-used non-monotonic activation functions. It is of much interest to extend our analysis to the agnostic learning setting (Frei et al., 2020), where no relationship between the label y and the input $\mathbf{x}$ is assumed. In such a case, one needs to deal with some extra hardness (Goel et al., 2019). For example, there may exist many bad local minima even if σ is strictly monotonic (Auer et al., 1996).

Acknowledgements

We thank Weinan E, Chao Ma, and Jihao Long for many helpful discussions and anonymous reviewers for valuable suggestions.

References

P. Auer, M. Herbster, M. K. Warmuth, et al. Exponentially many local minima for single neurons. *Advances in neural information processing systems*, pages 316–322, 1996.

A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, 1994.

M. L. Brady, R. Raghavan, and J. Slawny. Back propagation fails to separate where perceptrons succeed. *IEEE Transactions on Circuits and Systems*, 36(5): 665–674, 1989.

J. Chen, R. Du, and K. Wu. A comparison study of deep Galerkin method and deep Ritz method for elliptic problems with different boundary conditions. *arXiv e-prints*, pages arXiv–2005, 2020.

Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer ReLU networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020.

S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

S. Frei, Y. Cao, and Q. Gu. Agnostic learning of a single neuron with gradient descent. *arXiv preprint arXiv:2005.14426*, 2020.

S. Goel, S. Karmalkar, and A. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems*, pages 8584–8593, 2019.

D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUS). *arXiv preprint arXiv:1606.08415*, 2016.

S. Kakade, A. T. Kalai, V. Kanade, and O. Shamir. Efficient learning of generalized linear and single index models with isotonic regression. *arXiv preprint arXiv:1104.2018*, 2011.

A. T. Kalai and R. Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*. Citeseer, 2009.

S. M. M. Kalan, M. Soltanolkotabi, and A. S. Avestimehr. Fitting ReLUs via SGD and quantized SGD.

In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2469–2473. IEEE, 2019.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45 (6):983–1006, 1998.

X.-A. Li, Z.-Q. J. Xu, and L. Zhang. A multi-scale dnn algorithm for nonlinear elliptic equations with multiple scales. *Communications in Computational Physics*, 28(5):1886–1906, 2020.

S. Liang, L. Lyu, C. Wang, and H. Yang. Reproducing activation function for deep learning. *arXiv preprint arXiv:2101.04844*, 2021.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

R. Livni, S. Shalev-Shwartz, and O. Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pages 855–863, 2014a.

R. Livni, S. Shalev-Shwartz, and O. Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 855–863, 2014b.

A. Maillard, G. Ben Arous, and G. Biroli. Landscape complexity for the empirical risk of generalized linear models. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107, pages 287–327. PMLR, 20–24 Jul 2020.

E. Malach and S. Shalev-Shwartz. When hardness of approximation meets hardness of learning. *arXiv preprint arXiv:2008.08059*, 2020.

S. Mei, Y. Bai, A. Montanari, et al. The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774, 2018.

R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

S. Oymak and M. Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR, 2019.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

V. Ros, G. B. Arous, G. Biroli, and C. Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Physical Review X*, 9(1):011003, 2019.

O. Shamir. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research*, 19(1):1135–1163, 2018.

V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33, 2020.

M. Soltanolkotabi. Learning ReLUs via gradient descent. In *Advances in neural information processing systems*, pages 2007–2017, 2017.

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

Y. S. Tan and R. Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*, 2019.

K. Tessera, S. Hooker, and B. Rosman. Keep the gradients flowing: Using gradient flow to study sparse network optimization. *arXiv preprint arXiv:2102.01670*, 2021.

Y. Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

C. Xie, M. Tan, B. Gong, A. Yuille, and Q. V. Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

G. Yehudai and O. Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32:6598–6608, 2019.

G. Yehudai and O. Shamir. Learning a single neuron with gradient methods. *arXiv preprint arXiv:2001.05205*, 2020.

Supplementary Material: Learning a Single Neuron for Non-monotonic Activation Functions

A Proofs for Section 3

A.1 Proof of Proposition 3.1

Our proof needs the following technical lemma.

Lemma A.1 (Lemma B.1 in (Yehudai and Shamir, 2020)). *For some fixed α, and let $\mathbf{a}, \mathbf{b}$ be two unit vectors in $\mathbb{R}^2$ such that $\arccos(\mathbf{a}^T\mathbf{b}) \leq \pi - \delta$ for some $\delta \in (0, \pi]$. Then,*

$$\inf_{\mathbf{u}\in\mathbb{R}^2, \|\mathbf{u}\|=1} \int \mathbf{1}_{\mathbf{a}^T\mathbf{y}>0}\mathbf{1}_{\mathbf{b}^T\mathbf{y}>0}\mathbf{1}_{\|\mathbf{y}\|\leq\alpha}(\mathbf{u}^T\mathbf{y})^2 \mathrm{d}\mathbf{y} \geq \frac{\alpha^4}{8\sqrt{2}} \sin^3\left(\frac{\delta}{4}\right)$$

Proof of Proposition 3.1. Let $S(\mathbf{w}, \mathbf{w}^*) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^T\mathbf{x} \geq 0, \mathbf{w}^{*T}\mathbf{x} \geq 0, \|\mathbf{x}\| \leq \alpha\}$ where α is the constant defined in Assumption 1, and

$$A(\mathbf{w}, \mathbf{w}^*, \mathbf{x}) = (\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\mathbf{w}^T\mathbf{x})(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x}).$$

Denote by $S^c(\mathbf{w}, \mathbf{w}^*)$ be the complement of $S(\mathbf{w}, \mathbf{w}^*)$. Using Assumption 1 and the mean value theorem, we have

$$A(\mathbf{w}, \mathbf{w}^*, \mathbf{x}) \geq \begin{cases} \gamma^2(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2, & \text{if } \mathbf{x} \in S(\mathbf{w}, \mathbf{w}^*) \\ -\zeta^2(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2, & \text{if } \mathbf{x} \in S^c(\mathbf{w}, \mathbf{w}^*). \end{cases}$$

Then,

$$\begin{aligned}
\langle \nabla\mathcal{R}(\mathbf{w}), \mathbf{w} - \mathbf{w}^* \rangle &= \mathbb{E}_{\mathbf{x}}[A(\mathbf{w}, \mathbf{w}^*, \mathbf{x})] = \mathbb{E}_{\mathbf{x}}[A(\mathbf{w}, \mathbf{w}^*, \mathbf{x})\mathbf{1}_{S(\mathbf{w},\mathbf{w}^*)}] + \mathbb{E}_{\mathbf{x}}[A(\mathbf{w}, \mathbf{w}^*, \mathbf{x})\mathbf{1}_{S^c(\mathbf{w},\mathbf{w}^*)}] \\
&\geq \gamma^2 \mathbb{E}_{\mathbf{x}}[(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2\mathbf{1}_{S(\mathbf{w},\mathbf{w}^*)}] - \zeta^2 \mathbb{E}_{\mathbf{x}}[(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2\mathbf{1}_{S^c(\mathbf{w},\mathbf{w}^*)}] \\
&\geq (\gamma^2 + \zeta^2) \mathbb{E}_{\mathbf{x}}[(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2\mathbf{1}_{S(\mathbf{w},\mathbf{w}^*)}] - \zeta^2 \mathbb{E}_{\mathbf{x}}[(\mathbf{w}^T\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2] \\
&\geq (\gamma^2 + \zeta^2)\|\mathbf{w} - \mathbf{w}^*\|^2 \inf_{\mathbf{u}\in\mathrm{span}(\mathbf{w},\mathbf{w}^*), \|\mathbf{u}\|=1} \mathbb{E}_{\mathbf{x}}[(\mathbf{u}^T\mathbf{x})^2\mathbf{1}_{S(\mathbf{w},\mathbf{w}^*)}] - \zeta^2\tau\|\mathbf{w} - \mathbf{w}^*\|^2, \qquad (15)
\end{aligned}$$

where the last inequality uses the assumption that $\mathbb{E}[\mathbf{x}\mathbf{x}^T] \leq \tau I_d$. What remains is to bound the first term of the right hand side. Let $\mathbf{y} = (\mathbf{w}^T\mathbf{x}, \mathbf{w}^{*T}\mathbf{x}) \in \mathbb{R}^2$ be the projection of $\mathbf{x}$ into $\mathrm{span}\{\mathbf{w}, \mathbf{w}^*\}$. Then,

$$\begin{aligned}
\inf_{\mathbf{u}\in\mathrm{span}(\mathbf{w},\mathbf{w}^*), \|\mathbf{u}\|=1} \mathbb{E}_{\mathbf{x}}[(\mathbf{u}^T\mathbf{x})^2\mathbf{1}_{S(\mathbf{w},\mathbf{w}^*)}] &= \inf_{\mathbf{u}\in\mathrm{span}(\mathbf{w},\mathbf{w}^*), \|\mathbf{u}\|=1} \mathbb{E}_{\mathbf{x}}\left[(\mathbf{u}^T\mathbf{x})^2\mathbf{1}_{\|\mathbf{x}\|\leq\alpha}\mathbf{1}_{\mathbf{w}^T\mathbf{x}\geq 0}\mathbf{1}_{\mathbf{w}^{*T}\mathbf{x}\geq 0}\right] \\
&\geq \inf_{\mathbf{u}\in\mathbb{R}^2, \|\mathbf{u}\|=1} \int (\mathbf{u}^T\mathbf{y})^2\mathbf{1}_{\|\mathbf{y}\|\leq\alpha}\mathbf{1}_{y_1\geq 0}\mathbf{1}_{y_2\geq 0}p_{\mathbf{w},\mathbf{w}^*}(\mathbf{y})\mathrm{d}\mathbf{y} \\
&\geq \beta \inf_{\mathbf{u}\in\mathbb{R}^2, \|\mathbf{u}\|=1} \int (\mathbf{u}^T\mathbf{y})^2\mathbf{1}_{\|\mathbf{y}\|\leq\alpha}\mathbf{1}_{y_1\geq 0}\mathbf{1}_{y_2\geq 0}\mathrm{d}\mathbf{y} \\
&\geq \beta\frac{\alpha^4}{8\sqrt{2}}\sin^3(\delta/4), \qquad (16)
\end{aligned}$$

where the last inequality follows from Lemma A.1. Plugging (16) into (15) completes the proof.

A.2 Proof of Proposition 3.3

First, if the initialization satisfies $\|\mathbf{w}_0 - \mathbf{w}^*\| < 1$, then we must have $\|\mathbf{w}_t - \mathbf{w}^*\| < 1$ for any $t \geq 0$. Otherwise, we must have $t_0 = \inf\{t : \|\mathbf{w}_t - \mathbf{w}^*\| \geq 1\} < \infty$. Then, $\|\mathbf{w}_t - \mathbf{w}^*\| < 1$ for $t \in [0, t_0)$. According to Lemma 3.2,

$\lambda(\delta_t) > 0$ for $t \in [0, t_0)$. Hence, $\mathrm{d}\|\mathbf{w}_t - \mathbf{w}^*\|^2/\mathrm{d}t \geq -\lambda\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq 0$ for $t \in [0, t_0)$, which implies that for any $t < t_0$, $\|\mathbf{w}_t - \mathbf{w}^*\| \leq \|\mathbf{w}_0 - \mathbf{w}^*\| < 1 = \|\mathbf{w}_{t_0} - \mathbf{w}^*\|$. This is contradictory to the continuity of the GD trajectory. Thus, $\|\mathbf{w}_t - \mathbf{w}^*\|^2 \leq e^{-\lambda(\frac{\pi}{2})t}\|\mathbf{w}_0 - \mathbf{w}^*\|^2$.

Second, according to (Yehudai and Shamir, 2020, Lemma 5.1), with probability larger than $\frac{1}{2} - \frac{1}{4}\eta d - 1.2^{-d}$, we have $\|\mathbf{w}_0 - \mathbf{w}^*\|^2 \leq 1 - 2\eta^2 d < 1$. Therefore, we complete the proof.

B Proofs of Section 4

B.1 Proof of Lemma 4.5

Firstly, we can write $-r'_\sigma(\beta) = \frac{1}{\sqrt{2\pi}}\int_0^\infty a(\beta, z) z e^{-z^2/2}\mathrm{d}z$, where

$$a(\beta, z) = (\sigma(z) - \sigma(\beta z))\sigma'(\beta z) - (\sigma(-z) - \sigma(-\beta z))\sigma'(-\beta z).$$

- When $-\beta z \leq -z_0$, $(\sigma(-z) - \sigma(-\beta z))\sigma'(-\beta z) \leq 0$. Hence, $a(\beta, z) \geq (\sigma(z) - \sigma(\beta z))\sigma'(\beta z) \geq 0$.
- When $-\beta z \geq -z_0$, we have $\sigma'(\beta z) \geq \sigma'(-\beta z) \geq 0$. Hence, using the the monotonicity of $q(\cdot)$, we have

$$a(\beta, z) \geq [(\sigma(z) - \sigma(\beta z)) - (\sigma(-z) - \sigma(-\beta z))]\,\sigma'(-\beta z) = [q(z) - q(\beta z)]\sigma'(-\beta z) \geq 0,$$

Combining them together, $a(\beta, z) \geq 0$ for any $z \geq 0, \beta \in [0, 1]$. Hence,

$$\begin{aligned}
-r'(\beta) &= \frac{1}{\sqrt{2\pi}}\int_0^\infty a(\beta, z) z e^{-\frac{z^2}{2}}\mathrm{d}z \geq \frac{1}{\sqrt{2\pi}}\int_0^{z_0} a(\beta, z) z e^{-\frac{z^2}{2}}\mathrm{d}z \\
&\geq \frac{1}{\sqrt{2\pi}}\int_0^{z_0} (\sigma(z) - \sigma(\beta z))\sigma'(\beta z) z e^{-\frac{z^2}{2}}\mathrm{d}z \geq C\int_0^{z_0}(z - \beta z) z e^{-\frac{z^2}{2}}\mathrm{d}z \geq C(1 - \beta).
\end{aligned}$$

□

B.2 Calculation of $\hat{\sigma}_1$ for the Sine activation function

$$\begin{aligned}
\hat{\sigma}_1 &= \mathbb{E}_{z\sim\mathcal{N}(0,1)}[z\sigma(z)] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} z\sin(dz)e^{-z^2/2}\mathrm{d}z = \frac{d}{\sqrt{2\pi}}\int \cos(dz)e^{-t^2/2}\mathrm{d}z \\
&= \frac{d}{\sqrt{2\pi}}\sum_{n=0}^\infty \frac{(-1)^n}{(2n)!}\int (dt)^{2n} e^{-z^2/2}\mathrm{d}z \\
&= d\sum_{n=0}^\infty \frac{(-1)^n d^{2n}}{(2n)!}(2n-1)!! \\
&= d\sum_{n=0}^\infty \frac{(-d^2/2)^n}{n!} = de^{-d^2/2}.
\end{aligned}$$

Therefore, the first Hermite coefficient is exponentially small for the periodic activation function: $\sigma(z) = \sin(dz)$.

C Proofs for empirical GD

C.1 Proof of Proposition 5.3

Denote by $\mathbf{w}_t$ and $\hat{\mathbf{w}}_t$ the solutions of population and empirical GD, respectively, i.e., $\mathbf{w}_0 = \hat{\mathbf{w}}_0 = 0$ and $\dot{\mathbf{w}}_t = -\nabla\mathcal{R}(\mathbf{w}_t), \dot{\hat{\mathbf{w}}}_t = -\nabla\hat{\mathcal{R}}_n(\hat{\mathbf{w}}_t)$. By Theorem 4.3, we have

$$\|\mathbf{w}_t - \mathbf{w}^*\| \leq e^{-C_1 t}. \tag{17}$$

For the empirical GD, let $T_0 = \inf\{t : \|\hat{\mathbf{w}}_t - \mathbf{w}^*\| \geq 2\}$ and $\Delta_t = \mathbf{w}_t - \hat{\mathbf{w}}_t$. Then, for $t \leq T_0$,

$$\frac{\mathrm{d}\|\Delta_t\|^2}{\mathrm{d}t} = -2\langle\nabla\mathcal{R}(\mathbf{w}_t) - \nabla\mathcal{R}(\hat{\mathbf{w}}_t), \Delta_t\rangle - 2\langle\nabla\mathcal{R}(\hat{\mathbf{w}}_t) - \nabla\hat{\mathcal{R}}_n(\hat{\mathbf{w}}_t), \Delta_t\rangle$$

$$\lesssim \|\Delta_t\|^2 + \frac{\sqrt{d}\log^{3/2}(n/\delta)}{\sqrt{n}}\|\Delta_t\|,$$

where the last inequality follows from Lemma 5.2 and Proposition 5.1. Let $\epsilon_n = \frac{C_3\sqrt{d}\log^{3/2}(n/\delta)}{\sqrt{n}}$. Hence $\frac{\mathrm{d}\|\Delta_t\|}{\mathrm{d}t} \le C_2\|\Delta_t\| + \epsilon_n$, which yields to

$$\|\Delta_t\| \le \|\Delta_0\| + \epsilon_n(e^{C_2 t} - 1) = \epsilon_n(e^{C_2 t} - 1), \tag{18}$$

where we use the fact that $\Delta_0 = 0$. Combining (17) and (18) leads to

$$\|\hat{\mathbf{w}}_t - \mathbf{w}^*\| \le \|\hat{\mathbf{w}}_t - \mathbf{w}_t\| + \|\mathbf{w}_t - \mathbf{w}^*\| \le \epsilon_n(e^{C_2 t} - 1) + e^{-C_1 t} =: e(t) - \epsilon_n, \tag{19}$$

Taking $\epsilon_n e^{C_1 t} = e^{-C_2 t}$ gives $T = \frac{\log(1/\epsilon_n)}{C_1 + C_2}$. Obviously, $e(\cdot)$ is monotonically decreasing for $t \le T$. Thus, for $t \le T$, $\|\hat{\mathbf{w}}_t - \mathbf{w}^*\| \le e(t) - \epsilon_n \le e(0) - \epsilon_n = 1$. Therefore, we must have $T_1 \le T_0$. This means that the previous estimates hold for $t \le T$. Taking $t = T$, we have $\|\hat{\mathbf{w}}_T - \mathbf{w}^*\| \le e(T) - \epsilon_n \lesssim \epsilon_n^{\frac{C_1}{C_1+C_2}}$

C.2 Proof of Proposition 5.4

The empirical GD can be written as

$$\hat{\mathbf{w}}_t = -(I - \hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^T)\nabla\mathcal{R}(\hat{\mathbf{w}}_t) - (I - \hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^T)(\nabla\hat{\mathcal{R}}_n(\hat{\mathbf{w}}_t) - \nabla\mathcal{R}(\hat{\mathbf{w}}_t)).$$

Let $\hat{a}_t = \langle\hat{\mathbf{w}}_t, \mathbf{w}^*\rangle$ and $e_t = -\mathbf{w}^{*T}(I - \hat{\mathbf{w}}_t\hat{\mathbf{w}}_t^T)(\nabla\hat{\mathcal{R}}_n(\hat{\mathbf{w}}_t) - \mathcal{R}(\hat{\mathbf{w}}_t))$. Then,

$$\dot{\hat{a}}_t = f'(\hat{a}_t)(1 - \hat{a}_t^2) + e_t,$$

By Proposition 5.1 and $\|\hat{\mathbf{w}}_t\| = 1$, with probability $1 - \delta_2$, we have $e_t \le O(\sqrt{\frac{d\log^3(n/\delta_2)}{n}}) =: \epsilon_n$. Analogous to the proof of Proposition 4.9, we have with probability $1 - 0.5e^{-d\delta_1^2}$ that,

$$\frac{d}{dt}(1 - \hat{a}_t) \le \frac{q_\sigma(\delta_1)}{2}(1 - \hat{a}_t) + \delta_t \le \frac{q_\sigma(\delta_1)}{2}(1 - \hat{a}_t) + \epsilon_n.$$

By Gronwall's inequality, $1 - \hat{a}_t \le (1 - \hat{a}_0)e^{-q_\sigma(\delta_1)t/2} + \frac{2\epsilon_n}{q_\sigma(\delta_1)}$.

D Proof of Lemma 5.2

For any $\mathbf{w} \in E_Q$, consider the orthogonal decomposition: $\mathbf{w} = \beta\mathbf{w}^* + \alpha\mathbf{w}_\perp$ with $\langle\mathbf{w}_\perp, \mathbf{w}^*\rangle = 0$ and $\|\mathbf{w}_\perp\| = 1$. Let $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)^T \in \mathbb{R}^{d\times d}$ be an orthonormal matrix with $\mathbf{v}_1 = \mathbf{w}^*, \mathbf{v}_2 = \mathbf{w}_\perp$. Using change of variable $\mathbf{x} = V\mathbf{x}$ and the symmetry of $\mathcal{N}(0, I_d)$, we have

$$\nabla\mathcal{R}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}}[(\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{x}] = V^T\mathbf{u},$$

where $\mathbf{u} = \mathbb{E}_{\mathbf{x}}[(\sigma(\beta x_1 + \alpha x_2) - \sigma(x_1))\sigma'(\beta x_1 + \alpha x_2)\mathbf{x}] = (u_1, u_2, 0, \ldots, 0)$. Here $u_i = \mathbb{E}[(\sigma(\beta x_1 + \alpha x_2) - \sigma(x_1))\sigma'(\beta x_1 + \alpha x_2)x_i]$. Hence, it is easy to see that $\|\nabla\mathcal{R}(\mathbf{w})\| = \|\mathbf{u}\| \le CQ$.

In addition,

$$\begin{aligned}\nabla^2\mathcal{R}(\mathbf{w}) &= \mathbb{E}[\sigma'(\mathbf{w}^T\mathbf{x})\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{x}\mathbf{x}^T] + \mathbb{E}[(\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}_*^T\mathbf{x}))\sigma''(\mathbf{w}^T\mathbf{x})\mathbf{x}\mathbf{x}^T] \\ &:= H_1 + H_2.\end{aligned} \tag{20}$$

We then estimate H_1, H_2 separately. By the symmetry of the input distribution, $H_1 = \mathbb{E}[\sigma'(\|\mathbf{w}\|x_1)^2\mathbf{x}\mathbf{x}^T]$. Hence

$$(H_1)_{i,j} = \mathbb{E}[\sigma'(\|\mathbf{w}\|x_1)^2 x_i x_j] = \begin{cases} 0 & \text{if } i \ne j \\ \mathbb{E}[\sigma'(\|\mathbf{w}\|x_1)^2 x_i^2] & \text{if } i = j. \end{cases}$$

Therefore, H_1 is diagonal and $\lambda_{\max}(H_1) \le C$. Let us turn to H_2. Consider the orthogonal decomposition: $\mathbf{w} = \beta\mathbf{w}^* + \alpha\mathbf{w}_\perp$ with $\langle\mathbf{w}_\perp, \mathbf{w}^*\rangle = 0$ and $\|\mathbf{w}_\perp\| = 1$. By symmetry, $H_2 = \mathbb{E}[(\sigma(\alpha x_1 + \beta x_2) - \sigma(x_2))\sigma''(\alpha x_1 + \beta x_2)\mathbf{x}\mathbf{x}^T]$. Let

$$c_{s,t} = \mathbb{E}_{x_1,x_2\sim\mathcal{N}(0,1)}[(\sigma(\alpha x_1 + \beta x_2) - \sigma(x_2))\sigma''(\alpha x_1 + \beta x_2)x_s x_t], \quad s, t = 1, 2$$

$$q = \mathbb{E}_{x_1,x_2,x_3\sim\mathcal{N}(0,1)}[(\sigma(\alpha x_1+\beta x_2)-\sigma(x_2))\sigma''(\alpha x_1+\beta x_2)x_3^2] \tag{21}$$

Hence,

$$H_2 = \begin{pmatrix} c_{1,1} & c_{1,2} & 0 & \dots & 0 \\ c_{2,1} & c_{2,2} & 0 & \dots & 0 \\ 0 & 0 & q & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & q \end{pmatrix} \tag{22}$$

It is easy to obtain that

$$\lambda_{\max}(H_2) \le \max\{q, c_{1,1}+c_{2,2}\} \lesssim |\alpha| + |\beta-1| \lesssim \|\mathbf{w}-\mathbf{w}^*\|. \tag{23}$$

Combining the estimates of H_1 and H_2, we complete the proof. □

E Proof of Proposition 5.1

E.1 Tool box for bounding empirical processes

Definition E.1. Let ψ be a nondecreasing, convex function with $\psi(0) = 0$. The Orlicz norm of a random variable X is defined by

$$\|X\|_\psi := \inf\{t > 0 : \mathbb{E}[\psi(|X|/t)] \le 1\}.$$

For our purposes, Orlicz norms of interest are the ones given by $\psi_p(x) = e^{x^p} - 1$ for $p \ge 1$. In particular, the cases of $p = 1$ and $p = 2$ correspond to the sub-exponential and sub-gaussian random variables, respectively. A random variable with finite ψ_p-norm has the following control of the tail behavior

$$\mathbb{P}\{|X| \ge t\} \le C_1 e^{-C_2 \frac{t^p}{\|X\|_{\psi_p}^p}},$$

where C_1, C_2 are constant that may depend on the value of p.

Lemma E.1.
- *If $X \sim \mathcal{N}(0,\sigma^2)$, X is sub-gaussian with $\|X\|_{\psi_2} \le C\sigma$.*
- *Let X, Y be sub-gaussian random variables. Then, XY is sub-exponential and*

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

- *If $|X| \le |Y|$ a.s., then $\|X\|_\psi \le \|Y\|_\psi$ for any ψ that satisfies the condition in Definition E.1.*

Theorem E.2 (Bernstein's inequality)**.** *Let $X_1, \dots, X_n$ be independent sub-exponential random variables. Suppose $K = \max_i \|X_i\|_{\psi_1} < \infty$. Then, for any $t > 0$, we have*

$$\mathbb{P}\Big\{\big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\big| \ge t\Big\} \le 2\exp\left(-Cn\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right).$$

Proposition E.3 (Sums of independent sub-gaussians)**.** *Let $X_1, \dots, X_n$ be independent, mean zero, sub-gaussian random variables. Then, $\sum_{i=1}^n X_i$ is also a sub-gaussian random variable, and*

$$\|\sum_{i=1}^m X_i\|_{\psi_2}^2 \le C\sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

Lemma E.4 (Centering)**.** *For a random variable X, we have $\|X - \mathbb{E}[X]\|_{\psi_p} \le C\|X\|_{\psi_p}$ for a constant $C > 0$ that may depend on p.*

We refer the reader to (Vershynin, 2018, Section 2) and (Van Der Vaart and Wellner, 1996, Section 2) for the proof of the above properties and more information on Orlicz spaces.

Let (T, ρ) be a semi-metric space, i.e., $\rho(t_1, t_2) \le \rho(t_1, t_3) + \rho(t_3, t_2)$ and $\rho(t_1, t_2) = \rho(t_2, t_1)$ for any $t_1, t_2, t_3 \in T$. We denote the diameter of T with respect to ρ by $\mathrm{diam}(T) = \sup_{s,t\in T}\rho(s,t)$.

Definition E.2 (Sub-gaussian process)**.** Consider a random process $(X_t)_{t\in T}$ on a semi-metric space (T,ρ). We say that the process is a sub-gaussian process if there exits $K\geq 0$ such that

$$\|X_t - X_s\|_{\psi_2} \leq K\rho(t,s) \qquad \forall\, t,s\in T.$$

The following theorem gives a bound of a sub-gaussian process $(X_t)_{t\in T}$ in terms of the Dudley integral

$$J(\delta) = \int_{\delta}^{\mathrm{diam}(T)} \sqrt{\log N(T,\rho,\varepsilon)}\mathrm{d}\varepsilon,$$

where $N(T,\rho,\varepsilon)$ is the ε-covering number of T with respect to ρ.

Theorem E.5 (Theorem 8.1.6 in (Vershynin, 2018))**.** *Let $(X_t)_{t\in T}$ be a mean zero sub-gaussian process as in E.2 on a semi-metric space (T,ρ). Then, there exist $C>0$ such that for any $u>0$, we have with probability $1-2e^{-u^2}$ that*

$$\sup_{t\in T}|X_t| \leq CK\left(J(0)+\mathrm{diam}(T)u\right). \tag{24}$$

Some facts Here, we state some facts which will repeatedly used in the subsequent analysis. Consider the metric space $B_Q = \{\mathbf{w}\in\mathbb{R}^d : \|\mathbf{w}-\mathbf{w}^*\|\leq Q\}$ with $\|\cdot\|_2$. Following Corollary 4.2.13 of (Vershynin, 2018), we have

$$N(B_Q,\|\cdot\|_2,\varepsilon) \leq \left(\frac{2Q}{\varepsilon}+1\right)^d, \tag{25}$$

where we omit the dependence on $\mathbf{w}^*$ since it holds for any $\mathbf{w}^*\in\mathbb{R}^d$.

For any $M>0$ and $\mathbf{x}\in\mathbb{R}^d$, we define $\mathbf{x}^M := \mathbf{x}\min(1,\frac{M}{\|\mathbf{x}\|})$. Hence, if $\|\mathbf{x}\|\leq M$, $\mathbf{x}^M=\mathbf{x}$. Let $\boldsymbol{X}\sim\mathcal{N}(0,I_d)$. Then, $\|\boldsymbol{X}\|^2=\sum_{i=1}^d X_i^2$ follows the χ^2_d distribution. Following Eq. (3.1) in (Vershynin, 2018), we have for $M\geq 2d$,

$$\mathbb{P}\{\|\boldsymbol{X}\|^2\geq M\}\leq 2e^{-CM}. \tag{26}$$

Moreover, for any $\mathbf{u}\in\mathbb{R}^d$, $|\mathbf{u}^T\boldsymbol{X}^M| = |\mathbf{u}^T\boldsymbol{X}\min(1,M/\|\boldsymbol{X}\|)|\leq|\mathbf{u}^T\boldsymbol{X}|$. By Lemma E.1, we have

$$\|\mathbf{u}^T\boldsymbol{X}^M\|_{\psi_2}\leq\|\mathbf{u}^T\boldsymbol{X}\|_{\psi_2}\leq C\|\mathbf{u}\|. \tag{27}$$

E.2 Bounding the difference of loss function

In this subsection, we let $T_Q = B_Q(\mathbf{w}^*)$ and $\rho(\mathbf{w}_1,\mathbf{w}_2)=\|\mathbf{w}_1-\mathbf{w}_2\|$. Consider $f_{\mathbf{w}}(\mathbf{x}) = (\sigma(\mathbf{w}^T\mathbf{x})-\sigma(\mathbf{w}^{*T}\mathbf{x}))^2$ and define the empirical process $(Z_{\mathbf{w}})_{\mathbf{w}\in T_Q}$:

$$Z_{\mathbf{w}} := \hat{\mathcal{R}}_n(\mathbf{w})-\mathcal{R}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_{\mathbf{w}}(\boldsymbol{X}_i)-\mathbb{E}[f_{\mathbf{w}}(\boldsymbol{X})], \tag{28}$$

where $\boldsymbol{X}_i\overset{iid}{\sim}\mathcal{N}(0,I_d)$. Define the truncated version as follows

$$Z^M_{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^n f_{\mathbf{w}}(\boldsymbol{X}_i^M)-\mathbb{E}[f_{\mathbf{w}}(\boldsymbol{X}^M)], \tag{29}$$

Then, we can bound $(Z_{\mathbf{w}})_{\mathbf{w}\in T_Q}$ using the following decomposition

$$\sup_{\mathbf{w}\in T_Q}|Z_{\mathbf{w}}|\leq\sup_{\mathbf{w}\in T_Q}|Z_{\mathbf{w}}-Z^M_{\mathbf{w}}|+\sup_{\mathbf{w}\in T_Q}|Z^M_{\mathbf{w}}|. \tag{30}$$

We will estimate the two terms of right hand side separately.

Lemma E.6. *For any $\mathbf{w}\in T_Q$, we have*

$$\begin{aligned}|f_{\mathbf{w}}(\mathbf{x}_1)-f_{\mathbf{w}}(\mathbf{x}_2)| &\leq (Q+1)^2(\|\mathbf{x}_1\|+\|\mathbf{x}_2\|)\|\mathbf{x}_1-\mathbf{x}_2\|\\ |f_{\mathbf{w}_1}(\mathbf{x}^M)-f_{\mathbf{w}_2}(\mathbf{x}^M)| &\leq 2QM|(\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{x}^M|.\end{aligned} \tag{31}$$

Proof. We first have

$$\begin{aligned}|f_{\mathbf{w}}(\mathbf{x}_1) - f_{\mathbf{w}}(\mathbf{x}_2)| &= |(\sigma(\mathbf{w}^T\mathbf{x}_1) - \sigma(\mathbf{w}^{*T}\mathbf{x}_1))^2 - (\sigma(\mathbf{w}^T\mathbf{x}_2) - \sigma(\mathbf{w}^{*T}\mathbf{x}_2))^2| \\ &= (\sigma(\mathbf{w}^T\mathbf{x}_1) - \sigma(\mathbf{w}^{*T}\mathbf{x}_1) + \sigma(\mathbf{w}^T\mathbf{x}_2) - \sigma(\mathbf{w}^{*T}\mathbf{x}_2)) \\ &\quad \cdot (\sigma(\mathbf{w}^T\mathbf{x}_1) - \sigma(\mathbf{w}^{*T}\mathbf{x}_1) - \sigma(\mathbf{w}^T\mathbf{x}_2) + \sigma(\mathbf{w}^{*T}\mathbf{x}_2)) \\ &\leq (Q+1)^2(\|\mathbf{x}_1\| + \|\mathbf{x}_2\|)\|\mathbf{x}_1 - \mathbf{x}_2\|,\end{aligned}$$

where the last inequality is due to that σ is 1-Lipschitz and $\|\mathbf{w} - \mathbf{w}^*\| \leq Q$. Then,

$$\begin{aligned}|f_{\mathbf{w}_1}(\mathbf{x}^M) - f_{\mathbf{w}_2}(\mathbf{x}^M)| &= |(\sigma(\mathbf{w}_1^T\mathbf{x}^M) - \sigma(\mathbf{w}_1^{*T}\mathbf{x}^M))^2 - (\sigma(\mathbf{w}_2^T\mathbf{x}^M) - \sigma(\mathbf{w}^{*T}\mathbf{x}^M))^2| \\ &= |(\sigma(\mathbf{w}_1^T\mathbf{x}^M) + \sigma(\mathbf{w}_2^T\mathbf{x}^M) - 2\sigma(\mathbf{w}^{*T}\mathbf{x}^M))(\sigma(\mathbf{w}_1^T\mathbf{x}^M) - \sigma(\mathbf{w}_2^T\mathbf{x}^M))| \\ &\leq (|(\mathbf{w}_1 - \mathbf{w}^*)^T\mathbf{x}^M| + |(\mathbf{w}_2 - \mathbf{w}^*)^T\mathbf{x}^M|)|(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x}^M| \\ &\leq 2QM|(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x}^M|,\end{aligned}$$

where the third inequality follows from that σ is 1-Lipschitz continuous. □

We then have the following bound of the first term on the right hand side of (30).

Lemma E.7. *For any $\delta \in (0,1)$, with probability $1-\delta$ over the sampling of data, we have*

$$\sup_{\mathbf{w}\in T_Q} |Z_{\mathbf{w}} - Z_{\mathbf{w}}^M| \leq C_1(Q+1)^2\left(d\max\left\{\sqrt{\frac{\log(2/\delta)}{n}}, \frac{\log(2/\delta)}{n}\right\} + e^{-C_2M^2}\right) \tag{32}$$

Proof. Using Lemma E.6 and the fact, $\|\boldsymbol{X}_i^M\| \leq \|\boldsymbol{X}_i\|$, we have

$$\begin{aligned}|Z_{\mathbf{w}} - Z_{\mathbf{w}}^M| &\leq \frac{1}{n}\sum_{i=1}^n |f_{\mathbf{w}}(\boldsymbol{X}_i) - f_{\mathbf{w}}(\boldsymbol{X}_i^M)| + \mathbb{E}[|f_{\mathbf{w}}(\boldsymbol{X}) - f_{\mathbf{w}}(\boldsymbol{X}^M)|] \\ &\leq \frac{2(Q+1)^2}{n}\sum_{i=1}^n \|\boldsymbol{X}_i\|\|\boldsymbol{X}_i - \boldsymbol{X}_i^M\| + 2(Q+1)^2\,\mathbb{E}[\|\boldsymbol{X}\|\|\boldsymbol{X} - \boldsymbol{X}^M\|] \\ &= \frac{2(Q+1)^2}{n}\sum_{i=1}^n (V_i^M - \mathbb{E}[V^M]) + 4(Q+1)^2\,\mathbb{E}[V^M],\end{aligned} \tag{33}$$

where we let $V^M = \|\boldsymbol{X}\|\|\boldsymbol{X} - \boldsymbol{X}^M\| = \|\boldsymbol{X}\|^2(1 - \min(1, M/\|\boldsymbol{X}\|))$. Then,

$$\|V^M\|_{\psi_1} \leq \|\|\boldsymbol{X}\|^2\|_{\psi_1} \leq Cd.$$

By Theorem E.2, we have

$$\mathbb{P}\Big\{\Big|\frac{1}{n}\sum_{i=1}^n V_i^M - \mathbb{E}[V^M]\Big| \geq t\Big\} \leq 2\exp\left(-Cn\min\left(\frac{t^2}{d^2}, \frac{t}{d}\right)\right). \tag{34}$$

By (26), we have

$$\begin{aligned}\mathbb{E}[V^M] &= \int_0^\infty \mathbb{P}\{V^M \geq t\}\mathrm{d}t = \int_0^\infty \mathbb{P}\{\|\boldsymbol{X}\|\,(\|\boldsymbol{X}\| - \min\{M, \|\boldsymbol{X}\|\}) \geq t\}\mathrm{d}t \\ &= \int_0^\infty \mathbb{P}\{\|\boldsymbol{X}\|^2 - M\|\boldsymbol{X}\| \geq t\}\mathrm{d}t = \int_0^\infty \mathbb{P}\{\|\boldsymbol{X}\| \geq \sqrt{t + M^2/4} + M/2\}\mathrm{d}t \\ &\leq \int_0^\infty 2e^{-C\left(\sqrt{t+M^2/4}+M/2\right)^2}\mathrm{d}t \leq 2e^{-CM^2/2}\int_0^\infty e^{-Ct}\mathrm{d}t = \frac{2}{C}e^{-CM^2/2}.\end{aligned} \tag{35}$$

Combining (34) and (35) and taking RHS of (34)=δ, we complete the proof. □

We proceed to bound the second term on the right hand side of (30).

Lemma E.8. *For any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\sup_{\mathbf{w}\in T_Q} |Z_{\mathbf{w}}^M| \lesssim \frac{MQ^2}{\sqrt{n}}(\sqrt{d}+\sqrt{\log(\delta/2)}).$$

Proof. By Lemma E.1 and E.6 , we have

$$\|f_{\mathbf{w}_1}(\boldsymbol{X}^M) - f_{\mathbf{w}_2}(\boldsymbol{X}^M)\|_{\psi_2} \le 2QM\|(\mathbf{w}_1-\mathbf{w}_2)^T X^M\|_{\psi_2} \le CQM\|\mathbf{w}_1-\mathbf{w}_2\|.$$

where the last inequality is due to Eq. (27). By Proposition E.3, we have

$$\begin{aligned}
\|Z_{\mathbf{w}_1}^M - Z_{\mathbf{w}_2}^M\|_{\psi_2} &= \Big\|\frac{1}{n}\sum_{i=1}^n (f_{\mathbf{w}_1}(\boldsymbol{X}_i^M) - f_{\mathbf{w}_2}(\boldsymbol{X}_i^M) - \mathbb{E}[f_{\mathbf{w}_1}(\boldsymbol{X}^M)] - \mathbb{E}[f_{\mathbf{w}_2}(\boldsymbol{X}^M)])\Big\|_{\psi_2} && (36)\\
&\le \frac{1}{n}\sqrt{\sum_{i=1}^n \|f_{\mathbf{w}_1}(\boldsymbol{X}_i^M) - f_{\mathbf{w}_2}(\boldsymbol{X}_i^M) - \mathbb{E}[f_{\mathbf{w}_1}(\boldsymbol{X}^M)] - \mathbb{E}[f_{\mathbf{w}_2}(\boldsymbol{X}^M)\|_{\psi_2}^2} && (37)\\
&\le \frac{C}{n}\sqrt{\sum_{i=1}^n \|f_{\mathbf{w}_1}(\boldsymbol{X}_i^M) - f_{\mathbf{w}_2}(\boldsymbol{X}_i^M)\|_{\psi_2}^2} && (38)\\
&\le \frac{CQM\|\mathbf{w}_1-\mathbf{w}_2\|}{\sqrt{n}} = \frac{CQM}{\sqrt{n}}\rho(\mathbf{w}_1,\mathbf{w}_2). && (39)
\end{aligned}$$

It means that $(Z_{\mathbf{w}}^M)_{\mathbf{w}\in T}$ is a sub-gaussian process.

According to (25), the Dudley integral of (T_Q,ρ) satisfies

$$\begin{aligned}
J(0) &= \int_0^{\operatorname{diam}(T_Q)} \sqrt{\log N(T_Q,\rho,\varepsilon)}\mathrm{d}\varepsilon \le \int_0^{2Q}\sqrt{d\log\left(1+\frac{2Q}{\varepsilon}\right)}\mathrm{d}\varepsilon\\
&= 2Q\sqrt{d}\int_1^\infty \frac{\sqrt{\log(1+s)}}{s^2}\mathrm{d}s \le CQ\sqrt{d}. && (40)
\end{aligned}$$

By Theorem E.5 and (40), with probability $1-2e^{-u^2}$, we have

$$\sup_{\mathbf{w}\in T}|Z_{\mathbf{w}}^M| \lesssim \frac{QM}{\sqrt{n}}\left(J(0)+u\operatorname{diam}(T_Q)\right) \le \frac{Q^2M}{\sqrt{n}}(\sqrt{d}+u). \quad (41)$$

Let the failure probability $2e^{-u^2}=\delta$, and we complete the proof. □

Proposition E.9. *For any $\delta\in(0,1)$, with probability $1-\delta$, we have*

$$\sup_{\|\mathbf{w}-\mathbf{w}^*\|\le Q}|\hat{\mathcal{R}}_n(\mathbf{w})-\mathcal{R}(\mathbf{w})| \lesssim \frac{d(Q+1)^2\sqrt{\log n}}{\sqrt{n}}\max\left\{\sqrt{\log(4/\delta)},\frac{\log(4/\delta)}{\sqrt{n}}\right\}.$$

Proof. Combining Lemma E.7 and E.8, we have, with probability $1-\delta_1-\delta_2$, that

$$\sup_{\mathbf{w}\in T_Q}|Z_{\mathbf{w}}| \lesssim (Q+1)^2\left(d\max\left\{\sqrt{\frac{\log(2/\delta_1)}{n}},\frac{\log(2/\delta_1)}{n}\right\}+e^{-C_2M^2}+\frac{M}{\sqrt{n}}(\sqrt{d}+\sqrt{\log(2/\delta_2)})\right).$$

Taking $M=\sqrt{\frac{\log n}{2C_2}}, \delta_1=\delta_2=\delta/2$, we have

$$\sup_{\mathbf{w}\in T_Q}|Z_{\mathbf{w}}| \lesssim \frac{d(Q+1)^2\sqrt{\log n}}{\sqrt{n}}\max\left\{\sqrt{\log(4/\delta)},\frac{\log(4/\delta)}{\sqrt{n}}\right\}.$$

Noting that $Z_{\mathbf{w}}=\hat{\mathcal{R}}(\mathbf{w})-\mathcal{R}(\mathbf{w})$, we complete the proof. □

E.3 Bounding the difference between gradients

In this subsection, we let $T_Q = B_Q(\mathbf{w}^*) \times \mathbb{S}^{d-1}$ and $\rho(\boldsymbol{t}_1, \boldsymbol{t}_2) = \|\mathbf{w}_1 - \mathbf{w}_2\| + \|\mathbf{u}_1 - \mathbf{u}_2\|$ for $\boldsymbol{t}_1 = (\mathbf{w}_1, \mathbf{u}_1), \boldsymbol{t}_2 = (\mathbf{w}_2, \mathbf{u}_2) \in T_Q$. Let $Y_{\boldsymbol{t}}(\mathbf{x}) := (\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{u}^T\mathbf{x}$. Consider the empirical process $(O_{\boldsymbol{t}})_{\boldsymbol{t}\in T_Q}$:

$$O_{\boldsymbol{t}} = \langle \mathbf{u}, \nabla\hat{\mathcal{R}}_n(\mathbf{w}) - \nabla\mathcal{R}(\mathbf{w})\rangle = \frac{1}{n}\sum_{i=1}^{n} Y_{\boldsymbol{t}}(\boldsymbol{X}_i) - \mathbb{E}[Y_{\boldsymbol{t}}(\boldsymbol{X})]. \tag{42}$$

For any $M > 0$, make the following decomposition

$$\sup_{\boldsymbol{t}\in T_Q} |Q_{\boldsymbol{t}}| \le \sup_{\boldsymbol{t}\in T_Q} |Q_{\boldsymbol{t}} - Q^M_{\boldsymbol{t}}| + \sup_{\boldsymbol{t}\in T_Q} |Q^M_{\boldsymbol{t}}|, \tag{43}$$

where $Q^M_{\boldsymbol{t}}$ is the truncated empirical process defined by

$$O^M_{\boldsymbol{t}} := \frac{1}{n}\sum_{i=1}^{n} Y_{\boldsymbol{t}}(\boldsymbol{X}^M_i) - \mathbb{E}[Y_{\boldsymbol{t}}(\boldsymbol{X}^M)]. \tag{44}$$

We then estimate the two terms on the right hand slide of (43), separately.

Lemma E.10. *Assume $M \ge 1$. For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and $\boldsymbol{t}_1, \boldsymbol{t}_2 \in T_Q$, we have*

$$|Y_{\boldsymbol{t}}(\mathbf{x}_1) - Y_{\boldsymbol{t}}(\mathbf{x}_2)| \lesssim (Q+1)^2 \max_{i=1,2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_i\|)\|\mathbf{x}_1 - \mathbf{x}_2\| \tag{45}$$

$$|Y_{\boldsymbol{t}_1}(\mathbf{x}^M) - Y_{\boldsymbol{t}_2}(\mathbf{x}^M)| \lesssim M(1+QM)(|(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x}| + |(\mathbf{u}_1 - \mathbf{u}_2)^T\mathbf{x}|). \tag{46}$$

Proof. First,

$$\begin{aligned}
\|\nabla_{\mathbf{x}} Y_{\boldsymbol{t}}(\mathbf{x})\| &= \|(\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{w} - \sigma'(\mathbf{w}^{*T}\mathbf{x})\mathbf{w}^*)\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{u}^T\mathbf{x} \\
&\qquad + (\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))(\sigma''(\mathbf{w}^T\mathbf{x})\mathbf{u}^T\mathbf{x}\mathbf{w} + \sigma'(\mathbf{w}^T\mathbf{x})\mathbf{u})\| \\
&\le (\|\mathbf{w}\| + \|\mathbf{w}^*\|)\|\mathbf{x}\| + \|\mathbf{w} - \mathbf{w}^*\|\|\mathbf{x}\|(\|\mathbf{w}\|\|\mathbf{x}\| + 1) \\
&\le 2(Q+1)\|\mathbf{x}\| + Q(Q+1)\|\mathbf{x}\|^2.
\end{aligned}$$

Following the mean value theorem, we have

$$|Y_{\boldsymbol{t}}(\mathbf{x}_1) - Y_{\boldsymbol{t}}(\mathbf{x}_2)| \le 2(Q+1)^2 \max_{i=1,2}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_i\|)\|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Second,

$$\nabla_{\boldsymbol{t}} Y_{\boldsymbol{t}}(\mathbf{x}) = \begin{pmatrix} \left(\sigma'(\mathbf{w}^T\mathbf{x})^2 + (\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x})\sigma''(\mathbf{w}^T\mathbf{x})\right)\mathbf{u}^T\mathbf{x}\mathbf{x} \\ (\sigma(\mathbf{w}^T\mathbf{x}) - \sigma(\mathbf{w}^{*T}\mathbf{x}))\sigma'(\mathbf{w}^T\mathbf{x})\mathbf{x} \end{pmatrix} =: \begin{pmatrix} v_1(\boldsymbol{t}, \mathbf{x})\mathbf{x} \\ v_2(\boldsymbol{t}, \mathbf{x})\mathbf{x} \end{pmatrix}. \tag{47}$$

For $\|\mathbf{x}\| \le M$, it is easy to verify that

$$|v_1(\boldsymbol{t}, \mathbf{x})| \le M(1+QM), \qquad |v_2(\boldsymbol{t}, \mathbf{x})| \le QM.$$

By the mean value theorem, there exists $\boldsymbol{t}'$ such that

$$\begin{aligned}
|Y_{\boldsymbol{t}_1}(\mathbf{x}) - Y_{\boldsymbol{t}_2}(\mathbf{x})| &= |\nabla_{\boldsymbol{t}} Y_{\boldsymbol{t}'}(\mathbf{x})(\boldsymbol{t}_1 - \boldsymbol{t}_2)| = |v_1(\boldsymbol{t}', \mathbf{x})(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x} + v_2(\boldsymbol{t}', \mathbf{x})(\mathbf{u}_1 - \mathbf{u}_2)^T\mathbf{x}| \\
&\lesssim M(1+QM)(|(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x}| + |(\mathbf{u}_1 - \mathbf{u}_2)^T\mathbf{x}|).
\end{aligned} \tag{48}$$

□

We then estimate the first term on the right hand side of (43).

Lemma E.11. *There exists $C_1, C_2, C_3, C_4 > 0$ such that for $M > C_1 d$, with probability $1 - nC_2e^{-C_3M^2}$, we have*

$$\sup_{\boldsymbol{t}\in T} |Q_{\boldsymbol{t}} - Q^M_{\boldsymbol{t}}| \lesssim (Q+1)^2 e^{-C_4M^2}.$$

Proof. Using Lemma E.10 and the fact $\|\boldsymbol{X}_i^M\| \le \|\boldsymbol{X}_i\|$, we have

$$\begin{aligned}
|O_{\boldsymbol{t}} - O_{\boldsymbol{t}}^M| &\le \frac{1}{n}\sum_{i=1}^n |Y_{\boldsymbol{t}}(\boldsymbol{X}_i) - Y_{\boldsymbol{t}}(\boldsymbol{X}_i^M)| + \mathbb{E}[|Y_{\boldsymbol{t}}(\boldsymbol{X}) - Y_{\boldsymbol{t}}(\boldsymbol{X}^M)|] \\
&\lesssim \frac{(Q+1)^2}{n}\sum_{i=1}^n \|\boldsymbol{X}_i\|(1+\|\boldsymbol{X}_i\|)\|\boldsymbol{X}_i - \boldsymbol{X}_i^M\| + (Q+1)^2\,\mathbb{E}[\|\boldsymbol{X}\|(1+\|\boldsymbol{X}\|)\|\boldsymbol{X} - \boldsymbol{X}^M\|] \\
&\lesssim \frac{(Q+1)^2}{n}\sum_{i=1}^n V_i^M + (Q+1)^2\,\mathbb{E}[V^M], \qquad (49)
\end{aligned}$$

where we let

$$V^M = \|\boldsymbol{X}\|(\|\boldsymbol{X}\|+1)\|\boldsymbol{X} - \boldsymbol{X}^M\| = (1+\|\boldsymbol{X}\|)\|\boldsymbol{X}\|^2(1-\min(1, M/\|\boldsymbol{X}\|)).$$

Note that for any $i \in [n]$, $\mathbb{P}\{V_i^M > 0\} = \mathbb{P}\{\|\boldsymbol{X}\| > M\} \le C_1 e^{-C_2 M^2}$ for $M \ge C_3 d$. Taking the union bound, we have

$$\mathbb{P}\{\sum_{i=1}^n V_i^M = 0\} = 1 - \mathbb{P}\{\sum_{i=1}^n V_i^M > 0\} \ge 1 - \sum_i \mathbb{P}\{V_i^M > 0\} \ge 1 - nC_1 e^{-C_2 M^2}. \qquad (50)$$

Similar to (35), we can obtain that

$$\mathbb{E}[V^M] \le C_1 e^{-C_2 M^2}, \qquad (51)$$

for $M \ge C_4 d$ with C_4 large enough.

Combining (50) and (51) completes the proof. □

Before proceeding to the estimate of the second term on the right hand side of (43), we first bound the Dudley integral of the metric space.

Lemma E.12. *The Dudley integral of (T_Q, ρ) satisfies $J(0) \lesssim \sqrt{d}(Q+1)$.*

Proof. Note that

$$N(T_Q, \rho, \varepsilon) \le N(B_Q(\mathbf{w}^*), \|\cdot\|, \frac{\varepsilon}{2}) N(\mathbb{S}^{d-1}, \|\cdot\|, \frac{\varepsilon}{2}) \le \left(\frac{4Q}{\varepsilon} + 1\right)^d \left(\frac{2}{\varepsilon}\right)^d.$$

Moreover, $\mathrm{diam}(T) \le 2Q + 2$. Hence, the Dudley integral is given by

$$\begin{aligned}
J(0) &= \int_0^{2Q+2} \sqrt{\log N(T, \rho, \varepsilon)}\mathrm{d}\varepsilon \\
&\le \sqrt{d}\int_0^{2Q+2} \sqrt{\log(1 + \frac{4Q}{\varepsilon}) + \log(2/\varepsilon)}\mathrm{d}\varepsilon \\
&\le \sqrt{d}\int_0^{2Q+2} \sqrt{\log(1 + \frac{4Q}{\varepsilon})}\mathrm{d}\varepsilon + \sqrt{d}\int_0^{2Q+2} \sqrt{\log(2/\varepsilon)}\mathrm{d}\varepsilon \\
&\le \sqrt{d}4Q\int_{\frac{2Q}{Q+1}}^{\infty} \frac{\sqrt{\log(1+s)}}{s^2}\mathrm{d}s + 2\sqrt{d}\int_{\frac{1}{Q+1}}^{\infty} \frac{\sqrt{\log s}}{s^2}\mathrm{d}s \\
&\lesssim \sqrt{d}(Q+1). \qquad (52)
\end{aligned}$$

□

Lemma E.13. *For any $u > 0$, with probability $1 - 2e^{-u^2}$, we have*

$$\sup_{\mathbf{w}\in T} |Z_{\mathbf{w}}^M| \le \frac{(Q+1)^2 M^2}{\sqrt{n}}(\sqrt{d} + u).$$

Proof. By Lemma E.1 and E.10 , we have

$$\begin{aligned}\|Y_{\boldsymbol{t}_1}(\boldsymbol{X}^M) - Y_{\boldsymbol{t}_2}(\boldsymbol{X}^M)\|_{\psi_2} &\lesssim M(1+QM)\||(\mathbf{w}_1-\mathbf{w}_2)^T\boldsymbol{X}^M| + |(\mathbf{u}_1-\mathbf{u}_2)^T\boldsymbol{X}^M|\|_{\psi_2}\\ &\lesssim M(1+QM)(\|\mathbf{w}_1-\mathbf{w}_2\| + \|\mathbf{u}_1-\mathbf{u}_2\|)\\ &= M(1+QM)\rho(\boldsymbol{t}_1,\boldsymbol{t}_2). \end{aligned} \tag{53}$$

where the second inequality is due to Eq. (27). By Proposition E.3, we have

$$\begin{aligned}\|O^M_{\boldsymbol{t}_1} - O^M_{\boldsymbol{t}_2}\|_{\psi_2} &= \Big\|\frac{1}{n}\sum_{i=1}^n (Y_{\boldsymbol{t}_1}(\boldsymbol{X}_i^M) - Y_{\boldsymbol{t}_2}(\boldsymbol{X}_i^M) - \mathbb{E}[Y_{\boldsymbol{t}_1}(\boldsymbol{X}^M)] - \mathbb{E}[Y_{\boldsymbol{t}_2}(\boldsymbol{X}^M)])\Big\|_{\psi_2}\\ &\lesssim \frac{1}{n}\sqrt{\sum_{i=1}^n \|Y_{\boldsymbol{t}_1}(\boldsymbol{X}_i^M) - Y_{\boldsymbol{t}_2}(\boldsymbol{X}_i^M) - \mathbb{E}[Y_{\boldsymbol{t}_1}(\boldsymbol{X}^M)] - \mathbb{E}[Y_{\boldsymbol{t}_2}(\boldsymbol{X}^M)]\|_{\psi_2}^2}\\ &\lesssim \frac{1}{n}\sqrt{\sum_{i=1}^n \|Y_{\boldsymbol{t}_1}(\boldsymbol{X}_i^M) - Y_{\boldsymbol{t}_2}(\boldsymbol{X}_i^M)\|_{\psi_2}^2}\\ &\lesssim \frac{M(1+QM)}{\sqrt{n}}\rho(\boldsymbol{t}_1,\boldsymbol{t}_2). \end{aligned} \tag{54}$$

It means that $(O^M_{\boldsymbol{t}})_{\boldsymbol{t}\in T_Q}$ is a sub-gaussian process. Moreover, $\operatorname{diam}(T_Q) = 2(Q+1)$.

By Theorem E.5 and Lemma E.12, with probability $1-2e^{-u^2}$, we have

$$\sup_{\mathbf{w}\in T}|Z^M_{\mathbf{w}}| \lesssim \frac{M(1+QM)}{\sqrt{n}}\left(J(0) + u\operatorname{diam}(T_Q)\right) \le \frac{(Q+1)^2M^2}{\sqrt{n}}(\sqrt{d}+u). \tag{55}$$

□

Proposition E.14. *Assume $n \ge 3$. For any $\delta \in (0,1)$, with probability $1-\delta$, we have*

$$\sup_{\|\mathbf{w}-\mathbf{w}^*\|\le Q}|\nabla\hat{\mathcal{R}}_n(\mathbf{w}) - \nabla\mathcal{R}(\mathbf{w})| \lesssim \frac{\sqrt{d}\log^{3/2}(n/\delta)}{\sqrt{n}}(Q+1)^2.$$

Proof. Combining Lemma E.11 and E.13, with probability $(1-nC_2e^{-C_3M^2})(1-2e^{-u^2})$, we have for $M \ge C_1 d$,

$$\sup_{\mathbf{w}\in T_Q}|Q_{\boldsymbol{t}}| \lesssim (Q+1)^2\left(e^{-C_4M^2} + \frac{M^2}{\sqrt{n}}(\sqrt{d}+u)\right).$$

Taking $M^2 = \frac{\log(2nC_2/\delta)}{\min(C_3,C_4)}$ and $u = \sqrt{\log(4/\delta)}$, we have with probability $(1-\delta/2)^2 \ge 1-\delta$ that

$$\sup_{\mathbf{w}\in T_Q}|Q_{\boldsymbol{t}}| \le C(Q+1)^2\left(\frac{\delta}{2C_2n} + \frac{\log(2C_2n/\delta)}{\min(C_3,C_4)\sqrt{n}}(\sqrt{d}+\sqrt{\log(4/\delta)})\right).$$

Assuming that $n \ge 3$, the above inequality can be simplified as follows

$$\sup_{\mathbf{w}\in T_Q}|Q_{\boldsymbol{t}}| \lesssim \frac{\sqrt{d}\log^{3/2}(n/\delta)}{\sqrt{n}}(Q+1)^2.$$

Noting that

$$\sup_{\mathbf{w}\in B_Q(\mathbf{w}^*)}\|\nabla\hat{\mathcal{R}}_n(\mathbf{w}) - \nabla\mathcal{R}(\mathbf{w})\| = \sup_{\mathbf{w}\in B_Q(\mathbf{w}^*)}\sup_{\mathbf{u}\in\mathbb{S}^{d-1}}\mathbf{u}^T(\nabla\hat{\mathcal{R}}_n(\mathbf{w}) - \nabla\mathcal{R}(\mathbf{w})) = \sup_{\boldsymbol{t}\in T_Q}O_{\boldsymbol{t}}.$$

we complete the proof. □