Proceedings of Machine Learning Research vol 125:1–31, 2020 33rd Annual Conference on Learning Theory

Learning a Single Neuron with Gradient Methods

Gilad Yehudai and **Ohad Shamir**
{GILAD.YEHUDAI, OHAD.SHAMIR}@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider the fundamental problem of learning a single neuron $\mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x})$ in a realizable setting, using standard gradient methods with random initialization, and under general families of input distributions and activations. On the one hand, we show that some assumptions on both the distribution and the activation function are necessary. On the other hand, we prove positive guarantees under mild assumptions, which go significantly beyond those studied in the literature so far. We also point out and study the challenges in further strengthening and generalizing our results.

1. Introduction

In recent years, much effort has been devoted to understanding why neural networks are successfully trained with simple, gradient-based methods, despite the inherent non-convexity of the learning problem. However, our understanding of this is still partial at best.

In this paper, we focus on the simplest possible nonlinear neural network, composed of a single neuron, of the form $\mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x})$, where $\mathbf{w}$ is the parameter vector and $\sigma : \mathbb{R} \to \mathbb{R}$ is some fixed non-linear activation function. Moreover, we consider a realizable setting, where the inputs are sampled from some distribution $\mathcal{D}$, the target values are generated by some unknown target neuron $\mathbf{x} \mapsto \sigma(\mathbf{v}^\top \mathbf{x})$ (possibly corrupted by independent zero-mean noise, and where we generally assume $\|\mathbf{v}\| = 1$ for simplicity), and we wish to train our neuron with respect to the squared loss. Mathematically, this boils down to minimizing the following objective function:

$$F(\mathbf{w}) := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\frac{1}{2}\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)^2\right]. \tag{1}$$

For this problem, we are interested in the performance of gradient-based methods, which are the workhorse of modern machine learning systems. These methods initialize $\mathbf{w}$ randomly, and proceed by taking (generally stochastic) gradient steps w.r.t. F. If we hope to explain the success of such methods on complicated neural networks, it seems reasonable to expect a satisfying explanation for their convergence on single neurons.

Although the learning of single neurons was studied in a number of papers (see the related work section below for more details), the existing analyses all suffer from one or several limitations: Either they apply for a specific distribution $\mathcal{D}$, which is convenient to analyze but not very practical (such as a standard Gaussian distribution); Apply to gradient methods only with a specific initialization (rather than a standard random one); Require technical conditions on the input distribution which are not generally easy to verify; Or require smoothness and strict monotonicity conditions on the activation function $\sigma(\cdot)$ (which excludes, for example, the common ReLU function $\sigma(z) = \max\{0, z\}$). However, a bit of experimentation strongly suggests that none of these restrictions is really necessary for standard gradient methods to succeed on this simple problem. Thus, our understanding of this problem is probably still incomplete.

The goal of this paper is to study to what extent the limitations above can be removed, with the following contributions:

- We begin by asking whether positive results are possible without *any* explicit assumptions on the distribution $\mathcal{D}$ or the activation $\sigma(\cdot)$ (other than, say, bounded support for the former and Lipschitz continuity for the latter). Although this seems reasonable at first glance, we show in Sec. 3 that unfortunately, this is not the case: Even for the ReLU activation function, there are bounded distributions $\mathcal{D}$ on which gradient descent will fail to optimize Eq. (1) with probability exponentially close to 1. Moreover, even for $\mathcal{D}$ which is a standard Gaussian, there are Lipschitz activation functions on which gradient methods will likely fail.

- Motivated by the above, we ask whether it is possible to prove positive results with *mild and transparent* assumptions on the distribution and activation function, which does not exclude common setups. In Sec. 4, we prove a key technical result, which implies that if the distribution $\mathcal{D}$ is sufficiently "spread" and the activation function satisfies a weak monotonicity condition (satisfied by ReLU and all standard activation functions), then $\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{v} \rangle$ is positive in most of the domain. This implies that an exact gradient step with sufficiently small step size will bring us closer to $\mathbf{v}$ in "most" places. Building on this result, we prove in Sec. 5 a constant-probability convergence guarantee for several variants of gradient methods (gradient descent, stochastic gradient descent, and gradient flow) with random initialization.

- In Sec. 6, we consider more specifically the case where $\mathcal{D}$ is any spherically symmetric distribution (which includes the standard Gaussian as a special case) and the ReLU activation function. In this setting, we show that the convergence results can be made to hold with high probability, due to the fact that the angle between the parameter vector and the target vector $\mathbf{v}$ motonically decreases. As we discuss later on, the case of the ReLU function and a standard Gaussian distribution was also considered in Soltanolkotabi (2017); Kalan et al. (2019), but that analysis crucially relied on initialization at the origin and a Gaussian distribution, whereas our results apply to more generic initialization schemes and distributions.

- A natural question arising from these results is whether a high-probability result can be proved for non-spherically symmetric distributions. We study this empirically in Subsection 6.2, and show that perhaps surprisingly, the angle to the target function might *increase* rather than decrease, already when we consider unit-variance Gaussian distributions with a non-zero mean. This suggests that a fundamentally different approach would be required for a general high-probability guarantee.

Overall, we hope our work contributes to a better understanding of the dynamics of gradient methods on simple neural networks, and suggests some natural avenues for future research.

1.1. Related Work

First, we emphasize that learning a single target neuron is *not* an inherently difficult problem: Indeed, it can be efficiently performed with minimal assumptions, using the Isotron algorithm and its variants (Kalai and Sastry (2009); Kakade et al. (2011)). Also, other algorithms exist for even more complicated networks or more general settings, under certain assumptions (e.g., Goel et al. (2016); Janzamin et al. (2015)). However, these are non-standard algorithms, whereas our focus here is on standard, vanilla gradient methods.

For this setting, a positive result was provided in Mei et al. (2016), showing that gradient descent on the empirical risk function $\frac{1}{n}\sum_{i=1}^{n}(\sigma(\mathbf{x}_i^\top\mathbf{w})-\sigma(\mathbf{x}_i^\top\mathbf{v}))^2$ (with $\mathbf{x}_i$ sampled i.i.d. from $\mathcal{D}$ and n sufficiently large) successfully yields a good approximation of $\mathbf{v}$. However, the analysis requires σ to be strictly monotonic, and to have uniformly bounded derivatives up to the third order. This excludes standard activation functions such as the ReLU, which are neither strictly monotonic nor differentiable. Indeed, assuming that the activation is strictly monotonic makes the analysis much easier, as we show later on in Thm. 3.2. A related analysis under strict monotonicity conditions is provided in Oymak and Soltanolkotabi (2018).

For the specific case of a ReLU activation function $\sigma(\cdot)=\max\{\cdot,0\}$ and a standard Gaussian input distribution, Tian (2017) proved that with constant probability, gradient flow over Eq. (1) will asymptotically converge to the global minimum. Soltanolkotabi (2017) and Kalan et al. (2019) considered a similar setting, and proved a non-asymptotic convergence guarantee for gradient descent or stochastic gradient descent on the empirical risk function $\frac{1}{n}\sum_{i=1}^{n}(\sigma(\mathbf{x}_i^\top\mathbf{w})-\sigma(\mathbf{x}_i^\top\mathbf{v}))^2$. However, that analysis crucially relied on initialization at precisely $\mathbf{0}$, as well as a certain assumption on how the derivative of the ReLU function is computed at 0. In more details, we impose the convention that even though the ReLU function is not differentiable at 0, we take $\sigma'(0)$ to be some fixed positive number, and the gradient of the population objective F at $\mathbf{0}$ to be

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[(\sigma(0)-\sigma(\mathbf{v}^\top\mathbf{x}))\sigma'(0)\mathbf{x}\right] \;=\; -\,\sigma'(0)\cdot\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\sigma(\mathbf{v}^\top\mathbf{x})\mathbf{x}\right]\,.$$

Assuming $\sigma'(0)>0$, we get that the gradient is non-zero and proportional to $-\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\sigma(\mathbf{v}^\top\mathbf{x})\mathbf{x}]$. For a Gaussian distribution (and more generally, spherically symmetric distributions), this turns out to be proportional to $-\mathbf{v}$, so that an exact gradient step from $\mathbf{0}$ will lead us precisely in the direction of the target parameter vector $\mathbf{v}$. As a result, if we calculate a sufficiently precise approximation of this direction from a random sample, we can get *arbitrarily close to* $\mathbf{v}$ *in a single iteration* (see Kalan et al. (2019, Remark 1) for a discussion of this). Unfortunately, this unique behavior is specific to initialization at $\mathbf{0}$ with a certain convention about $\sigma'(0)$ (note that even locally around $\mathbf{0}$, the gradient may not approximate $\mathbf{v}$, since it is generally discontinuous around $\mathbf{0}$). Thus, although the analysis is important and insightful, it is difficult to apply more generally.

Du et al. (2017) considered conditions under which a single ReLU convolutional filter is learnable with gradient methods, a special case of which is a single ReLU neuron. The paper is closely related to our work, in the sense that they were also motivated by finding general conditions under which positive results are attainable. Moreover, some of the techniques they employed share similarities with ours (e.g., considering the gradient correlation as in Sec. 4). However, our results differ in several aspects: First, they consider only the ReLU activation function, while we also consider general activations. Second, their results assume a technical condition on the eigenvalues of certain distribution-dependent matrices, with the convergence rate depending on these eigenvalues. However, the question of when might this condition hold (for general distributions) is left unclear. In contrast, our assumptions are more transparent and have a clear geometric intuition. Third, their results hold with constant probability, even for a standard Gaussian distribution, while we employ a different analysis to prove high probability guarantees for general spherically symmetric distributions. Finally, we also provide negative results, showing the necessity of assumptions on both the activation function and the input distribution, as well as suggesting which approaches might not work for further generalizing our results.

A line of recent works established the effectiveness of gradient methods in solving non-convex optimization problems with a *strict saddle* property, which implies that all near-stationary points with nearly positive definite Hessians are close to global minima (see Jin et al. (2017); Ge et al. (2015); Sun et al. (2015)). A relevant example is phase retrieval, which actually fits our setting with $\sigma(\cdot)$ being the quadratic function $z \mapsto z^2$ (Sun et al. (2018)). However, these results can only be applied to smooth problems, where the objective function is twice differentiable with Lipschitz-continuous Hessians (excluding, for example, problems involving the ReLU activation function). An interesting recent exception is the work of Tan and Vershynin (2019), which considered the case $\sigma(z) = |z|$. However, their results are specific to that activation, and assumes a specific input distribution $\mathcal{D}$ (uniform on a scaled origin-centered sphere). In contrast, our focus here is on more general families of distributions and activations.

Brutzkus and Globerson (2017) show that gradient descent learns a simple convolutional network with non-overlapping patches, when the inputs have a standard Gaussian distribution. Similar to the analysis in Sec. 6 in our paper, they rely on showing that the angle between the learned parameter vector and a target vector monotonically decreases with gradient methods. However, the network architecture studied is different than ours, and their proof heavily relies on the symmetry of the Gaussian distribution.

Less directly related to our setting, a popular line of recent works showed how gradient methods on highly over-parameterized neural networks can learn various target functions in polynomial time (e.g., Allen-Zhu et al. (2019); Daniely (2017); Arora et al. (2019); Cao and Gu (2019)). However, as pointed out in Yehudai and Shamir (2019), this type of analysis cannot be used to explain learnability of single neurons.

2. Preliminaries

Notation. We use bold-faced letters to denote vectors. For a vector $\mathbf{w}$, we let w_i denote its i-th coordinate. We denote $[z]_+ := \max\{0, z\}$ to be the ReLU function. For a vector $\mathbf{w}$, we let $\bar{\mathbf{w}} := \frac{\mathbf{w}}{\|\mathbf{w}\|}$, and by $\mathbf{1}$ we denote the all-ones vector $(1, \ldots, 1)$. Given vectors $\mathbf{w}, \mathbf{v}$ we let $\theta(\mathbf{w}, \mathbf{v}) := \arccos\left(\frac{\mathbf{w}^\top \mathbf{v}}{\|\mathbf{w}\|\|\mathbf{v}\|}\right) = \arccos(\bar{\mathbf{w}}^\top \bar{\mathbf{v}}) \in [0, \pi]$ denote the angle between $\mathbf{w}$ and $\mathbf{v}$. We use $\mathcal{P}$ to denote probability. $\mathbb{1}(\cdot)$ denotes the indicator function, for example $\mathbb{1}(x > 0)$ equals 1 if $x > 0$ and 0 otherwise.

Target Neuron. Unless stated otherwise, we assume that the target vector $\mathbf{v}$ in Eq. (1) is unit norm, $\|\mathbf{v}\| = 1$.

Gradients. When $\sigma(\cdot)$ is differentiable, the gradient of the objective function in Eq. (1) is

$$\nabla F(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right) \cdot \sigma'(\mathbf{w}^\top \mathbf{x})\mathbf{x}\right] \tag{2}$$

When $\sigma(\cdot)$ is not differentiable, we will still assume that it is differentiable almost everywhere (up to a finite number of points), and that in every point of non-differentiability z, there are well-defined left and right derivatives. In that case, practical implementations of gradient methods fix $\sigma'(z)$ to be some number between its left and right derivatives (for example, for the ReLU function, $\sigma'(0)$ is defined as some number in $[0, 1]$). Following that convention, the expected gradient used by these methods still corresponds to Eq. (2), and we will follow the same convention here.

Algorithms. In our paper, we focus on the following three standard gradient methods:

- **Gradient Descent**: We initialize at some $\mathbf{w}_0$ and set a fixed learning rate η. At each iteration $t > 0$, we do a single step in the negative direction of the gradient: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla F(\mathbf{w}_t)$.
- **Stochastic Gradient Descent (SGD)**: We initialize at some $\mathbf{w}_0$ and set a fixed learning rate η. At each iteration $t > 0$, we sample an input $\mathbf{x}_t \sim \mathcal{D}$, and calculate a stochastic gradient:

$$g_t = \left(\sigma(\mathbf{w}_t^\top \mathbf{x}_t) - \sigma(\mathbf{v}^\top \mathbf{x}_t)\right) \cdot \sigma'(\mathbf{w}_t^\top \mathbf{x}_t)\mathbf{x}_t \tag{3}$$

  and do a single step in the negative direction of the stochastic gradient: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta g_t$. Note that here we consider SGD on the population loss, which is different from SGD on a fixed training set. We also note that our proof techniques easily extend to mini-batch SGD, where g_t is taken to be the average of B stochastic gradients w.r.t. $\mathbf{x}_t^1, \ldots, \mathbf{x}_t^B$ sampled i.i.d. from $\mathcal{D}$. However, for simplicity we will focus on $B = 1$.
- **Gradient Flow**: We initialize at some $\mathbf{w}(0)$, and for every $t > 0$, we set $\mathbf{w}(t)$ to be the solution of the differential equation: $\dot{\mathbf{w}}(t) = -\nabla F(\mathbf{w}(t))$. This can be thought of as a continuous form of gradient descent, where we consider an infinitesimal learning rate. We note that strictly speaking, gradient flow is not an algorithm. However, it approximates the behavior of gradient descent in many cases, and has the advantage that its analysis is often simpler.

3. Assumptions on the Distribution and Activation are Necessary

The main concern of this paper is under what assumptions can a single neuron be provably learned with gradient methods. In this section, we show that perhaps surprisingly, this is not possible unless we make non-trivial assumptions on *both* the input distribution and the activation function.

3.1. Assumptions on the Input Distribution are Necessary

We begin by asking whether Eq. (1) can be minimized by gradient methods in a distribution-free manner (with no assumptions beyond, say, bounded support), as in learning problems where the population objective is convex. Perhaps surprisingly, we show that the answer is negative, even if we consider specifically the ReLU activation, and a distribution supported on the unit Euclidean ball. This is based on the following key result:

Theorem 3.1 *Suppose that σ is the ReLU function (with the convention that $\sigma'(z) = \mathbb{1}(z > 0)$), and assume that $\mathbf{w}$ is sampled from a product distribution $D_{\mathbf{w}}$ (namely, each w_i is sampled independently from some distribution $D_{\mathbf{w}}^i$). Then there exists a distribution $\mathcal{D}$ over the inputs, supported on $\{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$, and $\mathbf{v}$ with $\|\mathbf{v}\| = 1$ such that the following holds: With probability at least $1 - \exp\left(-\frac{d}{4}\right)$ over the initialization point sampled from $D_{\mathbf{w}}$, if we run gradient flow, gradient descent or stochastic gradient descent, then for every $t > 0$ we have $F(\mathbf{w}_t) - \inf_{\mathbf{w}} F(\mathbf{w}) \geq \frac{1}{8d}$ (and for gradient flow, $F(\mathbf{w}(t)) - \inf_{\mathbf{w}} F(\mathbf{w}) \geq \frac{1}{8d}$).*

The full proof can be found in Appendix A. Thm. 3.1 applies to any product initialization scheme, which includes most standard initializations used in practice (e.g., the standard Xavier initialization, see Glorot and Bengio (2010)). The theorem implies that it is impossible to prove positive guarantees in our setting without distributional assumptions on ths inputs. Inspecting the construction, the source of the problem (at least for the ReLU neuron) appears to be the fact that the distribution is supported on a small number of well-separated regions. Thus, in our positive results, we will assume that the distribution is sufficiently "spread", as formalized later on in Sec. 4

3.2. Assumptions on the Activation Function

We now turn to discuss the activation function, explaining why even if the activation is Lipschitz and the input distribution $\mathcal{D}$ is a standard Gaussian, this is likely insufficient for positive guarantees in our setting.

In particular, let us consider the case that $\sigma(\cdot)$ is a 1-Lipschitz *periodic* function. Then Theorem 3 in Shamir (2018) implies that for a large family of input distributions $\mathcal{D}$ on $\mathbb{R}^d$ (including a standard Gaussian), if we assume that the vector $\mathbf{v}$ in the target neuron $\sigma(\mathbf{v}^\top \mathbf{x})$ is a uniformly distributed unit vector, then for any fixed $\mathbf{w}$,$Var_{\mathbf{v}}(\nabla F(\mathbf{w})) \leq \mathcal{O}(\exp(-d))$. This implies that the gradient at $\mathbf{w}$ is virtually independent of the underlying target vector $\mathbf{v}$: In fact, it is extremely concentrated around a fixed value which does not depend on $\mathbf{v}$. Theorem 4 from Shamir (2018) goes further and shows that for any gradient method, even an exponentially small amount of noise will be enough to make its trajectory (after at most $\exp(\mathcal{O}(d))$ iterations) independent of $\mathbf{v}$, in which case it cannot possibly succeed in this setting. We note that their result is even more general as they consider a general function $f(\mathbf{w}, \mathbf{x})$ instead of $\sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$, so our setting can be seen as a private case.

When considering a standard Gaussian distribution, the above argument can be easily extended to activations σ which are periodic only in a segment of length $\Omega(d)$ around the origin. This can be seen by extending the activation to $\tilde{\sigma}$ which is periodic on $\mathbb{R}$, applying the above argument to it, and noting that the probability mass outside of a ball of radius $\Omega(d)$ is exponentially small (for example, see Yehudai and Shamir (2019) Proposition 4.2, where they consider an activation which is a finite sum of ReLU functions and periodic in a segment of length $O(d^2)$).

The above discussion motivates us to impose some condition on the activation function which excludes periodic functions. One such mild assumptions, which we will adopt in the rest of the paper (and corresponds to virtually all activations used in practice) is that the activation is monotonically non-decreasing. Before continuing, we remark that by assuming a slight strengthening of this assumption, namely that the function is *strictly* monotonically increasing, it is easy to prove a positive guarantee, as evidenced by Thm. 3.2. However, this excludes popular activations such as the ReLU function.

Theorem 3.2 *Assume* $\inf_z \sigma'(z) \geq \gamma > 0$ *for some* $\gamma > 0$*, and the following for some* λ, c_1, c_2*:* $\Sigma := \mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\right]$ *is positive definite with minimal eigenvalue* $\lambda > 0$*,* $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\|\mathbf{x}\|^2\right] \leq c_1$*, and* $\sup_z \sigma'(z) \leq c_2$*. Then starting from any point* $\mathbf{w}_0$*, after doing* t *iterations of gradient descent with learning rate* $\eta < \frac{\lambda\gamma^2}{c_1^2 c_2^4}$*, we have that:*$\|\mathbf{w}_t - \mathbf{v}\|^2 \leq \|\mathbf{w}_0 - \mathbf{v}\|(1-\lambda\gamma^2\eta)^t$.

The proof can be found in Appendix A, and can be easily generalized to apply also to gradient flow and SGD. The above shows that if we assume strict monotonicity of the activation, then under very mild assumptions on the data $\mathbf{w}_t$ will converge exponentially fast to $\mathbf{v}$. We note that this kind of analysis on strictly monotonic activations is not novel in itself (see e.g. Foster et al. (2018); Oymak (2018)), the purpose of the theorem is merely to point out that using a strictly monotonically increasing function makes the analysis dramatically easier. In the rest of the paper, however, we focus on results which only require weak monotonicity.

4. Under Mild Assumptions, the Gradient Points in a Good Direction

Motivated by the results in Sec. 3, we use the following assumptions on the distribution and activation:

Assumption 4.1 *The following holds for some fixed $\alpha, \beta, \gamma > 0$:*

1. *The distribution $\mathcal{D}$ satisfies the following: For any vector $\mathbf{w} \neq \mathbf{v}$, let $\mathcal{D}_{\mathbf{w},\mathbf{v}}$ denote the marginal distribution of $\mathbf{x}$ on the subspace spanned by $\mathbf{w}, \mathbf{v}$ (as a distribution over $\mathbb{R}^2$). Then any such distribution has a density function $p_{\mathbf{w},\mathbf{v}}(\mathbf{x})$ such that $\inf_{\mathbf{x}:\|\mathbf{x}\|\leq\alpha} p_{\mathbf{w},\mathbf{v}}(\mathbf{x}) \geq \beta$.*
2. *$\sigma : \mathbb{R} \mapsto \mathbb{R}$ is monotonically non-decreasing, and satisfies $\inf_{0<z<2\alpha} \sigma'(z) \geq \gamma$.*

The distributional assumption is such that in every 2-dimensional subspace, the marginal distribution is sufficiently "spread" in any direction close to the origin. For example, for a standard Gaussian distribution, this is true for $\alpha, \beta = \Theta(1)$ regardless of the dimension d (as the marginal distribution of a standard Gaussian on the subspace is a standard 2-dimensional Gaussian). Also, for any distribution, it can be made to hold by mixing it with a bit of a Gaussian or uniform distribution if possible. The assumption on the activation function is very mild, and covers most activations used in practice such as ReLU and ReLU-like functions (e.g. leaky-ReLU, Softplus), as well as standard sigmoidal activations (for which the derivative in any bounded interval is lower bounded by a positive constant).

With these assumptions, we prove the following key technical result, which implies that the gradient of the objective has a positive correlation with the direction of the global minimum (at $\mathbf{w} = \mathbf{v}$), if the angle between $\mathbf{w}$ and $\mathbf{v}$ and the norm of $\mathbf{w}$ are not too large:

Theorem 4.2 *Under Assumptions 4.1, for any $\mathbf{w}$ such that $\|\mathbf{w}\| \leq 2$ and $\theta(\mathbf{w}, \mathbf{v}) \leq \pi - \delta$ for some $\delta \in (0, \pi]$, it holds that $\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{v} \rangle \geq \frac{\alpha^4\beta\gamma^2}{8\sqrt{2}} \sin^3\left(\frac{\delta}{4}\right) \|\mathbf{w} - \mathbf{v}\|^2$.*

The theorem implies that for suitable values of $\mathbf{w}$, gradient methods (which move in the negative gradient direction) will decrease the distance from $\mathbf{v}$. When this behavior occurs, it is easy to show that gradient methods succeed in learning the target neuron, like in the previous Thm. 3.2 for the strictly monotonic case. The main challenge is to guarantee that the trajectory of the algorithm will indeed never violate the theorem's conditions, in particular that the angle between $\mathbf{w}$ and $\mathbf{v}$ indeed remains bounded away from π (and in fact, later on we will show that such a guarantee is not always possible).

The formal proof of the theorem can be found in Appendix B, but its intuition can be described as follows: we want to bound below the term

$$\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{v} \rangle = \mathbb{E}_{\mathbf{x}} \left[\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x}) \right) \cdot \sigma'(\mathbf{w}^\top \mathbf{x}) \cdot (\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x}) \right] .$$

Note that:

1. Using the assumption on σ, the term inside the above expectation is nonnegative for every $\mathbf{x}$. This is because $\sigma'(x) \geq 0$, and for any monotonically non-decreasing function f we have $(f(x) - f(y))(x - y) \geq 0$. Thus, viewing the expectation as an integral over a nonnegative function, we can lower bound it by taking the integral over the smaller set $\left\{\mathbf{x} \in \mathbb{R}^d : \ \mathbf{w}^\top\mathbf{x} > 0, \ \mathbf{v}^\top\mathbf{x} > 0\right\}$. Note that on this set, $\sigma(\mathbf{w}^\top\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$ and $\sigma(\mathbf{v}^\top\mathbf{x}) = \mathbf{v}^\top\mathbf{x}$.

2. The resulting integral depends only on dot products of $\mathbf{x}$ with $\mathbf{w}$ and $\mathbf{v}$. Thus, it is enough to consider the marginal distribution on the 2-dimensional plane spanned by $\mathbf{w}$ and $\mathbf{v}$.

3. By the assumption on the distribution, the density function of this marginal distribution is always at least β on any $\mathbf{x}$ such that $\|\mathbf{x}\| \leq \alpha$. This means we can lower bound the integral above by integrating over $\mathbf{w}$ with a uniform distribution on this set and multiplying by β.

In total, the expression above can be lower bounded by a certain 2-dimensional integral (with uniform measure and with no σ terms) on the set $\left\{\mathbf{y} \in \mathbb{R}^2 : \ \hat{\mathbf{w}}^\top\mathbf{y} > 0, \ \hat{\mathbf{v}}^\top\mathbf{y} > 0, \|\mathbf{y}\| \leq \alpha\right\}$ where $\hat{\mathbf{w}}, \hat{\mathbf{v}}$ are the 2-dimensional vectors representing $\mathbf{w}, \mathbf{v}$ on the 2-dimensional plane spanned by them. We lower bound this integral by a term that scales with the angle $\theta(\mathbf{w}, \mathbf{v})$.

Remark 4.3 (Implication on Optimization Landscape) *The proof of the theorem can be shown to imply that for the ReLU activation, under the theorem's conditions, the only stationary point that is not the global minimum* $\mathbf{v}$ *must be at the origin. In particular, the proof implies that any stationary point (with* $\nabla F(\mathbf{w}) = 0$*) must be along the ray* $\{\mathbf{w} = -a \cdot \mathbf{v} : a \geq 0\}$. *For the ReLU activation (which satisfies* $\sigma(z)\sigma'(-a \cdot z) = 0$ *for any* $a \geq 0$ *and* z*), the gradient at such points equals*

$$\nabla F(-a \cdot \mathbf{v}) = \mathbb{E}_{\mathbf{x}}\left[(\sigma(-a\mathbf{v}^\top\mathbf{x}) - \sigma(\mathbf{v}^\top\mathbf{x}))\sigma'(-a\mathbf{v}^\top\mathbf{x})\mathbf{x}\right] \ = \ \mathbb{E}_{\mathbf{x}}\left[(-a\mathbf{v}^\top\mathbf{x})\sigma'(-a\mathbf{v}^\top\mathbf{x})\mathbf{x}\right] \ .$$

In particular, $\langle\nabla F(-a \cdot \mathbf{v}), \mathbf{v}\rangle = -a \cdot \mathbb{E}_{\mathbf{x}}\left[\sigma'(-a\mathbf{v}^\top\mathbf{x})(\mathbf{v}^\top\mathbf{x})^2\right]$ *. This implies that* $\nabla F(-a \cdot \mathbf{v})$ *might be zero only if either* $a = 0$ *(i.e., at the origin), or* $\mathbf{v}^\top\mathbf{x} \geq 0$ *with probability* 1*, which cannot happen according to Assumption 4.1.*

Remark 4.4 (Impossible to generalize to $\mathbf{w} = -c \cdot \mathbf{v}$) *Thm. 4.2 does not cover the case when* $\mathbf{w}$ *is in the opposite direction of* $\mathbf{v}$*, i.e. there is* $c > 0$ *such that* $\mathbf{w} = -c \cdot \mathbf{v}$*. However, it is impossible to generalize the theorem in this direction, even if the distribution is standard Gaussian and for the ReLU activation. The reason is that in this case, using the closed form for the gradient from Brutzkus and Globerson (2017) we get that* $\nabla F(\mathbf{w}) = \frac{1}{2}\mathbf{w}$ *and in particular, gradient descent would converge to the suboptimal stationary point at the origin.*

5. Convergence with Constant Probability Under Mild Assumptions

In this section, we use Thm. 4.2 in order to show that under some assumption on the initialization of $\mathbf{w}$, gradient methods will be able to learn a single neuron with probability at least (close to) $\frac{1}{2}$. Note that the loss surface of $F(\mathbf{w})$ is not convex, and as explained in Remark 4.3, there may be a stationary point at $\mathbf{w} = \mathbf{0}$. This stationary point can cause difficulties, as it is not obvious how to control the angle between $\mathbf{v}$ and $\mathbf{w}$ close to the origin (which is required for Thm. 4.2 to apply). But, if we assume $\|\mathbf{w} - \mathbf{v}\|^2 < 1$ at initialization, then we are bounded away from the origin, and we can ensure that it will remain that way throughout the optimization process. One such initialization, which guarantees this with at least constant probability, is a zero-mean Gaussian initialization with small enough variance:

Lemma 5.1 *Assume $\|\mathbf{v}\| = 1$. If we sample $\mathbf{w} \sim \mathcal{N}\left(0, \tau^2 I\right)$ for $\tau \leq \frac{1}{d\sqrt{2}}$ then w.p $> \frac{1}{2} - \frac{1}{4}\tau d - 1.2^{-d}$ we have that $\|\mathbf{w} - \mathbf{v}\|^2 \leq 1 - 2\tau^2 d$*

In order to bound each gradient step we will need these additional assumptions:

Assumption 5.2 *The following holds for some positive c_1, c_2:*

1. *$\|\mathbf{x}\|^2 \leq c_1$ almost surely over $\mathbf{x} \sim \mathcal{D}$*
2. *$\sigma'(z) \leq c_2$ for all $z \in \mathbb{R}$*

With these assumptions, we show convergence for gradient flow, gradient descent and stochastic gradient descent:

Theorem 5.3 *Under assumptions 4.1 and 5.2 we have:*

1. *(Gradient Flow) Assume that $\|\mathbf{w}(0) - \mathbf{v}\|^2 < 1$. Running gradient flow, then for every time $t > 0$ we have*
$$\|\mathbf{w}(t) - \mathbf{v}\|^2 \leq \|\mathbf{w}(0) - \mathbf{v}\|^2 \exp(-t\lambda)$$
where $\lambda = \frac{\alpha^4 \beta \gamma^2}{210}$.

2. *(Gradient Descent) Assume that $\|\mathbf{w}_0 - \mathbf{v}\|^2 < 1$. Let $\eta \leq \frac{\lambda}{2c}$ for $\lambda = \min\left\{1, \frac{\alpha^4\beta\gamma^2}{210}\right\}$ and $c = c_1^2 c_2^4$. Running gradient descent with step size η, we have that for every $T > 0$, after T iterations:*
$$\|\mathbf{w}_T - \mathbf{v}\|^2 \leq \|\mathbf{w}_0 - \mathbf{v}\|^2 \left(1 - \frac{\eta\lambda}{2}\right)^T$$

3. *(Stochastic Gradient Descent) Let $\epsilon_1, \epsilon_2, \delta > 0$, and assume that $\|\mathbf{w}_0 - \mathbf{v}\|^2 \leq 1 - \epsilon_1$. Let $\eta \leq \frac{\lambda \epsilon_1^2 \epsilon_2^2 c_3^2}{60 c_1^3 c_2^6 \log\left(\frac{2}{\delta}\right)}$ where $\lambda = \frac{\alpha^4\beta\gamma^2}{210}$ and $c_3 = \left(\frac{1}{2}\right)^{\frac{\lambda}{20 c_1 c_2^2}} - \left(\frac{1}{2}\right)^{\frac{\lambda}{18 c_1 c_2^2}}$. Then w.p $1 - \left\lceil \frac{20 c_1 c_2^2 \log\left(\frac{1}{\epsilon_2}\right)}{\lambda} \right\rceil \delta$, after $T \geq \frac{2\log\left(\frac{1}{\epsilon_2}\right)}{\lambda\eta}$ iterations we have that: $\|\mathbf{w}_T - \mathbf{v}\|^2 \leq \epsilon_2$.*

Combined with Lemma 5.1, Thm. 5.3 shows that with proper initialization, gradient flow, gradient descent as well as stochastic gradient descent successfully minimize Eq. (1) with probability (close to) $\frac{1}{2}$, and for the first two algorithms, the distance to $\mathbf{v}$ decays exponentially fast.

The full proof of the theorem can be found in Appendix C, and its intuition for gradient flow and gradient is as described above (namely, that if $\|\mathbf{w} - \mathbf{v}\| < 1$, it will stay that way and $\|\mathbf{w} - \mathbf{v}\|$ will just continue to shrink over time, using Thm. 4.2). The proof for stochastic gradient descent is much more delicate. This is because the update at each iteration is noisy, so we need to ensure we remain in the region where Thm. 4.2 is applicable. Here we give a short proof intuition:

1. Assume we initialized with $\|\mathbf{w}_0 - \mathbf{v}\|^2 \leq 1 - \epsilon$ for some $\epsilon > 0$. In order for the analysis to work we need that $\|\mathbf{w}_t - \mathbf{v}\| < 1$ throughout the algorithm's run. Thus, we show (using a maximal version of Azuma's inequality) that if η is small enough (depending on ϵ), and we take at most $m = O\left(\frac{1}{\eta}\right)$ gradient steps then w.h.p for every $t = 1, \dots, m$: $\|\mathbf{w}_t - \mathbf{v}\|^2 \leq 1 - \frac{\epsilon}{2}$

2. The next step is to show that if $\|\mathbf{w}_t - \mathbf{v}\|^2 < 1$, then $\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 | \mathbf{w}_t\right] \leq (1-\eta\lambda)\|\mathbf{w}_t - \mathbf{v}\|^2$ for an appropriate λ. This is done using Thm. 4.2, as in the gradient descent case, but note that here this only holds in expectation over the sample selected at iteration t.

3. Next, we use Azuma's inequality again on $m = O\left(1/\eta\right)$ iterations for a small enough η, to show that w.h.p $\mathbf{w}_m$ does not move too far away from $\tilde{\mathbf{w}}_m := \mathbb{E}[\mathbf{w}_m]$ where the expectation is taken over $\mathbf{x}_1, \dots, \mathbf{x}_m$. Also, we show that after m iterations $\|\tilde{\mathbf{w}}_m - \mathbf{v}\|^2 \leq \rho\|\mathbf{w}_0 - \mathbf{v}\|^2$ for a constant ρ smaller than 1. This shows that w.h.p., after a single epoch of m iterations, $\|\mathbf{w}_m - \mathbf{v}\|$ shrinks by a constant factor.

4. We then repeat this analysis across t epochs (each consisting of m iterations), and use a union bound. Overall, we get that after sufficiently many iterations, with high probability, the iterates get as close as we want to zero.

We note the optimization analysis for stochastic gradient descent is inspired by the analysis in Shamir (2015) for the different non-convex problem of principal component analysis (PCA), which also attempts to avoid a problematic stationary point. An interesting question for future research is to understand to what extent the polynomial dependencies in the problem parameters can be improved.

Remark 5.4 *Our assumption on the data that $\|\mathbf{x}\|^2 \leq c_1$ is made for simplicity. For the gradient descent case, it is easy to verify that the proof only requires that the fourth moment of the data is bounded by some constant, which ensures that the gradients of the objective function used by the algorithm are bounded. For SGD it is enough to assume that the input distribution is sub-Gaussian. The proof proceeds in the same manner, by using a concentration bound for martingales with sub-Gaussian tails.*

6. High-Probability Convergence

The results in the previous section hold under mild conditions, but unfortunately only guarantee a constant probability of success. In this section, we consider the possibility of proving guarantees which hold with high probability (arbitrarily close to 1). On the one hand, in Subsection 6.1, we provide such a result for the ReLU activation, assuming the input distribution $\mathcal{D}$ is spherically symmetric. On the other hand, in Subsection 6.2, we point out non-trivial obstacles to extending such a result to non-spherically symmetric distributions. Overall, we believe that getting high-probability convergence guarantees for non-spherically symmetric distributions is an interesting avenue for future research.

6.1. Convergence for Spherically Symmetric Distributions

In this subsection, we make the following assumptions:

Assumption 6.1 *Assume that:*

1. $\mathbf{x} \sim \mathcal{D}$ has a spherically symmetric distribution. That is, for any orthogonal matrix A, it holds that $A\mathbf{x} \sim \mathcal{D}$.

2. The activation function $\sigma(\cdot)$ is the standard ReLU function $\sigma(z) = \max\{0, z\}$.

These assumptions are significantly stronger than Assumptions 4.1, but allow us to prove a stronger high-probability convergence result. Note that even with these assumptions the loss surface is still not convex, and may contain a spurious stationary point (see Remark 4.3). For simplicity, we will focus on proving the result for gradient flow. The result can then be extended to gradient descent and stochastic gradient descent, along similar lines as in the proof of Thm. 5.3.

The proof strategy in this case is quite different from that of the constant-probability guarantee, and relies on the following key technical result:

Lemma 6.2 *If $\mathbf{w}(t) \neq 0$, then $\frac{\partial}{\partial t}\theta(\mathbf{w}(t), \mathbf{v}) \leq 0$*

The lemma (which relies on the spherical symmetry of the distribution) implies that if we initialize at any point $\mathbf{w}(0) \notin \text{span}\{\mathbf{v}\}$, then the angle between $\mathbf{w}(0)$ and $\mathbf{v}$ is strictly less than π, and will remain so as long as $\mathbf{w}(t) \neq 0$. As a result, we can apply Thm. 4.2 to prove that $\|\mathbf{w}(t) - \mathbf{v}\|$ decays exponentially fast. The only potential difficulty is that $\mathbf{w}(t)$ may converge to the potential stationary point at the origin (at which the angle is not well-defined), but fortunately this cannot happen due to the following lemma:

Lemma 6.3 *Let $\theta = \theta(\mathbf{w}(t), \mathbf{v})$ and assume that $\mathbf{w}(t) \neq 0$. If $\|\mathbf{w}(t)\|$ is at most $\max\left\{\frac{\sin(\theta)+\cos(\theta)}{2}, \frac{\sin(\theta)(1+\cos(\theta))}{2}\right\}$ then $\frac{\partial}{\partial t}\|\mathbf{w}(t)\|^2 \geq 0$.*

The lemma can be shown to imply that as long as θ remains bounded away from π, then $\|\mathbf{w}(t)\|^2$ cannot decrease below some positive number (as its derivative is positive close enough to zero, and $\|\mathbf{w}(t)\|^2$ is a continuous function of t). The proof idea of both lemmas is based on a technical calculation, where we project the spherically symmetric distribution on the 2-dimensional subspace spanned by $\mathbf{w}$ and $\mathbf{v}$.

Using the lemmas above, we can get the following convergence guarantee:

Theorem 6.4 *Assume we initialize $\mathbf{w}(0)$ such that $0 < \|\mathbf{w}(0)\| \leq 2$, $\theta(\mathbf{w}(0), \mathbf{v}) \leq \pi - \epsilon$ for some $\epsilon > 0$ and that Assumption 4.1(1) holds. Then running gradient flow, for all $t \geq 0$*

$$\|\mathbf{w}(t) - \mathbf{v}\|^2 \leq \|\mathbf{w}(0) - \mathbf{v}\| \exp(-\lambda t)$$

where $\lambda = \frac{\alpha^4\beta}{8\sqrt{2}} \sin^3\left(\frac{\epsilon}{8}\right)$.

We now note that the assumption of the theorem holds with exponentially high probability under standard initialization schemes. For example, if we use a Gaussian initialization $\mathbf{w}(0) \sim \mathcal{N}(0, \frac{1}{d}I)$, then by standard concentration of measure arguments, it holds w.p $> 1 - e^{-\Omega(d)}$ that $\theta(\mathbf{w}(0), \mathbf{v})$ is at most (say) $\frac{3\pi}{4}$, and w.p $> 1 - e^{-\Omega(d)}$ that $\|\mathbf{w}(0)\| \leq 2$. As a result, by Thm. 6.4, w.p $> 1 - e^{-\Omega(d)}$ over the initialization we have $\|\mathbf{w}(t) - \mathbf{v}\|^2 \leq \|\mathbf{w}(0) - \mathbf{v}\|^2 e^{-\Omega(t)}$ for all t. The full proof of the theorem can be found in Appendix D.

Remark 6.5 *If we further assume that the distribution is a standard Gaussian, then it is possible to prove Lemma 6.2 and Lemma 6.3 in a much easier fashion. The reason is that specifically for a standard Gaussian distribution there is a closed-form expression (without the expectation) for the loss and the gradient, see Brutzkus and Globerson (2017), Safran and Shamir (2017). We provide the relevant versions of the lemmas, as well as their proofs, in Subsection D.1.*

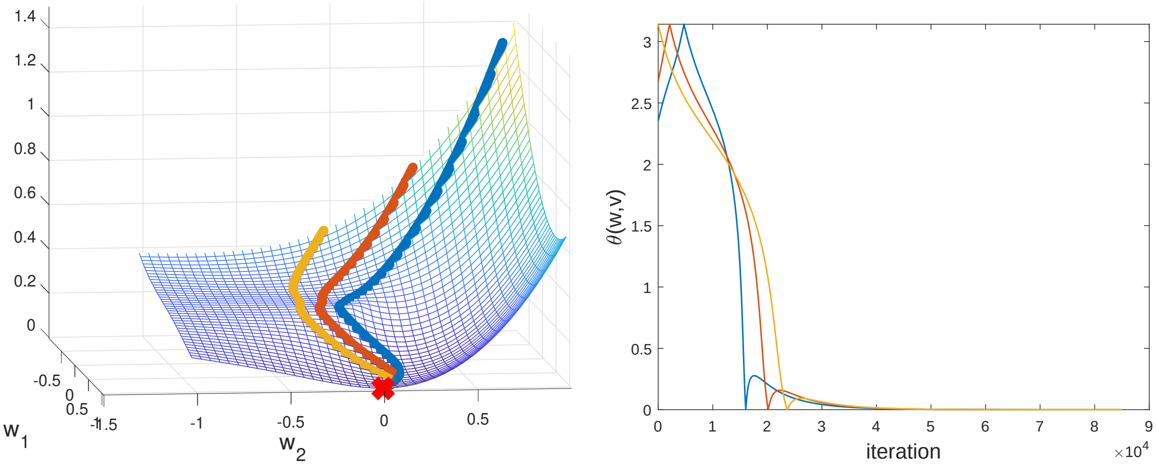

Figure 1: Gradient descent for 2-dimensional data (best viewed in color). The left figure represents the trajectory of gradient descent over the loss surface. The red ”x” marker represents the global minimum at $\mathbf{w} = \mathbf{v} = (1, 0)$. The right figure shows the angle between $\mathbf{w}$ and $\mathbf{v}$ as a function of the number of iterations, where the angle ranges from 0 to π. The plot colors in the right figure correspond to the trajectory colors in the left figure.

6.2. Non-monotonic Angle Behavior

The results in the previous subsection crucially relied on the fact that at almost any point $\mathbf{w}$, the angle $\theta(\mathbf{w}, \mathbf{v})$ decreases. This type of analysis was also utilized in works on related settings (e.g., Brutzkus and Globerson (2017)).

Based on this, it might be tempting to conjecture that this monotonically decreasing angle property (and as a result, high-probability guarantees) can be shown to hold more generally, not just for symmetrically spherical distributions. Perhaps surprisingly, we show empirically that this may not be the case, already when we discuss the simple setting of unit variance Gaussian with a *non-zero* mean. We emphasize that this does not necessarily mean that gradient methods will not succeed, only that an analysis based on showing monotonic behavior of the relevant geometric quantity will not work in general.

In particular, in Figure 1 we report the result of running gradient descent (with constant step size $\eta = 10^{-3}$) on our objective function F in $\mathbb{R}^2$, where the input distribution $\mathcal{D}$ is a unit-variance Gaussian with mean at $(0, 1)$, and our target vector is $\mathbf{v} = (1, 0)$. We initialize at three different locations: $w_1 = (-1\ 1)$, $w_2 = (-1, 0.5)$, $w_3 = (-1, 0)$. Although the algorithm eventually reaches the global minimum $\mathbf{w} = \mathbf{v}$, the angle between them is clearly non-monotonic, and actually is initially increasing rather than decreasing. Even worse, the angle appears to attain every value in $(0, \pi]$, so it appears that any analysis using angle-based “safe regions” is bound to fail.

Overall, we conclude that proving a high-probability convergence guarantee for gradient methods appears to be an interesting open problem, already in the case of unit-variance, non-zero-mean Gaussian input distributions. We leave tackling this problem to future work.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, 2019.

[2] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[3] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[4] Yuan Cao and Quanquan Gu. A generalization theory of gradient descent for learning overparameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019.

[5] Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[6] Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

[7] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, pages 8745–8756, 2018.

[8] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[9] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[10] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[12] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of nonconvexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[13] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[14] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[15] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT*. Citeseer, 2009.

[16] Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A Salman Avestimehr. Fitting relus via sgd and quantized sgd. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2469–2473. IEEE, 2019.

[17] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

[18] Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. *arXiv preprint arXiv:1809.03019*, 2018.

[19] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv preprint arXiv:1812.10004*, 2018.

[20] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[21] Ohad Shamir. A stochastic pca and svd algorithm with an exponential convergence rate. In *International Conference on Machine Learning*, pages 144–152, 2015.

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

[23] Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

[24] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[25] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[26] Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*, 2019.

[27] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3404–3413. JMLR. org, 2017.

[28] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Appendix A. Proofs from Sec. 3

Proof For each distribution $\mathcal{D}^i_{\mathbf{w}}$, let $p_i = \mathcal{P}(w_i > 0)$. We define the following dataset:

$$S = \{\mathbf{x}_i = b_i\mathbf{e}_i : i = 1 \dots, d\}$$

where $\mathbf{e}_i$ is the standard i-th unit vector, and $b_i = 1$ if $p_i < \frac{1}{2}$ and -1 otherwise. Take $\mathcal{D}$ to be the uniform distribution on S.

Informally, the proof idea is the following: With overwhelming probability, we will initialize at a point $\mathbf{w}$ such that for at least $\Omega(d)$ coordinates i, it holds that $\sigma'(\mathbf{w}^\top\mathbf{x}_i) = 0$, and as a result, $\nabla F(\mathbf{w})$ is zero on those coordinates. Based on this, we show that these coordinates will not change from their initialized values. However, a point $\mathbf{w}$ with $\Omega(d)$ coordinates with this property is suboptimal by a fixed factor, so the algorithm does not converge to an optimal solution.

More formally, using Eq. (2) and the fact that σ is the ReLU function, we get

$$\nabla F(\mathbf{w}) = \frac{1}{d}\sum_{i=1}^{d}\left(\sigma(\mathbf{w}^\top\mathbf{x}_i) - \sigma(\mathbf{v}^\top\mathbf{x}_i)\right)\cdot \mathbb{1}\left(\mathbf{w}^\top\mathbf{x}_i > 0\right)\mathbf{x}_i\,.$$

In particular, for every index i for which $\mathbb{1}\left(\mathbf{w}^\top\mathbf{x}_i > 0\right) = 0$ we have that $\left(\nabla F(\mathbf{w})\right)_i = 0$. Next, we define $\mathbf{v}$ with $\mathbf{v}_i = b_i\frac{1}{\sqrt{d}}$ (note that $\|\mathbf{v}\| = 1$). For every $d/4$ indices $i_1, \dots, i_{d/4}$ for which $\mathbb{1}\left(\mathbf{w}^\top\mathbf{x}_i \geq 0\right) = 0$ we have that:

$$\begin{aligned} F(\mathbf{w}) = \frac{1}{2d}\sum_{i=1}^{d}\left(\sigma(\mathbf{w}^\top\mathbf{x}_i) - \sigma(\mathbf{v}^\top\mathbf{x}_i)\right)^2 &\geq \frac{1}{2d}\sum_{i\in\{i_1,\dots,i_{d/4}\}}\left(\sigma(\mathbf{w}^\top\mathbf{x}_i) - \sigma(\mathbf{v}^\top\mathbf{x}_i)\right)^2 \\ =\frac{1}{2d}\sum_{i\in\{i_1,\dots,i_{d/4}\}}\sigma(\mathbf{v}^\top\mathbf{x}_i)^2 = \frac{1}{2d}\sum_{i\in\{i_1,\dots,i_{d/4}\}}\sigma\left(b_i^2\frac{1}{\sqrt{d}}\right)^2 &= \frac{1}{8d} \end{aligned} \tag{4}$$

Denote the random variable $Z_i = \mathbb{1}\left(\mathbf{w}_0^\top\mathbf{x}_i > 0\right)$ and $Z = \sum_{i=1}^d Z_i$ (for gradient flow we denote $Z_i = \mathbb{1}\left(\mathbf{w}(0)^\top\mathbf{x}_i \geq 0\right)$). It is easily verified that $\mathbb{E}[Z_i] = \Pr(\mathbf{w}_0^\top\mathbf{x}_i > 0) = \Pr(w_{0,i}b_i > 0) \leq \frac{1}{2}$. We have that $Z_1, \dots, Z_d$ are independent, $\max_i |Z_i| \leq 1$, and $\mathbb{E}[Z] = \sum_{i=1}^d \mathbb{E}[Z_i] \leq \frac{d}{2}$. Using Hoeffding's inequality, we get that w.p $\geq 1 - \exp\left(-\frac{d}{4}\right)$ it holds that $Z \leq \frac{3}{4}d$, which means that there are at least $\frac{d}{4}$ indices such that $Z_i = 0$. We condition on this event and let these indices be $i_1, \dots, i_{d/4}$. We will now show that for every index $i \in \{i_1, \dots, i_{d/4}\}$, using gradient methods will not change the i-th coordinate of $\mathbf{w}_t$ ($\mathbf{w}(t)$ for gradient flow) from its initial value. Let i be such a coordinate.

For gradient descent, we will show by induction that for every iteration t we have that $\mathbb{1}\left(\mathbf{w}_t^\top\mathbf{x}_i > 0\right) = 0$. The base case is true, because we conditioned on this event. Assume for $t-1$, then $\left(\nabla F(\mathbf{w}_{t-1})\right)_i = 0$, which means that $(\mathbf{w}_t)_i = (\mathbf{w}_{t-1})_i - \eta(\nabla F(\mathbf{w}_{t-1}))_i = (\mathbf{w}_{t-1})_i$, and in particular $\mathbb{1}\left(\mathbf{w}_t^\top\mathbf{x}_i > 0\right) = \mathbb{1}\left(\mathbf{w}_{t-1}^\top\mathbf{x}_i > 0\right) = 0$. This proves that for every iteration t, the i-th coordinate of $\nabla F(\mathbf{w}_t)$ is zero, which mean that $(\mathbf{w}_t)_i = (\mathbf{w}_0)_i$.

For stochastic gradient descent, at each iteration t we sample $\mathbf{x}_t \sim \mathcal{D}$, and define the stochastic gradient g_t as in Eq. (3). If $\mathbf{x}_t \neq \mathbf{x}_i$ then $(\mathbf{x}_t)_i = 0$ hence $(g_t)_i = 0$, otherwise, if $\mathbf{x}_t = \mathbf{x}_i$ then by $(g_t)_i = (\nabla F(\mathbf{w}_t))_i$ and by the same induction argument as in gradient descent we have that $(g_t)_i = 0$. In both cases the i-th coordinate of the stochastic gradient is zero, hence $(\mathbf{w}_t)_i = (\mathbf{w}_0)_i$.

For gradient flow, assume by contradiction that for some $t > 0$ that $\mathbb{1}\left(\mathbf{w}(t)^\top \mathbf{x}_i > 0\right) \neq 0$ and let t_1 be the first time that this happen. Then for all $0 < t < t_1$ we have that $\mathbb{1}\left(\mathbf{w}(t)^\top \mathbf{x}_i > 0\right) = 0$, and in particular $\left(\nabla F(\mathbf{w}(t))\right)_i = 0$. Hence for all $0 < t < t_1$ running gradient flow we get $\left(\dot{\mathbf{w}}(t)\right)_i = \left(\nabla F(\mathbf{w}(t))\right)_i = 0$, and in particular $\mathbb{1}\left(\mathbf{w}(t)^\top \mathbf{x}_i > 0\right) = \mathbb{1}\left(\mathbf{w}(0)^\top \mathbf{x}_i > 0\right) = 0$, a contradiction to the fact that $\mathbf{w}(t)$ is continuous. Thus for all $t > 0$ we showed that $\mathbb{1}\left(\mathbf{w}(t)^\top \mathbf{x}_i > 0\right) = 0$, hence $\left(\nabla F(\mathbf{w}(t))\right)_i = 0$ which shows that $(\mathbf{w}(t))_i = (\mathbf{w}(0))_i$.

By the conditioned event, Eq. (4) applies at initialization. Since in all the gradient methods above the i-th coordinate of $\mathbf{w}$ did not change from its initial value for $i \in \{i_1, \dots, i_{d/4}\}$, we can apply Eq. (4) to get that for every iteration $t > 0$ for gradient descent or SGD we have that $F(\mathbf{w}_t) \geq \frac{1}{8d}$ (and for gradient flow, for every time $t > 0$, we have $F(\mathbf{w}(t)) \geq \frac{1}{8d}$).

We end by noting that although the distribution defined here is discrete over a finite dataset, the same argument can also be made for a non-discrete distribution, by considering a mixture of smooth distributions concentrated around the support points of the discrete distribution above. ■

Proof [Thm. 3.2] We have that:

$$\begin{aligned}
\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{v} \rangle &= \mathbb{E}_{\mathbf{x}}\left[(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x}))\sigma'(\mathbf{w}^\top \mathbf{x})(\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})\right] \\
&\overset{(*)}{\geq} \mathbb{E}_{\mathbf{x}}\left[\gamma \cdot (\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x}))(\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})\right] \\
&\overset{(**)}{\geq} \mathbb{E}_{\mathbf{x}}\left[\gamma^2 (\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})^2\right] \;=\; \gamma^2 (\mathbf{w} - \mathbf{v})^\top \Sigma (\mathbf{w} - \mathbf{v}) \geq \gamma^2 \lambda \|\mathbf{w} - \mathbf{v}\|^2
\end{aligned}$$

where $(*)$ is by monotonicity of σ (hence $(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x}))(\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x}) \geq 0$ always), and $(**)$ is by the assumption that $\sigma'(z) \geq \gamma$. Next, we bound the gradient $\nabla F(\mathbf{w})$:

$$\begin{aligned}
\|\nabla F(\mathbf{w}_t)\|^2 &= \mathbb{E}_{\mathbf{x}}\left[\left(\sigma(\mathbf{w}_t^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)^2 \cdot \sigma'(\mathbf{w}^\top \mathbf{x})^2 \mathbf{x}^\top \mathbf{x}\right] \\
&\leq c_2^4 \mathbb{E}_{\mathbf{x}}\left[\left(\mathbf{w}_t^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x}\right)^2 \cdot \mathbf{x}^\top \mathbf{x}\right] \\
&\leq c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2 \mathbb{E}_{\mathbf{x}}\left[\|\mathbf{x}\|^2 \cdot \mathbf{x}^\top \mathbf{x}\right] \leq c_1^2 c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2.
\end{aligned}$$

At iteration $t + 1$ we have that:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 &= \|\mathbf{w}_t - \eta \nabla F(\mathbf{w}_t) - \mathbf{v}\|^2 \\
&= \|\mathbf{w}_t - \mathbf{v}\|^2 - 2\eta \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v} \rangle + \eta^2 \|\nabla F(\mathbf{w}_t)\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{v}\|^2 - 2\gamma^2 \lambda \eta \|\mathbf{w}_t - \mathbf{v}\|^2 + \eta^2 c_1^2 c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{v}\|^2 \left(1 - \gamma^2 \lambda \eta\right).
\end{aligned}$$

Using induction over the above proves the lemma.

■

Appendix B. Proofs from Sec. 4

We will first need the following lemma:

Lemma B.1 *Fix some $\alpha \geq 0$, and let $\mathbf{a}, \mathbf{b}$ be two vectors in $\mathbb{R}^2$ such that $\theta(\mathbf{a}, \mathbf{b}) \leq \pi - \delta$ for some $\delta \in (0, \pi]$. Then*

$$\inf_{\mathbf{u}:\|\mathbf{u}\|=1} \int \mathbb{1}_{\mathbf{a}^\top \mathbf{y}>0} \mathbb{1}_{\mathbf{b}^\top \mathbf{y}>0} \mathbb{1}_{\|\mathbf{y}\|\leq\alpha} (\mathbf{u}^\top \mathbf{y})^2 d\mathbf{y} \;\geq\; \frac{\alpha^4}{8\sqrt{2}} \sin^3\left(\frac{\delta}{4}\right) .$$

Proof It is enough to lower bound

$$\inf_{\mathbf{u}} \inf_{\mathbf{b}:\theta(\mathbf{a},\mathbf{b})\leq\pi-\delta} \int \mathbb{1}_{\mathbf{a}^\top \mathbf{y}>0, \mathbf{b}^\top \mathbf{y}>0, \|\mathbf{y}\|\leq\alpha} (\bar{\mathbf{u}}^\top \mathbf{y})^2 d\mathbf{y} .$$

The inner infimum is attained at some $\mathbf{b}$ such that $\theta(\mathbf{a}, \mathbf{b}) = \pi - \delta$. This is because $\bar{\mathbf{u}}^\top \mathbf{y}$ does not depend on $\mathbf{a}$ and $\mathbf{b}$, and the volume for which the indicator function inside the integral is non-zero is smallest when the angle $\theta(\mathbf{a}, \mathbf{b})$ is largest. Setting this and switching the order of the infima, we get

$$\inf_{\mathbf{b}:\theta(\mathbf{a},\mathbf{b})=-\pi+\delta} \inf_{\mathbf{u}} \int \mathbb{1}_{\mathbf{a}^\top \mathbf{y}>0} \mathbb{1}_{\mathbf{b}^\top \mathbf{y}>0} \mathbb{1}_{\|\mathbf{y}\|\leq\alpha} (\bar{\mathbf{u}}^\top \mathbf{y})^2 d\mathbf{y} .$$

When $\theta(\mathbf{a}, \mathbf{b}) = -\pi + \delta$, we note that the set $\{\mathbf{y} \in \mathbb{R}^2 : \mathbf{a}^\top \mathbf{y} > 0, \mathbf{b}^\top \mathbf{y} > 0, \|\mathbf{y}\| \leq \alpha\}$ is simply a "pie slice" of radial width δ out of a ball of radius α. Since the expression is invariant to rotating the coordinates, we will consider without loss of generality the set $P = \{\mathbf{y} : \theta(\mathbf{y}, \mathbf{e}_1) \leq \delta/2, \|\mathbf{y}\| \leq \alpha\}$, and the expression above reduces to

$$\begin{aligned}
\inf_{\mathbf{u}} \int_{\mathbf{y}\in P} (\bar{\mathbf{u}}^\top \mathbf{y})^2 d\mathbf{y} &= \inf_{\mathbf{u}:\|\mathbf{u}\|=1} \int_{\mathbf{y}\in P} \left((u_1 y_1)^2 + (u_2 y_2)^2 + 2u_1 u_2 y_1 y_2\right) d\mathbf{y} \\
&\overset{(*)}{=} \inf_{\mathbf{u}:\|\mathbf{u}\|=1} \int_{\mathbf{y}\in P} \left((u_1 y_1)^2 + (u_2 y_2)^2\right) d\mathbf{y} \\
&= \inf_{u_1, u_2 : u_1^2 + u2^2 = 1} u_1^2 \int_{\mathbf{y}\in P} y_1^2 d\mathbf{y} + u_2^2 \int_{\mathbf{y}\in P} y_2^2 d\mathbf{y} \\
&= \min\left\{\int_{\mathbf{y}\in P} y_1^2 d\mathbf{y} \,,\, \int_{\mathbf{y}\in P} y_2^2 d\mathbf{y}\right\} \\
&\geq \int_{\mathbf{y}\in P} \min\{y_1^2, y_2^2\} d\mathbf{y} \,, \qquad (5)
\end{aligned}$$

where $(*)$ is from the fact that P is symmetric around the x-axis (namely, $(y_1, y_2) \in P$ if and only if $(y_1, -y_2) \in P$).

We now note that the set P contains the two (disjoint and equally-sized) rectangular sets

$$P_1' := \left[\frac{\alpha}{2}\cos\left(\frac{\delta}{4}\right), \alpha\cos\left(\frac{\delta}{4}\right)\right] \times \left[\frac{\alpha}{2}\sin\left(\frac{\delta}{4}\right), \alpha\sin\left(\frac{\delta}{4}\right)\right]$$

and

$$P_2' := \left[\frac{\alpha}{2}\cos\left(\frac{\delta}{4}\right), \alpha\cos\left(\frac{\delta}{4}\right)\right] \times \left[-\alpha\sin\left(\frac{\delta}{4}\right), -\frac{\alpha}{2}\sin\left(\frac{\delta}{4}\right)\right]$$

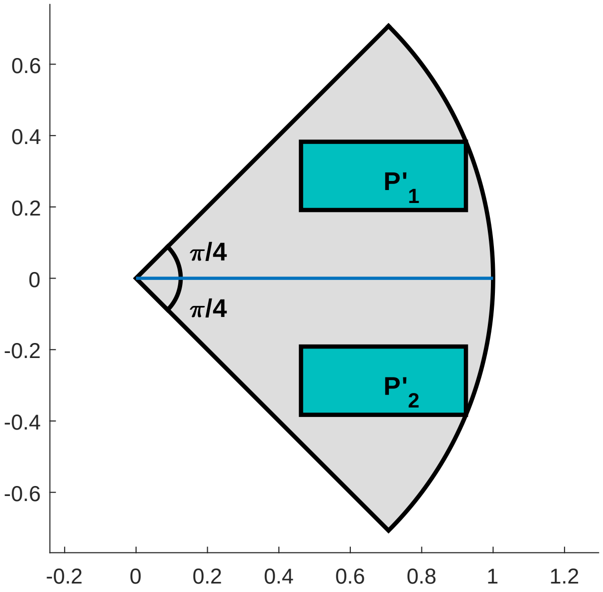

Figure 2: An illustration of the sets P, P_1', P_2' for the case of $\alpha = 1,\ \delta = \frac{\pi}{2}$. The set P, colored in gray, is a ”pie slice” and the rectangles P_1', P_2' are contained in P.

(see Figure 2 for an illustration). Therefore, we can lower bound Eq. (5) by

$$\begin{aligned}
\int_{\mathbf{y}\in P_1'\cup P_2'} \min\{y_1^2, y_2^2\} d\mathbf{y} &= \left(\min_{\mathbf{y}\in P_1'\cup P_2'} \min\{y_1^2, y_2^2\}\right) \int_{\mathbf{y}\in P_1'\cup P_2'} 1 d\mathbf{y} \\
&= \frac{\alpha^2}{4} \min\left\{\cos^2\left(\frac{\delta}{4}\right), \sin^2\left(\frac{\delta}{4}\right)\right\} \cdot \int_{\mathbf{y}\in P_1'\cup P_2'} 1 d\mathbf{y} \\
&= \frac{\alpha^2}{4} \sin^2\left(\frac{\delta}{4}\right) \cdot \int_{\mathbf{y}\in P_1'\cup P_2'} 1 d\mathbf{y}\ ,
\end{aligned}$$

where we used the fact that $\frac{\delta}{4} \in \left[0, \frac{\pi}{4}\right]$ and therefore $\cos^2(\delta/4) \geq \sin^2(\delta/4)$. The integral is simply the volume of $P_1' \cup P_2'$, and since P_1' and P_2' are disjoint and equally sized rectanges, this equals twice the volume of P_1', namely $2 \cdot \frac{\alpha}{2} \cos\left(\frac{\delta}{4}\right) \cdot \frac{\alpha}{2} \sin\left(\frac{\delta}{4}\right)$. Plugging into the above, we get

$$\frac{\alpha^2}{4} \sin^2\left(\frac{\delta}{4}\right) \cdot \frac{\alpha^2}{2} \cos\left(\frac{\delta}{4}\right) \sin\left(\frac{\delta}{4}\right) = \frac{\alpha^4}{8} \sin^3\left(\frac{\delta}{4}\right) \cos\left(\frac{\delta}{4}\right) \geq \frac{\alpha^4}{8\sqrt{2}} \sin^3\left(\frac{\delta}{4}\right)\ ,$$

where again we used the fact that $\delta/4 \in [0, \pi/4]$.

■

We now turn to prove the theorem:

Proof [Thm. 4.2]

We have:

$$\langle \nabla F(\mathbf{w}), \mathbf{w}-\mathbf{v}\rangle = \mathbb{E}_{\mathbf{x}}\left[\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)\cdot \sigma'(\mathbf{w}^\top \mathbf{x})\cdot (\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})\right] . \tag{6}$$

We note that since σ is monotonically non-decreasing, then for any $\mathbf{x}$, $\sigma'(\mathbf{w}^\top \mathbf{x}) \geq 0$ and $(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top))(\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x}) \geq 0$. As a result, we can lower bound Eq. (6) by

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\mathbf{w}^\top \mathbf{x}>0}\mathbb{1}_{\mathbf{v}^\top \mathbf{x}>0}\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)\cdot \sigma'(\mathbf{w}^\top \mathbf{x})\cdot (\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})\right] \\
&\geq \mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\|\mathbf{x}\|\leq\alpha}\mathbb{1}_{\mathbf{w}^\top \mathbf{x}>0}\mathbb{1}_{\mathbf{v}^\top \mathbf{x}>0}\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)\cdot \gamma\cdot (\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})\right] \\
&= \gamma\cdot\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\|\mathbf{x}\|\leq\alpha}\mathbb{1}_{\mathbf{w}^\top \mathbf{x}>0}\mathbb{1}_{\mathbf{v}^\top \mathbf{x}>0}\left(\sigma(\mathbf{w}^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)(\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})\right] ,
\end{aligned}$$

where we used that $\|\mathbf{w}\| \leq 2$, hence for $\|\mathbf{x}\| \leq \alpha$ we have $\langle \mathbf{x}, \mathbf{w}\rangle \leq 2\alpha$ which by our assumption means that $\sigma'(\langle \mathbf{w}, \mathbf{x}\rangle) > \gamma$. By the assumption that $\sigma'(z) \geq \gamma$ for any $0 < z < 2\alpha$, it follows that $(\sigma(z') - \sigma(z))\cdot(z'-z) \geq \gamma(z'-z)^2$ for any $0 < z, z' < 2\alpha$ As a result, the displayed equation above is at least

$$\begin{aligned}
&\gamma^2\cdot\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\|\mathbf{x}\|\leq\alpha}\mathbb{1}_{\mathbf{w}^\top>0}\mathbb{1}_{\mathbf{v}^\top \mathbf{x}>0}(\mathbf{w}^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x})^2\right] \\
&= \gamma^2\|\mathbf{w}-\mathbf{v}\|^2\cdot\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\|\mathbf{x}\|\leq\alpha}\mathbb{1}_{\mathbf{w}^\top \mathbf{x}>0}\mathbb{1}_{\mathbf{v}^\top \mathbf{x}>0}((\overline{\mathbf{w}-\mathbf{v}})^\top \mathbf{x})^2\right] \\
&\geq \gamma^2\|\mathbf{w}-\mathbf{v}\|^2\cdot\inf_{\mathbf{u}\in\text{span}\{\mathbf{w},\mathbf{v}\},\|\mathbf{u}\|=1}\mathbb{E}_{\mathbf{x}}\left[\mathbb{1}_{\|\mathbf{x}\|\leq\alpha}\mathbb{1}_{\mathbf{w}^\top \mathbf{x}>0}\mathbb{1}_{\mathbf{v}^\top \mathbf{x}>0}(\mathbf{u}^\top \mathbf{x})^2\right]
\end{aligned}$$

Since the expression inside the expectation above depends just on inner products of $\mathbf{x}$ with $\mathbf{w}, \mathbf{v}$, we can consider the marginal distribution $\mathcal{D}_{\mathbf{w},\mathbf{v}}$ of $\mathbf{x}$ on the 2-dimensional subspace spanned by $\mathbf{w}, \mathbf{v}$ (with density function $p_{\mathbf{w},\mathbf{v}}$), and letting $\hat{\mathbf{w}}, \hat{\mathbf{v}}$ denote the projections of $\mathbf{w}, \mathbf{v}$ on that subspace, write the above as

$$\begin{aligned}
&\gamma^2\|\mathbf{w}-\mathbf{v}\|^2\cdot\inf_{\mathbf{u}\in\mathbb{R}^2,\|\mathbf{u}\|=1}\mathbb{E}_{\mathbf{y}\sim\mathcal{D}_{\mathbf{w},\mathbf{v}}}\left[\mathbb{1}_{\hat{\mathbf{w}}^\top \mathbf{y}>0}\mathbb{1}_{\hat{\mathbf{v}}^\top \mathbf{y}>0}\mathbb{1}_{\|\mathbf{y}\|\leq\alpha}(\mathbf{u}^\top \mathbf{y})^2\right] \\
&= \gamma^2\|\mathbf{w}-\mathbf{v}\|^2\cdot\inf_{\mathbf{u}\in\mathbb{R}^2,\|\mathbf{u}\|=1}\int \mathbb{1}_{\hat{\mathbf{w}}^\top \mathbf{y}>0}\mathbb{1}_{\hat{\mathbf{v}}^\top \mathbf{y}>0}\mathbb{1}_{\|\mathbf{y}\|\leq\alpha}(\mathbf{u}^\top \mathbf{y})^2 p_{\mathbf{w},\mathbf{v}}(\mathbf{y})d\mathbf{y} \\
&\geq \beta\gamma^2\|\mathbf{w}-\mathbf{v}\|^2\cdot\inf_{\mathbf{u}\in\mathbb{R}^2,\|\mathbf{u}\|=1}\int \mathbb{1}_{\hat{\mathbf{w}}^\top \mathbf{y}>0}\mathbb{1}_{\hat{\mathbf{v}}^\top \mathbf{y}>0}\mathbb{1}_{\|\mathbf{y}\|\leq\alpha}(\mathbf{u}^\top \mathbf{y})^2 d\mathbf{y} ,
\end{aligned}$$

where the last step is by our assumptions (note that if $\mathbf{w} = \mathbf{v}$, the theorem statement is trivially true by Eq. (6) which implies that the inner product is non-negative). The theorem now follows from Lemma B.1. ■

Appendix C. Proofs from Sec. 5

Proof [Lemma 5.1] Fix some $\epsilon > 0$ to be determined later. We have that:

$$\begin{aligned}
\mathcal{P}\left(\|\mathbf{w}-\mathbf{v}\|^2 \leq 1-\epsilon\right) &= \mathcal{P}\left(\|\mathbf{w}\|^2 - 2\langle \mathbf{w}, \mathbf{v}\rangle \leq -\epsilon\right) \\
&= \mathcal{P}\left(\langle \mathbf{w}, \mathbf{v}\rangle \geq \frac{\|\mathbf{w}\|^2+\epsilon}{2}\right) .
\end{aligned}$$

Since the distribution of $\mathbf{w}$ is spherically symmetric, we can assume w.l.o.g that $\mathbf{v} = (1, 0)$, so that $\langle \mathbf{w}, \mathbf{v}\rangle = w_1$. Thus, the above probability can be written as:

$$\begin{aligned} \mathcal{P}\left(\langle \mathbf{w}, \mathbf{v}\rangle \geq \frac{\|\mathbf{w}\|^2 + \epsilon}{2}\right) &= \mathcal{P}\left(w_1 \geq \frac{\|\mathbf{w}\|^2 + \epsilon}{2}\right) \\ &\geq \mathcal{P}\left(w_1 \geq 2\mathbb{E}\left[\|\mathbf{w}\|^2\right]\right) - \mathcal{P}\left(\frac{\|\mathbf{w}\|^2 + \epsilon}{2} \geq 2\mathbb{E}\left[\|\mathbf{w}\|^2\right]\right) \end{aligned} \tag{7}$$

where we used the fact that for every two random variable A, B and constant c we have that $\mathcal{P}(A \geq B) \geq \mathcal{P}(A \geq c) - \mathcal{P}(B \geq c)$. For the first term of Eq. (7), we know that $\mathbb{E}\left[\|\mathbf{w}\|^2\right] = \tau^2 d$, hence:

$$\mathcal{P}\left(w_1 \geq 2\mathbb{E}\left[\|\mathbf{w}\|^2\right]\right) = \mathcal{P}\left(w_1 \geq 2\tau^2 d\right) = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\sqrt{2}\tau d\right)$$

where erf is the error function. For any $0 < z < 1$ it can be easily verified that $\text{erf}(z) \geq \frac{z}{3}$. Combining this and using the assumption that $\tau \leq \frac{1}{d\sqrt{2}}$ we can bound :

$$\mathcal{P}\left(w_1 \geq 2\mathbb{E}\left[\|\mathbf{w}\|^2\right]\right) \geq \frac{1}{2} - \frac{1}{3\sqrt{2}}\tau d \geq \frac{1}{2} - \frac{1}{4}\tau d$$

For the second term of Eq. (7) take $\epsilon = 2\tau^2 d$ to get:

$$\begin{aligned} \mathcal{P}\left(\frac{\|\mathbf{w}\|^2 + \epsilon}{2} \geq 2\mathbb{E}\left[\|\mathbf{w}\|^2\right]\right) &= \mathcal{P}\left(\|\mathbf{w}\|^2 \geq 4\tau^2 d - \epsilon\right) \\ &\leq \mathcal{P}\left(\|\mathbf{w}\|^2 \geq 2\tau^2 d\right) \leq \left(2e^{-1}\right)^{d/2} \leq 1.2^{-d} \end{aligned}$$

where in the second inequality we used a standard tail bound on Chi-squared distributions. Combining the above with Eq. (7) we get that:

$$\mathcal{P}\left(\|\mathbf{w} - \mathbf{v}\|^2 \leq 1 - 2\tau^2 d\right) \geq \frac{1}{2} - \frac{1}{4}\tau d - 1.2^{-d}.$$

■

C.1. Gradient Flow

Proof [Thm. 5.3(1)] First we show that at every time t_0 for which $\|\mathbf{w}(t_0) - \mathbf{v}\| < 1$ the conditions of Thm. 4.2 hold. We have that $\|\mathbf{w}(t_0)\| \leq \|\mathbf{w}(t_0) - \mathbf{v}\| + \|\mathbf{v}\| < 2$, hence $\|\mathbf{w}(t_0)\| < 2$. Next $\|\mathbf{w}(t_0) - \mathbf{v}\|^2 < 1$ and $\|\mathbf{v}\|^2 = 1$ hence $\langle \mathbf{w}(t_0), \mathbf{v}\rangle \geq \frac{1}{2}\|\mathbf{w}(t_0)\|^2 > 0$ which means that $\theta(\mathbf{w}(t_0), \mathbf{v}) < \frac{\pi}{2}$. This shows that we can use Thm. 4.2 at time $t = t_0$ to get that:

$$\frac{\partial}{\partial t}\|\mathbf{w}(t) - \mathbf{v}\|^2 = 2\langle \mathbf{w}(t) - \mathbf{v}, \frac{\partial}{\partial t}\mathbf{w}(t)\rangle = -2\langle \mathbf{w}(t) - \mathbf{v}, \nabla F(\mathbf{w}(t))\rangle \leq 0. \tag{8}$$

By the assumptions of the theorem, the above holds for time $t_0 = 0$. Assume on the way of contradiction that for some time $t > 0$ we have that $\|\mathbf{w}(t) - \mathbf{v}\| \geq 1$, and let t_1 be the first time that this happens. Then for every $t_0 < t < t_1$ we have that $\|\mathbf{w}(t) - \mathbf{v}\| < 1$. But because $\|\mathbf{w}(t_1) - \mathbf{v}\| \geq 1$ we have that for some time $t_0 < t < t_1$: $\frac{\partial}{\partial t}\|\mathbf{w}(t) - \mathbf{v}\| > 0$, a contradiction to Eq. (8). Hence for every $t \geq 0$ we have that $\|\mathbf{w}(t) - \mathbf{v}\| < 1$ and the conditions of Thm. 4.2 hold.

Using Thm. 4.2 again we get that for every $t > 0$:

$$\langle \nabla F(\mathbf{w}(t)), \mathbf{w}(t) - \mathbf{v} \rangle \geq \frac{\alpha^4 \beta \gamma^2}{8\sqrt{2}} \sin\left(\frac{\pi}{8}\right)^3 \|\mathbf{w}(t) - \mathbf{v}\|^2 \geq \frac{\alpha^4 \beta \gamma^2}{210} |\mathbf{w}(t) - \mathbf{v}\|^2.$$

Set $\lambda = \frac{\alpha^4 \beta \gamma^2}{210}$, in total we have that:

$$\frac{\partial}{\partial t} \|\mathbf{w}(t) - \mathbf{v}\|^2 = -2\langle \nabla F(\mathbf{w}(t)), \mathbf{w}(t) - \mathbf{v} \rangle \leq -\lambda \|\mathbf{w}(t) - \mathbf{v}\|^2 \, .$$

Using Grönwall's inequality, this proves that for every $t > 0$ we get:

$$\|\mathbf{w}(t) - \mathbf{v}\|^2 \leq \|\mathbf{w}(0) - \mathbf{v}\|^2 \exp(-\lambda t).$$

■

C.2. Gradient Descent

Proof [Thm. 5.3(2)] Assume that $\|\mathbf{w}_t - \mathbf{v}\|^2 < 1$ for some $t \geq 0$, then we have that $\theta(\mathbf{w}_t, \mathbf{v}) \leq \frac{\pi}{2}$. Thus, we can use Thm. 4.2 with $\delta = \frac{\pi}{2}$ to get that:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{v}\|^2 &= \|\mathbf{w}_t - \eta \nabla F(\mathbf{w}_t) - \mathbf{v}\|^2 \\ &= \|\mathbf{w}_t - \mathbf{v}\|^2 - 2\eta \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v} \rangle + \eta^2 \|\nabla F(\mathbf{w}_t)\|^2 \\ &\leq \|\mathbf{w}_t - \mathbf{v}\|^2 (1 - \eta\lambda) + \eta^2 \|\nabla F(\mathbf{w}_t)\|^2. \end{aligned}$$

Now to bound the second term of the above expression recall the definition of $\nabla F(\mathbf{w}_t)$ to get:

$$\begin{aligned} \|\nabla F(\mathbf{w}_t)\|^2 &= \mathbb{E}_{\mathbf{x}} \left[\left(\sigma(\mathbf{w}_t^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x}) \right)^2 \cdot \sigma'(\mathbf{w}^\top \mathbf{x})^2 \mathbf{x}^\top \mathbf{x} \right] \\ &\leq c_2^4 \mathbb{E}_{\mathbf{x}} \left[\left(\mathbf{w}_t^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x} \right)^2 \cdot \mathbf{x}^\top \mathbf{x} \right] \\ &\leq c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2 \mathbb{E}_{\mathbf{x}} \left[\|\mathbf{x}\|^2 \cdot \mathbf{x}^\top \mathbf{x} \right] \leq c_1^2 c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2 \end{aligned}$$

where in the first inequality we used that σ is monotonic with bounded derivative, and in the second inequality we used Cauchy-Schwartz. Note that by our choice of η:

$$1 - \eta\lambda + \eta^2 c < 1 - \frac{\eta\lambda}{2} < 1,$$

this proves that:

$$\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 \leq (1 - \eta\lambda + \eta^2 c) \|\mathbf{w}_t - \mathbf{v}\|^2 \leq \left(1 - \frac{\eta\lambda}{2}\right) \|\mathbf{w}_t - \mathbf{v}\|^2 \tag{9}$$

and in particular $\|\mathbf{w}_{t+1} - \mathbf{v}\| < 1$. Now after T iterations we can use Eq. (9) iteratively to get that:

$$\begin{aligned} \|\mathbf{w}_T - \mathbf{v}\|^2 \ &\leq \ \left(1 - \frac{\eta\lambda}{2}\right) \|\mathbf{w}_{T-1} - \mathbf{v}\|^2 \\ &\leq \ \dots \ \leq \ \left(1 - \frac{\eta\lambda}{2}\right)^T \|\mathbf{w}_0 - \mathbf{v}\|^2 \, . \end{aligned}$$

■

C.3. Stochastic Gradient Descent

First, we prove a recursion relation similar to the one in the gradient descent step. Only here since each gradient step is stochastic we can only prove that the recursion relation holds in expectation over the example selected in each iteration.

Lemma C.1 *Suppose that $\|\mathbf{w}_t - \mathbf{v}\|^2 \leq 1 - \epsilon$. Then*

$$\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 | \mathbf{w}_t\right] \leq (1 - 2\eta\lambda + \eta^2 c)\|\mathbf{w}_t - \mathbf{v}\|^2$$

where $c = c_1^2 c_2^4$.

Proof We can use Thm. 4.2 with $\delta = \frac{\pi}{2}$ to get that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{v}\|^2 | \mathbf{w}_t\right] &= \mathbb{E}\left[\|\mathbf{w}_t - \eta g_t - \mathbf{v}\|^2 | \mathbf{w}_t\right] \\ &= \|\mathbf{w}_t - \mathbf{v}\|^2 - 2\eta\mathbb{E}[\langle g_t, \mathbf{w}_t - \mathbf{v}\rangle | \mathbf{w}_t] + \eta^2\mathbb{E}[\|g_t\|^2 | \mathbf{w}_t] \\ &= \|\mathbf{w}_t - \mathbf{v}\|^2 - 2\eta\langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{v}\rangle + \eta^2\|\nabla F(\mathbf{w}_t)\|^2 \\ &\leq \|\mathbf{w}_t - \mathbf{v}\|^2(1 - 2\eta\lambda) + \eta^2\|\nabla F(\mathbf{w}_t)\|^2 .\end{aligned}$$

Now to bound the second term recall the definition of $\nabla F(\mathbf{w}_t)$ to get:

$$\begin{aligned}\|\nabla F(\mathbf{w}_t)\|^2 &= \mathbb{E}_{\mathbf{x}}\left[\left(\sigma(\mathbf{w}_t^\top \mathbf{x}) - \sigma(\mathbf{v}^\top \mathbf{x})\right)^2 \cdot \sigma'(\mathbf{w}^\top \mathbf{x})^2 \mathbf{x}^\top \mathbf{x}\right] \\ &\leq c_2^4 \mathbb{E}_{\mathbf{x}}\left[\left(\mathbf{w}_t^\top \mathbf{x} - \mathbf{v}^\top \mathbf{x}\right)^2 \cdot \mathbf{x}^\top \mathbf{x}\right] \\ &\leq c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2 \mathbb{E}_{\mathbf{x}}\left[\|\mathbf{x}\|^2 \cdot \mathbf{x}^\top \mathbf{x}\right] \leq c_1^2 c_2^4 \|\mathbf{w}_t - \mathbf{v}\|^2\end{aligned}$$

where in the first inequality we used that σ is monotonic with bounded derivative, and in the second inequality we used Cauchy-Schwartz. This proves the required bound. ■

The recursion relation above only works if $\mathbf{w}_t$ is in a "safe zone", that is $\|\mathbf{w}_t - \mathbf{v}\|^2 \leq 1 - \epsilon$. Although in expectation the distance between $\mathbf{w}_t$ and $\mathbf{v}$ only decrease, taking a stochastic step may take $\mathbf{w}_{t+1}$ outside of the safe zone. The following lemma shows that if η is small enough, then taking at most $m = O(1/\eta)$ steps keeps $\mathbf{w}_t$ in the "safe zone" w.h.p for every $t = 1, \dots, m$.

Lemma C.2 *Assume that $\|\mathbf{w}_0 - \mathbf{v}\|^2 \leq 1 - \epsilon$, and Let $\delta > 0$. Then w.p $> 1 - \delta$, if $\eta < \frac{\epsilon^2\lambda}{3c_1^2 c_2^4 \log\left(\frac{1}{\delta}\right)}$ and $m \leq \frac{1}{9\eta c_1 c_2^2}$ then for every $i = 1, \dots, m$ we have that $\|\mathbf{w}_i - \mathbf{v}\|^2 \leq 1 - \frac{\epsilon}{2}$.*

Proof Denote $X_i = \|\mathbf{w}_i - \mathbf{v}\|^2$, then we have:

$$\begin{aligned}|X_i - X_{i-1}| &= \left|\|\mathbf{w}_i - \mathbf{v}\|^2 - \|\mathbf{w}_{i-1} - \mathbf{v}\|^2\right| = \left|\|\mathbf{w}_{i-1} - \eta\mathbf{g}_{i-1} - \mathbf{v}\|^2 - \|\mathbf{w}_{i-1} - \mathbf{v}\|^2\right| \\ &= \left|-2\eta\langle \mathbf{g}_{i-1}, \mathbf{w}_{i-1} - \mathbf{v}\rangle + \eta^2\|\mathbf{g}_{i-1}\|^2\right| \leq 2\eta|\langle \mathbf{g}_{i-1}, \mathbf{w}_{i-1} - \mathbf{v}\rangle| + \eta^2\|\mathbf{g}_{i-1}\|^2 \quad (10)\end{aligned}$$

We will bound the norm of the gradient at each step:

$$\|\mathbf{g}_i\|^2 = \mathbf{x}_i^\top \mathbf{x}_i \sigma'\left(\mathbf{w}_i^\top \mathbf{x}_i\right)^2 \left(\sigma\left(\mathbf{w}_i^\top \mathbf{x}_i\right) - \sigma\left(\mathbf{v}^\top \mathbf{x}_i\right)\right)^2 \leq c_1^2 c_2^4 \|\mathbf{w}_i - \mathbf{v}\|^2$$

thus we can bound Eq. (10) with:

$$|X_i - X_{i-1}| \leq \|\mathbf{w}_{i-1} - \mathbf{v}\|^2 c_1^2 c_2^4 (2\eta + \eta^2) \leq 3\eta c_1^2 c_2^4 \|\mathbf{w}_{i-1} - \mathbf{v}\|^2 \tag{11}$$

Denote $\eta' = 3\eta c_1^2 c_2^4$. Using Eq. (10) we can bound:

$$\|\mathbf{w}_i - \mathbf{v}\|^2 \leq \|\mathbf{w}_{i-1} - \mathbf{v}\|^2 + \eta' \|\mathbf{w}_{i-1} - \mathbf{v}\|^2 \leq (1 + \eta')\|\mathbf{w}_{i-1} - \mathbf{v}\|^2 \tag{12}$$

Thus, combining Eq. (11) and Eq. (12) we get:

$$\begin{aligned} |X_i - X_{i-1}| &\leq \eta'(1+\eta')\|\mathbf{w}_{i-2} - \mathbf{v}\|^2 \\ &\leq \ldots \leq \eta'(1+\eta')^{i-2}\|\mathbf{w}_0 - \mathbf{v}\|^2 \leq \eta'(1+\eta')^i(1-\epsilon) \end{aligned}$$

We would like to use Azuma's inequality on X_i, but in order to prove that they are supermartingales we need to use Lemma C.1. The problem here is that the condition of the lemma, that $\|\mathbf{w}_t - \mathbf{v}\|^2 < 1 - \epsilon$, does not necessarily holds, hence the series X_i may not be supermartingales. Instead, we consider a dual series of random variables $\tilde{X}_i = \min\left\{X_i, 1 - \frac{\epsilon}{2}\right\}$, and prove that they are supermartingales. First we have that:

$$\left|\tilde{X}_i - \tilde{X}_{i-1}\right| \leq |X_i - X_{i-1}| \leq \eta'(1+\eta')^i(1-\epsilon).$$

Next, we have for every i that $\tilde{X}_i \leq 1 - \frac{\epsilon}{2}$, thus we can use Lemma C.1 (note that the result of the lemma does not depend on the value of ϵ) and choose $\eta' \leq \frac{\lambda}{c_1^2 c_2^4}$ to get that:

$$\mathbb{E}[\tilde{X}_i | \mathbf{w}_{i-1}] \leq \min\{(1 - 2\eta'\lambda + \eta'^2 c_1^2 c_2^4)X_{i-1}, 1 - \epsilon\} \leq \tilde{X}_{i-1}$$

this proves that the series $\tilde{X}_i$ are supermartingales. Now we use a maximal version of Azuma-Hoeffding inequality (see (11)) on $\tilde{X}_i$ to show that after m iterations we have that:

$$\begin{aligned} \mathcal{P}\left(\sup_{1\leq i\leq m} \tilde{X}_i - \tilde{X}_0 > \frac{\epsilon}{2}\right) &\leq \exp\left(\frac{-\epsilon^2}{2\sum_{i=0}^m \left(\eta'(1+\eta')^i(1-\epsilon)\right)^2}\right) \\ &\leq \exp\left(\frac{-\epsilon^2}{2\eta'^2(1-\epsilon)^2\frac{(1+\eta')^{2m+2}-1}{(1+\eta')^2-1}}\right) \\ &\leq \exp\left(\frac{-\epsilon^2}{2\eta'^2(1-\epsilon)^2\frac{2}{(1+\eta')^2-1}}\right) \leq \exp\left(\frac{-\epsilon^2(2+\eta')}{4\eta'(1-\epsilon)^2}\right) \end{aligned} \tag{13}$$

where in the second to last inequality we used that $\eta' \leq \frac{1}{2m+2}$ to bound $(1+\eta')^{2m+2} < 3$ for every m. Substituting the r.h.s of Eq. (13) with δ and simplifying the term we get that if $\eta' \leq \frac{\epsilon^2}{\log\left(\frac{1}{\delta}\right)}$ then w.p $> 1 - \delta$, for every $i = 1, \ldots, m$ (note that $\tilde{X}_0 = X_0$):

$$\min\left\{X_i, 1 - \frac{\epsilon}{2}\right\} \leq X_0 - \frac{\epsilon}{2} \leq 1 - \epsilon + \frac{\epsilon}{2} = 1 - \frac{\epsilon}{2}.$$

In particular, the above shows that w.p $> 1 - \delta$ for every $i = 1, \ldots, m$: $X_i = \|\mathbf{w}_i - \mathbf{v}\|^2 \leq 1 - \frac{\epsilon}{2}$. ■

Next we show that taking a single epoch of $m = O(1/\eta)$ iterations w.h.p will decrease the distance between $\mathbf{w}$ and $\mathbf{v}$ by a constant that does not depend on the epoch length or the step size.

Lemma C.3 *Let $\delta > 0$, take $\eta \leq \frac{\lambda \epsilon_1^2 \epsilon_2^2 c_3^2}{60 c_1^3 c_2^6 \log\left(\frac{2}{\delta}\right)}$ where $c_3 = \left(\frac{1}{2}\right)^{\frac{\lambda}{20 c_1 c_2^2}} - \left(\frac{1}{2}\right)^{\frac{\lambda}{18 c_1 c_2^2}}$, and $m = \frac{1}{9 \eta c_1 c_2^2}$. Assume $\epsilon_2 \leq \|\mathbf{w}_0 - \mathbf{v}\|^2 \leq 1 - \epsilon_1$. Then w.p $1 - \delta$ we have that $\|\mathbf{w}_m - \mathbf{v}\|^2 \leq \left(\frac{1}{2}\right)^{\frac{\lambda}{20 c_1 c_2^2}} \|\mathbf{w}_0 - \mathbf{v}\|^2$.*

Proof Denote $\tilde{\mathbf{w}}_i = \mathbb{E}[\mathbf{w}_i]$ where the expectation is over $\mathbf{x}_1, \dots, \mathbf{x}_i$, and let $Z_i = \|\mathbf{w}_i - \tilde{\mathbf{w}}_i\|^2$, then we have that:

$$\begin{aligned}
|Z_i - Z_{i-1}| &= \left| \|\mathbf{w}_i - \tilde{\mathbf{w}}_i\|^2 - \|\mathbf{w}_{i-1} - \tilde{\mathbf{w}}_{i-1}\|^2 \right| \\
&= \left| \|\mathbf{w}_{i-1} - \eta \mathbf{g}_{i-1} - \tilde{\mathbf{w}}_{i-1} + \eta \nabla F(\tilde{\mathbf{w}}_{i-1})\|^2 - \|\mathbf{w}_{i-1} - \tilde{\mathbf{w}}_{i-1}\|^2 \right| \\
&\leq 2\eta \left| \langle \nabla F(\tilde{\mathbf{w}}_{i-1}) - \mathbf{g}_{i-1}, \mathbf{w}_{i-1} - \tilde{\mathbf{w}}_{i-1} \rangle \right| + \eta^2 \|F(\tilde{\mathbf{w}}_{i-1}) - \mathbf{g}_{i-1}\|^2 \\
&\leq 2\eta \|F(\tilde{\mathbf{w}}_{i-1}) - \mathbf{g}_{i-1}\| \cdot \|\mathbf{w}_{i-1} - \tilde{\mathbf{w}}_{i-1}\| + \eta^2 \|F(\tilde{\mathbf{w}}_{i-1}) - \mathbf{g}_{i-1}\|^2 \\
&\leq 2\eta \left(\|\nabla F(\tilde{\mathbf{w}}_{i-1})\| + \|\mathbf{g}_{i-1}\| \right) \cdot \left(\|\mathbf{w}_{i-1}\| + \|\tilde{\mathbf{w}}_{i-1}\| \right) + \eta \left(\|\nabla F(\tilde{\mathbf{w}}_{i-1})\|^2 + \|\mathbf{g}_{i-1}\|^2 \right)
\end{aligned} \tag{14}$$

As in the proof of the previous lemma we can bound:

$$\|\mathbf{g}_i\|^2 \leq c_1 c_2^2 \|\mathbf{w}_i - \mathbf{v}\|^2 \leq c_1^2 c_2^4$$

where we used our assumption that $\|\mathbf{w}_i - \mathbf{v}\|^2 \leq 1$. In the same manner we can bound $\|\nabla F(\tilde{\mathbf{w}}_i)\| \leq c_1^2 c_2^4$. Again using our assumption we have that:

$$\|\mathbf{w}_i\| \leq \|\mathbf{v}\| + \|\mathbf{w}_i - \mathbf{v}\| \leq 1 + 1 - \epsilon \leq 2$$

and in the same manner $\|\tilde{\mathbf{w}}_i\| \leq 2$. In total we can bound Eq. (14) by:

$$|Z_i - Z_{i-1}| \leq 16 \eta c_1^2 c_2^4$$

Set $c_3 = \left(\frac{1}{2}\right)^{\frac{\lambda}{20 c_1 c_2^2}} - \left(\frac{1}{2}\right)^{\frac{\lambda}{18 c_1 c_2^2}}$, we now us Azuma's inequality and $Z_0 = 0$ to get that:

$$\mathcal{P}\left(Z_m \geq \epsilon_2 c_3\right) \leq \exp\left(\frac{-\epsilon_2^2 c_3^2}{256 m \eta^2 c_1^4 c_2^8} \right)$$

Substituting the r.h.s with $\frac{\delta}{2}$ we have that for :

$$m \leq \frac{\epsilon_2^2 c_3^2}{512 c_1^4 c_2^8 \eta^2 \log\left(\frac{2}{\delta}\right)} \tag{15}$$

then w.p $> 1 - \frac{\delta}{2}$: $\|\mathbf{w}_m - \tilde{\mathbf{w}}_m\|^2 \leq \epsilon_2 c_3$.

Take $m = \frac{1}{9 \eta c_1 c_2^2}$, by taking $\eta \leq \frac{\lambda \epsilon_1^2 \epsilon_2^2 c_3^2}{60 c_1^3 c_2^6 \log\left(\frac{2}{\delta}\right)}$ we have that Eq. (15) is satisfied and $1 - \eta\lambda + \eta^2 c \leq 1 - \frac{\eta\lambda}{2}$. Finally, using Lemma C.2 with $\frac{\delta}{2}$ and using a union bound, we get that after m

iterations w.p $> 1 - \delta$:

$$\begin{aligned}
\|\mathbf{w}_m - \mathbf{v}\|^2 &\le \|\tilde{\mathbf{w}}_m - \mathbf{v}\|^2 + \|\mathbf{w}_m - \tilde{\mathbf{w}}_m\|^2 \\
&\le \left(1 - \eta\lambda + \eta^2 c\right)^m \|\mathbf{w}_0 - \mathbf{v}\|^2 + \epsilon_2 c_3 \\
&\le \left(1 - \frac{\eta\lambda}{2}\right)^m \|\mathbf{w}_0 - \mathbf{v}\|^2 + \left(\left(\frac{1}{2}\right)^{\frac{\lambda}{20c_1c_2^2}} - \left(\frac{1}{2}\right)^{\frac{\lambda}{18c_1c_2^2}}\right) \|\mathbf{w}_0 - \mathbf{v}\|^2 \\
&\le \left(\left(1 - \frac{\eta\lambda}{2}\right)^{\frac{2}{\lambda\eta}}\right)^{\frac{\lambda}{18c_1c_2^2}} \|\mathbf{w}_0 - \mathbf{v}\|^2 + \left(\left(\frac{1}{2}\right)^{\frac{\lambda}{20c_1c_2^2}} - \left(\frac{1}{2}\right)^{\frac{\lambda}{18c_1c_2^2}}\right) \|\mathbf{w}_0 - \mathbf{v}\|^2 \\
&\le \left(\frac{1}{2}\right)^{\frac{\lambda}{18c_1c_2^2}} \|\mathbf{w}_0 - \mathbf{v}\|^2 + \left(\left(\frac{1}{2}\right)^{\frac{\lambda}{20c_1c_2^2}} - \left(\frac{1}{2}\right)^{\frac{\lambda}{18c_1c_2^2}}\right) \|\mathbf{w}_0 - \mathbf{v}\|^2 \\
&\le \left(\frac{1}{2}\right)^{\frac{\lambda}{20c_1c_2^2}} \|\mathbf{w}_0 - \mathbf{v}\|^2
\end{aligned}$$

where in the second to last inequality we used that $(1+x)^{\frac{1}{x}} \le \frac{1}{2}$ for $0 \le x \le 1$.

■

Now we are ready to prove the main theorem, by taking enough epochs with m iterations, and applying union bound:

Proof [Thm. 5.3(3)] We use Lemma C.3 to get that after $m = \frac{1}{9\eta c_1 c_2^2}$ iterations we have w.p $1 - \delta$

$$\|\mathbf{w}_m - \mathbf{v}\|^2 \le \left(\frac{1}{2}\right)^{\frac{\lambda}{20c_1c_2^2}} \|\mathbf{w}_0 - \mathbf{v}\|^2.$$

Using the above iteratively for t epochs and applying union bound, we have that after $T = t \cdot m$ iterations w.p $1 - t\delta$:

$$\|\mathbf{w}_{t\cdot m} - \mathbf{v}\|^2 \le \left(\frac{1}{2}\right)^{\frac{t\lambda}{20c_1c_2^2}} \|\mathbf{w}_0 - \mathbf{v}\|^2 \le \left(\frac{1}{2}\right)^{\frac{t\lambda}{20c_1c_2^2}}.$$

Setting $t = \left\lceil \frac{20c_1c_2^2 \log\left(\frac{1}{\epsilon_2}\right)}{\lambda} \right\rceil$ we have w.p $> 1 - \left\lceil \frac{20c_1c_2^2 \log\left(\frac{1}{\epsilon_2}\right)}{\lambda} \right\rceil \delta$, after $T = t \cdot m = \frac{2\log\left(\frac{1}{\epsilon_2}\right)}{\lambda\eta}$ iterations we have:

$$\|\mathbf{w}_T - \mathbf{v}\|^2 \le \left(\frac{1}{2}\right)^{\frac{t\lambda}{20c_1c_2^2}} \le \epsilon_2$$

■

Appendix D. Proofs from Sec. 6

In the proofs of this section, we follow the convention that for the ReLU function $\sigma(\cdot)$, it holds that $\sigma'(z) = \mathbf{1}(z \ge 0)$ (and in particular, that $\sigma'(0) = 1$). However, the same proofs will hold assuming any other value of $\sigma'(0)$ in $[0, 1]$.

Proof [Lemma 6.2] Using the chain rule and the lemma assumption that $\|\mathbf{w}(t)\| > 0$ (hence the angle expression is well-defined), we have

$$\begin{aligned}\frac{\partial}{\partial t}\theta(\mathbf{w}(t),\mathbf{v}) &= \frac{\partial}{\partial t}\arccos\left(\frac{\mathbf{w}(t)^\top\bar{\mathbf{v}}}{\|\mathbf{w}(t)\|}\right)\\ &= -\frac{1}{\sqrt{1-\left(\frac{\mathbf{w}(t)^\top\bar{\mathbf{v}}}{\|\mathbf{w}(t)\|}\right)^2}}\cdot\left(\frac{\|\mathbf{w}(t)\|\bar{\mathbf{v}}-(\mathbf{w}(t)^\top\bar{\mathbf{v}})\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}}{\|\mathbf{w}(t)\|^2}\right)^\top(-\nabla F(\mathbf{w}(t)))\\ &= \frac{1}{\sqrt{1-(\bar{\mathbf{w}}(t)^\top\bar{\mathbf{v}})^2}}\cdot\left(\frac{\bar{\mathbf{v}}-(\bar{\mathbf{w}}(t)^\top\bar{\mathbf{v}})\bar{\mathbf{w}}(t)}{\|\mathbf{w}(t)\|}\right)^\top\nabla F(\mathbf{w}(t))\,.\end{aligned}$$

Thus, it is enough to show that:

$$\left(\mathbf{v}-\frac{(\bar{\mathbf{w}}(t)^\top\mathbf{v})}{\|\mathbf{w}(t)\|}\mathbf{w}(t)\right)^\top\nabla F(\mathbf{w}(t))\le 0.$$

We fix $\mathbf{w}=\mathbf{w}(t)$, and denote $a=\frac{\bar{\mathbf{w}}^\top\mathbf{v}}{\|\mathbf{w}\|}$. Plugging in the definition of $\nabla F(\mathbf{w})$, we want to show that

$$\mathbb{E}_{\mathbf{x}}\left[\Big(\sigma(\mathbf{w}^\top\mathbf{x})-\sigma(\mathbf{v}^\top\mathbf{x})\Big)\cdot\sigma'(\mathbf{w}^\top\mathbf{x})\cdot(\mathbf{v}^\top\mathbf{x}-a\mathbf{w}^\top\mathbf{x})\right]\le 0\,.$$

Using the assumption that σ is ReLU, the above can be rewritten as

$$\mathbb{E}_{\mathbf{x}}\left[\Big(\sigma(\mathbf{w}^\top\mathbf{x})-\sigma(\mathbf{v}^\top\mathbf{x})\Big)\cdot(\mathbf{v}^\top\mathbf{x}-a\mathbf{w}^\top\mathbf{x})\cdot\mathbb{1}(\mathbf{w}^\top\mathbf{x}\ge 0)\right]\le 0\,. \tag{16}$$

We now note that the expression above depends only on inner products of $\mathbf{x}$ with $\mathbf{w},\mathbf{v}$, so we can rewrite the inequality as

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{D}_{\mathbf{w},\mathbf{v}}}\left[\Big(\sigma(\hat{\mathbf{w}}^\top\mathbf{y})-\sigma(\hat{\mathbf{v}}^\top\mathbf{y})\Big)\cdot(\hat{\mathbf{v}}^\top\mathbf{y}-a\hat{\mathbf{w}}^\top\mathbf{y})\cdot\mathbb{1}(\hat{\mathbf{w}}^\top\mathbf{y}\ge 0)\right]\le 0\,,$$

where $\mathcal{D}_{\mathbf{w},\mathbf{v}}$ is the marginal distribution of $\mathbf{x}$ on the 2-dimensional subspace $\text{span}\{\mathbf{w},\mathbf{v}\}$, and $\hat{\mathbf{w}},\hat{\mathbf{v}}\in\mathbb{R}^2$ are the representations of $\mathbf{w},\mathbf{v}$ in that subspace. Moreover, by the spherical symmetry of the distribution, the expression above is invariant to rotating the coordinate frame, so we can assume without loss of generality that $\hat{\mathbf{w}}=\|\mathbf{w}\|\begin{pmatrix}1\\0\end{pmatrix}$, in which case the above reduces to

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{D}_{\mathbf{w},\mathbf{v}}}\left[\left(\|\mathbf{w}\|\begin{pmatrix}1\\0\end{pmatrix}^\top\mathbf{y}-\sigma(\hat{\mathbf{v}}^\top\mathbf{y})\right)\cdot\left(\hat{\mathbf{v}}^\top\mathbf{y}-\langle\bar{\mathbf{w}},\mathbf{v}\rangle\begin{pmatrix}1\\0\end{pmatrix}^\top\mathbf{y}\right)\cdot\mathbb{1}(y_1>0)\right]\le 0\,.$$

Denote $g(\mathbf{y})=\left(\|\mathbf{w}\|\begin{pmatrix}1\\0\end{pmatrix}^\top\mathbf{y}-\sigma(\hat{\mathbf{v}}^\top\mathbf{y})\right)\cdot\left(\hat{\mathbf{v}}^\top\mathbf{y}-\langle\bar{\mathbf{w}},\mathbf{v}\rangle\begin{pmatrix}1\\0\end{pmatrix}^\top\mathbf{y}\right)$, so that the inequality above is

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{D}_{\mathbf{w},\mathbf{v}}}[\,g(\mathbf{y})\cdot\mathbb{1}(y_1>0)]\le 0\,. \tag{17}$$

The function $g(\mathbf{y})$ can be simplified as:

$$g(\mathbf{y})=(\|\mathbf{w}\|y_1-\sigma(y_1\hat{v}_1+y_2\hat{v}_2))\cdot(y_1\hat{v}_1+y_2\hat{v}_2-\hat{v}_1y_1)=(\|\mathbf{w}\|y_1-\sigma(y_1\hat{v}_1+y_2\hat{v}_2))\cdot y_2\hat{v}_2\,,$$

where we used the fact that $\langle \bar{\mathbf{w}}, \mathbf{v}\rangle = \langle \frac{1}{\|\mathbf{w}\|}\hat{\mathbf{w}}, \hat{\mathbf{v}}\rangle = v_1$.

We now perform a case analysis to justify Eq. (17), depending on the value of a (which by definition, equals $\frac{\bar{\mathbf{w}}^\top \mathbf{v}}{\|\mathbf{w}\|} = \frac{\mathbf{w}^\top \mathbf{v}}{\|\mathbf{w}\|^2} = \frac{\hat{\mathbf{w}}^\top \hat{\mathbf{v}}}{\|\mathbf{w}\|^2} = \frac{\hat{v}_1}{\|\mathbf{w}\|}$). In all the cases we assume $y_1 > 0$, otherwise the expression in the expectation is zero.

- $0 \leq a \leq 1$: In this case $\hat{v}_1 \geq 0$, and also $\langle \bar{\mathbf{w}}, \mathbf{v}\rangle \leq \|\mathbf{w}\|$. Assume w.l.o.g that $\hat{v}_2 \geq 0$ (the other case is similar), and for $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ denote $\tilde{\mathbf{y}} = \begin{pmatrix} y_1 \\ -y_2 \end{pmatrix}$. If $y_2 < 0$ then $g(\mathbf{y}) \leq 0$, on the other hand if $y_2 > 0$ then we can rewrite:

  $$g(\mathbf{y}) = y_2\hat{v}_2 \cdot (y_1(\|\mathbf{w}\| - \hat{v}_1) - y_2\hat{v}_2) = y_2\hat{v}_2 \cdot (y_1(\|\mathbf{w}\| - \langle \bar{\mathbf{w}}, \mathbf{v}\rangle) - y_2\hat{v}_2),$$

  where we have two cases:

  1. if $y_1(\|\mathbf{w}\| - \langle \bar{\mathbf{w}}, \mathbf{v}\rangle) > y_2\hat{v}_2$ then $|g(\tilde{\mathbf{y}})| \geq g(\mathbf{y})$ and also $g(\tilde{\mathbf{y}}) \leq 0$
  2. If $y_1(\|\mathbf{w}\| - \langle \bar{\mathbf{w}}, \mathbf{v}\rangle) \leq y_2\hat{v}_2$ then $g(\mathbf{y}) \leq 0$.

  We showed that for every $\mathbf{y} \in \mathbb{R}^2$ either $g(\mathbf{y}) \leq 0$ or there is a unique $\tilde{\mathbf{y}} \in \mathbb{R}^2$ with the same norm as $\mathbf{y}$ such that $g(\tilde{\mathbf{y}}) \leq 0$ and $|g(\tilde{\mathbf{y}})| \geq g(\mathbf{y})$. Since $\mathcal{D}$ has a spherical symmetric distribution this shows that Eq. (17) holds for these values of a.

- $a \leq 0$: In this case $\hat{v}_1 \leq 0$, we also assume w.l.o.g that $\hat{v}_2 \geq 0$ (the other case is similar). Here for every $\mathbf{y}$ with $y_2 \leq 0$ we have that:

  $$g(\mathbf{y}) = (\|\mathbf{w}\|y_1 - \sigma(y_1\hat{v}_1 + y_2\hat{v}_2)) \cdot y_2\hat{v}_2 = \|\mathbf{w}\|y_1 \cdot y_2\hat{v}_2 \leq 0,$$

  because $y_1 \geq 0$. On the other hand, if $y_2 \geq 0$ we have two cases:

  1. If also $\hat{v}_1 y_1 + \hat{v}_2 y_2 \leq 0$ then $g(\mathbf{y}) = \|\mathbf{w}\|y_1 \cdot y_2\hat{v}_2 \geq 0$, and then $g(\tilde{\mathbf{y}}) = -g(\mathbf{y})$.
  2. If $\hat{v}_1 y_1 + \hat{v}_2 y_2 \geq 0$ then $g(\mathbf{y}) = (\|\mathbf{w}\|y_1 - \hat{v}_1 y_1 - \hat{v}_2 y_2) \cdot y_2\hat{v}_2$. If $g(\mathbf{y}) \geq 0$, then $g(\tilde{\mathbf{y}}) \leq 0$ and also $|g(\tilde{\mathbf{y}})| \geq g(\mathbf{y})$.

  Hence we proved that for every $\mathbf{y}$ with $y_1 > 0$ either $g(\mathbf{y}) \leq 0$ or there is $\tilde{\mathbf{y}}$ with $|g(\tilde{\mathbf{y}})| \geq g(\mathbf{y})$ and $g(\tilde{\mathbf{y}}) \leq 0$. Since $\mathcal{D}$ has a spherical symmetric distribution this shows that Eq. (17) holds for these values of a.

- $a \geq 1$: In this case $\hat{v}_1 \geq 0$ and $\langle \hat{\mathbf{w}}, \mathbf{v}\rangle \geq \|\mathbf{w}\|$. Assume w.l.o.g that $\hat{v}_2 \geq 0$ (the other case is similar). If $y_2 > 0$ then

  $$g(\mathbf{y}) = (y_1(\|\mathbf{w}\| - \hat{v}_1) - y_2\hat{v}_2) \cdot y_2\hat{v}_2 \leq 0.$$

  If $y_2 < 0$ then we have two case:

  1. $y_1\hat{v}_1 + y_2\hat{v}_2 \leq 0$, then $g(\mathbf{y}) = \|\mathbf{w}\|y_1 \cdot y_2\hat{v}_2 < 0$
  2. $y_1\hat{v}_2 + y_2\hat{v}_2 > 0$, in which case if $g(\mathbf{y}) > 0$ then $g(\tilde{\mathbf{y}}) < 0$ and $g(\tilde{\mathbf{y}}) \geq g(\mathbf{y})$.

  Hence for every $\mathbf{y}$ with $y_1 > 0$ either $g(\mathbf{y}) < 0$ or there is $\tilde{\mathbf{y}}$ with $g(\tilde{\mathbf{y}}) < 0$ and $g(\tilde{\mathbf{y}}) \geq g(\mathbf{y})$. This shows that Eq. (17) holds for these values of a.

■

Proof [Lemma 6.3] By our assumption $\mathbf{w}(t) \neq 0$, hence the gradient of the objective is well-defined and we have that

$$\frac{\partial}{\partial t}\|\mathbf{w}(t)\|^2 = -\mathbf{w}(t)^\top \nabla F(\mathbf{w}(t)) = \mathbb{E}_{\mathbf{x}}\left[\left(\sigma(\mathbf{v}^\top\mathbf{x}) - \sigma(\mathbf{w}(t)^\top\mathbf{x})\right)\sigma'(\mathbf{w}(t)^\top\mathbf{x})\mathbf{w}(t)^\top\mathbf{x}\right] . \tag{18}$$

Fix $\mathbf{w} = \mathbf{w}(t)$. Using the assumption that σ is the ReLU function we can rewrite Eq. (18) as:

$$\mathbb{E}_{\mathbf{x}}\left[\left(\sigma(\mathbf{v}^\top\mathbf{x}) - \mathbf{w}^\top\mathbf{x}\right)\cdot \mathbf{w}^\top\mathbf{x}\cdot\mathbb{1}(\mathbf{w}^\top\mathbf{x}\geq 0)\right] . \tag{19}$$

Since the function inside the expectation in Eq. (19) depends only on the inner product of $\mathbf{x}$ with $\mathbf{w}$ and $\mathbf{v}$, we can consider the marginal distribution $\mathcal{D}_{\mathbf{w},\mathbf{v}}$ on the 2-dimensional subspace span$\{\mathbf{w},\mathbf{v}\}$, we also denote $\hat{\mathbf{w}},\hat{\mathbf{v}}\in\mathbb{R}^2$ as the representations of $\mathbf{w},\mathbf{v}$ on this 2-dimensional subspace. We can now rewrite Eq. (19) as:

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{D}_{\mathbf{w},\mathbf{v}}}\left[\left(\sigma(\hat{\mathbf{v}}^\top\mathbf{y}) - \hat{\mathbf{w}}^\top\mathbf{y}\right)\cdot\hat{\mathbf{w}}^\top\mathbf{y}\cdot\mathbb{1}(\hat{\mathbf{w}}^\top\mathbf{y}\geq 0)\right] . \tag{20}$$

Note that the function inside the expectation in Eq. (20) is homogeneous with respect to the norm of $\mathbf{y}$. Also, by our assumption $\mathcal{D}$ is a spherically symmetric distribution, hence also $\mathcal{D}_{\mathbf{w},\mathbf{v}}$ is spherically symmetric. Thus, in order to prove that Eq. (20) is non-negative, it is enough to consider the conditional distribution $\mathcal{D}_{\mathbf{w},\mathbf{y},1}$ of $\mathbf{y}$ on the set $\{\mathbf{y}:\|\mathbf{y}\|=1\}$. Since $\mathcal{D}_{\mathbf{w},\mathbf{v},1}$ (as a distribution on $\mathbb{R}^2$) is still spherically symmetric, it is invariant to a rotation of the coordinate system, so we can assume w.l.o.g that $\hat{\mathbf{w}} = \|\mathbf{w}\|\begin{pmatrix}1\\0\end{pmatrix}$. Overall, in order to prove that Eq. (20) is non-negative it is enough to show that:

$$\mathbb{E}_{\mathbf{y}\sim\mathcal{D}_{\mathbf{w},\mathbf{v},1}}\left[(\sigma(\hat{v}_1y_1+\hat{v}_2y_2) - \|\mathbf{w}\|y_1)\mathbb{1}(y_1\geq 0)\cdot y_1\|\mathbf{w}\|\right]\geq 0 . \tag{21}$$

Since $\mathcal{D}$ is spherically symmetrical and the function inside Eq. (21), the marginal distribution $\mathcal{D}_{\mathbf{w},\mathbf{v},1}$ is actually a uniform distribution on $\{\mathbf{y}\in\mathbb{R}^2:\|\mathbf{y}\|=1\}$. Thus, in order to show that Eq. (21) is non-negative, we can divide it by $\|\mathbf{w}\|$ (which is positive), and show that the following integral is non-negative:

$$\begin{aligned}&\int_0^1\left(\sigma\left(v_1y_1+v_2\sqrt{1-y_1^2}\right)-\|\mathbf{w}\|y_1\right)y_1+\left(\sigma\left(v_1y_1-v_2\sqrt{1-y_1^2}\right)-\|\mathbf{w}\|y_1\right)y_1dy_1\\ =&\int_0^1 y_1\left(\sigma\left(v_1y_1+v_2\sqrt{1-y_1^2}\right)+\sigma\left(v_1y_1-v_2\sqrt{1-y_1^2}\right)\right)-2\|\mathbf{w}\|y_1^2dy_1,\end{aligned}$$

where we wrote $y_2=\pm\sqrt{1-y_1^2}$ since $\|\mathbf{y}\|=1$. We can assume w.l.o.g that $v_2\geq 0$ (the other direction is similar) and write $v_2=\sqrt{1-v_1^2}$, and thus it is enough to prove that:

$$\int_0^1 y_1\sigma\left(v_1y_1+\sqrt{(1-y_1^2)(1-v_1^2)}\right)-2\|\mathbf{w}\|y_1^2dy_1\geq 0 . \tag{22}$$

Denote $\theta=\theta(\mathbf{w},\mathbf{v})$, since $\langle\bar{\mathbf{w}},\mathbf{v}\rangle=\langle\hat{\bar{\mathbf{w}}},\hat{\mathbf{v}}\rangle=v_1$ then $v_1=\cos(\theta)$ and $\sqrt{1-v_1^2}=\sin(\theta)$. Now we split into cases for the different values of v_1:

- $v_1 \geq 0$: In this case, if $0 \leq y_1 \leq 1$ then $v_1 y_1 + \sqrt{(1-y_1^2)(1-v_1^2)} \geq 0$, hence the integral in Eq. (22) can be calculated as:
$$\int_0^1 y_1 \left(v_1 y_1 + \sqrt{(1-y_1^2)(1-v_1^2)} \right) - 2\|\mathbf{w}\| y_1^2 dy_1 = \frac{v_1}{3} + \frac{\sqrt{1-v_1^2}}{3} - \frac{2\|\mathbf{w}\|}{3}. \qquad (23)$$
Thus, the above term is non-negative if:
$$\|\mathbf{w}\| \leq \frac{v_1 + \sqrt{1-v_1^2}}{2} = \frac{\sin(\theta) + \cos(\theta)}{2}.$$

- $v_1 \leq 0$: In this case, if $0 \leq y_1 \leq \sqrt{1-v_1^2}$ then $v_1 y_1 + \sqrt{(1-y_1^2)(1-v_1^2)} \geq 0$, and if $\sqrt{1-v_1^2} < y_1 \leq 1$ then $v_1 y_1 + \sqrt{(1-y_1^2)(1-v_1^2)} \leq 0$. Thus, the integral in Eq. (22) can be calculated as:
$$\begin{aligned}&\int_0^{\sqrt{1-v_1^2}} y_1 \left(v_1 y_1 + \sqrt{(1-y_1^2)(1-v_1^2)} \right) - 2\|\mathbf{w}\| y_1^2 dy_1 - \int_{\sqrt{1-v_1^2}}^1 2\|\mathbf{w}\| y_1^2 dy_1 \\ &= -\frac{2\|\mathbf{w}\|}{3} + \frac{v_1^3\sqrt{1-v_1^2}}{3} + \frac{v_1\left(\sqrt{1-v_1^2}\right)^3}{3} + \frac{\sqrt{1-v_1^2}}{3} \\ &= -\frac{2\|\mathbf{w}\|}{3} + \frac{\sqrt{1-v_1^2}(1+v_1)}{3}.\end{aligned}$$
Thus, the above term is non-negative if:
$$\|\mathbf{w}\| \leq \frac{\sqrt{1-v_1^2}(1+v_1)}{2} = \frac{\sin(\theta)(1+\cos(\theta))}{2}$$

■

Proof [Thm. 6.4] Assume we initialized with $\theta(\mathbf{w}(0), \mathbf{v}) \leq \pi - \epsilon$ and $0 < \|\mathbf{w}(0)\| \leq 2$. First we will show that $\mathbf{w}(t) \neq 0$ for all $t > 0$. Assume on the way of contradiction that for some $t > 0$ we have $\mathbf{w}(t) = 0$, and let t_1 be the first time for which it happens. For $t_0 = 0$ we know that $\mathbf{w}(t_0) \neq 0$, and also that for all $t \in [t_0, t_1]$, $\mathbf{w}(t) \neq 0$ and the gradient of the objective is well defined. Hence by Lemma 6.2 we know that $\theta(\mathbf{w}(t), \mathbf{v}) \leq \pi - \epsilon$ for all $t \in [t_0, t_1]$, because the angle can only decrease unless $\mathbf{w}(t) = 0$. But, by Lemma 6.3 we know that if $\|\mathbf{w}(t)\| \leq \max\left\{\frac{\sin(\epsilon)-\cos(\epsilon)}{2}, \frac{\sin(\epsilon)(1-\cos(\epsilon))}{2}\right\}$ then $\frac{\partial}{\partial t}\|\mathbf{w}(t)\| \geq 0$. In particular, for $\epsilon \in (0, \pi]$ and for all $t_0 \leq t < t_1$, we have that $\|\mathbf{w}(t)\|$ is bounded below by $\max\left\{\frac{\sin(\epsilon)-\cos(\epsilon)}{2}, \frac{\sin(\epsilon)(1-\cos(\epsilon))}{2}\right\} > 0$, a contradiction to $\mathbf{w}(t_1) = 0$. This shows that for all $t > 0$ we have that $\mathbf{w}(t) \neq 0$, hence by Lemma 6.2 we know that for every $t > 0$ we will have $\theta(\mathbf{w}(t), \mathbf{v}) \leq \pi - \epsilon$.

Now we can use Thm. 4.2 (where $\gamma = 1$ because of Assumption 6.1(3)) to get:
$$\langle \nabla F(\mathbf{w}(t)), \mathbf{w}(t) - \mathbf{v} \rangle \geq \frac{\alpha^4 \beta}{8\sqrt{2}} \sin\left(\frac{\epsilon}{8}\right)^3 \|\mathbf{w}(t) - \mathbf{v}\|^2.$$

Set $\lambda = \frac{\alpha^4\beta}{8\sqrt{2}} \sin\left(\frac{\epsilon}{8}\right)^3$, as explained above for all $t > 0$, $\nabla F(\mathbf{w}(t))$ is continuous since $\mathbf{w}(t) \neq 0$ and we have that:
$$\frac{\partial}{\partial t}\|\mathbf{w}(t) - \mathbf{v}\|^2 = 2\langle \mathbf{w}(t) - \mathbf{v}, \frac{\partial}{\partial t}\mathbf{w}(t) \rangle = -2\langle \mathbf{w}(t) - \mathbf{v}, \nabla F(\mathbf{w}(t)) \rangle \leq -\lambda \|\mathbf{w}(t) - \mathbf{v}\|^2,$$

Using Grönwall's inequality, this proves that for every $t > 0$ we get:

$$\|\mathbf{w}(t) - \mathbf{v}\|^2 \leq \|\mathbf{w}(0) - \mathbf{v}\|^2 \exp(-\lambda t).$$

■

D.1. Standard Gaussian Distribution

In this subsection we assume that $\mathcal{D} = \mathcal{N}(0, I)$, and that σ is the ReLU function.

Lemma D.1 *If* $\mathbf{w}(t) \neq 0$*, then* $\frac{\partial}{\partial t}\theta(\mathbf{w}(t), \mathbf{v}) \leq 0$

Proof Similar to the proof of Lemma 6.2, it is enough to prove that

$$\left(\bar{\mathbf{v}} - (\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})\bar{\mathbf{w}}(t)\right)^\top \nabla F(\mathbf{w}(t)) \ \leq\ 0\,, \tag{24}$$

where we used that $\|\mathbf{w}(t)\| > 0$ hence the angle expression is differentiable. In the standard Gaussian case, $\nabla F(\mathbf{w}(t))$ has a closed-form expression (see (3), (20)), namely

$$\nabla F(\mathbf{w}) = \frac{1}{2}\mathbf{w} - \frac{1}{2\pi}\left(\|\mathbf{v}\| \sin(\theta(\mathbf{w}, \mathbf{v}))\bar{\mathbf{w}} + (\pi - \theta(\mathbf{w}, \mathbf{v})\mathbf{v})\right)\ . \tag{25}$$

Multiplying this by $\left(\bar{\mathbf{v}} - (\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})\bar{\mathbf{w}}(t)\right)$, and noting that this vector is orthogonal to $\mathbf{w}(t)$ (as it is simply the component of $\bar{\mathbf{v}}$ orthogonal to $\bar{\mathbf{w}}(t)$, we get that

$$\begin{aligned}
\left(\bar{\mathbf{v}} - (\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})\bar{\mathbf{w}}(t)\right)^\top \nabla F(\mathbf{w}(t)) \ &= \ \left(\bar{\mathbf{v}} - (\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})\bar{\mathbf{w}}(t)\right)^\top \left(-\frac{\pi - \theta(\mathbf{w}(t), \mathbf{v})}{2\pi}\mathbf{v}\right) \\
&= \ -\frac{\pi - \theta(\mathbf{w}(t), \mathbf{v})}{2\pi}\left(\bar{\mathbf{v}}^\top \mathbf{v} - (\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})(\bar{\mathbf{w}}(t)^\top \mathbf{v})\right) \\
&= \ -\frac{\pi - \theta(\mathbf{w}(t), \mathbf{v})}{2\pi}\left(\|\mathbf{v}\| - \|\mathbf{v}\|(\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})^2\right) \\
&= \ -\frac{\pi - \theta(\mathbf{w}(t), \mathbf{v})}{2\pi}\left(1 - (\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}})^2\right)\|\mathbf{v}\|\ .
\end{aligned}$$

Since $\theta(\mathbf{w}(t), \mathbf{v}) \in [-\pi, \pi]$ and $\bar{\mathbf{w}}(t)^\top \bar{\mathbf{v}} \in [-1, 1]$, it follows that this expression is non-negative, establishing Eq. (24) and hence the lemma.

■

Lemma D.2 *Let* $\theta(\mathbf{w}(t), \mathbf{v}) = \pi - \alpha$ *and assume that* $\mathbf{w}(t) \neq 0$*. If* $\|\mathbf{w}(t)\| \leq \frac{\|\mathbf{v}\|}{\pi^4}\alpha^3$*, then* $\frac{\partial}{\partial t}\|\mathbf{w}(t)\|^2 \geq 0$

Proof

Using the closed-form expression for $\nabla F(\mathbf{w})$ (see Eq. (25)), we have

$$\begin{aligned}
\frac{\partial}{\partial t}\|\mathbf{w}(t)\|^2 &= \mathbf{w}(t)^\top \frac{\partial}{\partial t}\mathbf{w}(t) \;=\; -\,\mathbf{w}(t)^\top \nabla F(\mathbf{w}(t)) \\
&= -\frac{\|\mathbf{w}(t)\|^2}{2} + \frac{1}{2\pi}\Big(\|\mathbf{v}\|\|\mathbf{w}(t)\|\sin(\theta(\mathbf{w}(t),\mathbf{v})) + (\pi - \theta(\mathbf{w}(t),\mathbf{v})\mathbf{w}(t)^\top\mathbf{v})\Big) \\
&= \frac{\|\mathbf{w}(t)\|\|\mathbf{v}\|}{2\pi}\left(\sin(\theta(\mathbf{w}(t),\mathbf{v})) + (\pi - \theta(\mathbf{w}(t),\mathbf{v}))\bar{\mathbf{w}}(t)^\top\bar{\mathbf{v}} - \frac{\pi\|\mathbf{w}(t)\|}{\|\mathbf{v}\|}\right) \\
&= \frac{\|\mathbf{w}(t)\|\|\mathbf{v}\|}{2}\left(\sin(\theta(\mathbf{w}(t),\mathbf{v})) + (\pi - \theta(\mathbf{w}(t),\mathbf{v}))\cos(\theta(\mathbf{w}(t),\mathbf{v})) - \frac{\pi\|\mathbf{w}(t)\|}{\|\mathbf{v}\|}\right)
\end{aligned}$$

The expression $\sin(\theta)+(\pi-\theta)\cos(\theta)$ can be easily verified to be strictly monotonically decreasing in $\theta \in (0,\pi)$, and equal 0 at $\theta = \pi$. Therefore, if $\theta \le \pi - \alpha$, then the expression above can be lower bounded by

$$\begin{aligned}
&\frac{\|\mathbf{w}(t)\|\|\mathbf{v}\|}{2}\left(\sin(\pi-\alpha) + \alpha\cos(\pi-\alpha) - \frac{\pi\|\mathbf{w}(t)\|}{\|\mathbf{v}\|}\right) \\
&= \frac{\|\mathbf{w}(t)\|\|\mathbf{v}\|}{2}\left(\sin(\alpha) - \alpha\cos(\alpha) - \frac{\pi\|\mathbf{w}(t)\|}{\|\mathbf{v}\|}\right)\;. \qquad (26)
\end{aligned}$$

To slightly simplify this expression, we will now argue that

$$\sin(\alpha) - \alpha\cos(\alpha) \ge \left(\frac{\alpha}{\pi}\right)^3 \quad \forall \alpha \in [0,\pi]\;. \qquad (27)$$

Assuming this inequality holds, we get that Eq. (26) is at least

$$\frac{\|\mathbf{w}(t)\|\|\mathbf{v}\|}{2}\left(\left(\frac{\alpha}{\pi}\right)^3 - \frac{\pi\|\mathbf{w}(t)\|}{\|\mathbf{v}\|}\right)\;,$$

which is non-negative as long as $\|\mathbf{w}(t)\| \le \|\mathbf{v}\|\alpha^3/\pi^4$, proving the lemma. It only remains to establish Eq. (27). We consider two cases:

- If $\alpha \in [0,\pi/2]$, then by a Taylor expansion of $\sin(\alpha), \cos(\alpha)$ around 0, we have that $\sin(\alpha) - \alpha\cos(\alpha)$ is at least

$$\alpha - \frac{\alpha^3}{3!} - \alpha\left(1 - \frac{\alpha^2}{2!} + \frac{\alpha^4}{4!}\right) \;=\; \alpha^3\left(\frac{1}{2!} - \frac{1}{3!} - \frac{\alpha^2}{4!}\right) \;\ge\; \alpha^3\left(\frac{1}{2!} - \frac{1}{3!} - \frac{(\pi/2)^2}{4!}\right)\;,$$

  which is at least $\alpha^3/5$.

- If $\alpha \in \left[\frac{\pi}{2},\pi\right]$, it is easily verified via differentiation that $\sin(\alpha) - \alpha\cos(\alpha) \ge \sin(\alpha)$ is monotonically increasing in α. Therefore, it can be lower bounded by $\sin(\pi/2) - (\pi/2)\cos(\pi/2) = 1 \ge \alpha^3/\pi^3$.

Combining the two cases, Eq. (27) follows. ■