UNDERSTANDING HOW OVER-PARAMETRIZATION LEADS TO ACCELERATION: A CASE OF LEARNING A SINGLE TEACHER NEURON

Anonymous authors
Paper under double-blind review

ABSTRACT

Over-parametrization has become a popular technique in deep learning. It is observed that by over-parametrization, a larger neural network needs a fewer training iterations than a smaller one to achieve a certain level of performance — namely, over-parametrization leads to acceleration in optimization. However, despite that over-parametrization is widely used nowadays, little theory is available to explain the acceleration due to over-parametrization. In this paper, we propose understanding it by studying a simple problem first. Specifically, we consider the setting that there is a single teacher neuron with quadratic activation, where over-parametrization is realized by having multiple student neurons learn the data generated from the teacher neuron. We provably show that over-parametrization helps the iterate generated by gradient descent to enter the neighborhood of a global optimal solution that achieves zero testing error faster. On the other hand, we also point out an issue regarding the necessity of over-parametrization and study how the scaling of the output neurons affects the convergence time.

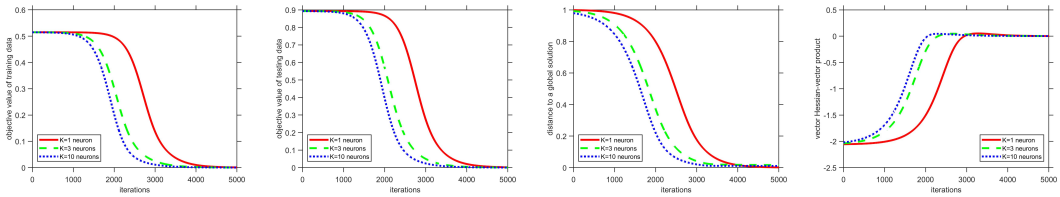
1 INTRODUCTION

Over-parametrization has become a popular technique in deep learning, as it is now widely observed larger neural nets can achieve better performance. Furthermore, a larger network can be trained to achieve a certain level of prediction performance with fewer iterations than that of a smaller net. This observation, to our knowledge, can be dated back as early as the work of Livni et al. (2014), who try different levels of over-parametrization and report that SGD converges much faster and finds a better solution when it is used to train a larger network. However, the reason why over-parametrization can lead to an acceleration still remains a mystery, and very little theory has helped explain the observation, with perhaps the notable exception of (Arora et al., 2018). Arora et al. (2018) consider over-parametrizing a single-output linear regression with l_p loss for $p > 2$–the square loss corresponds to $p = 2$–and they study the linear regression problem by replacing the model $w \in \mathbb{R}^d$ by another model $w_1 \in \mathbb{R}^d$ times a scalar $w_2 \in \mathbb{R}$. They show that the dynamics of gradient descent on the new over-parametrized model are equivalent to the dynamics of gradient descent on the original objective function with an adaptive learning rate plus some momentum terms. However, in practice, people actually use the techniques of over-parametrization, adaptive learning rate, and momentum simultaneously in deep learning (see e.g. (Hoffer et al., 2017; Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Lucas et al., 2019; Sutskever et al., 2013)), as each technique appears to contribute to performance and they may, to some extent, be complementary. It has been suggested that over-parameterizing a model leads implicitly to an adaptive learning rate or momentum, but this does not appear to fully explain the performance improvement.

To understand the benefits of overparameterization, let us begin by studying a simple canonical problem: a single teacher neuron $w_* \in \mathbb{R}^d$ with quadratic activation function. Specifically, the label y_i and the design vector $x_i \sim \mathcal{N}(0, \mathcal{I}_d)$ of sample i satisfies $y_i = (x_i^\top w_*)^2$. Therefore, the *standard* objective function for learning the teacher neuron w_* is

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{4n} \sum_{i=1}^n \left((x_i^\top w)^2 - y_i\right)^2, \tag{1}$$

where n denotes the number of samples. Problem (1) is called *phase retrieval* in signal processing literature, which has real applications in physical science such as astronomy and microscopy (see



(a) Obj. (2) on training data. (b) Obj. (2) on testing data. (c) dist$(W_t^{\#K}, w_*)$ vs. t. (d) Quantity (9) over t.

Figure 1: In the experiment, we set the dimension $d = 10$ and the number of training samples $n = 200$. Additional 200 samples are sampled and served as testing data. We let $w_* = e_1$ with e_1 being the unit vector. Each neuron $w^{(k)} \in \mathbb{R}^d$ ($k \in [K]$) of the student network is initialized by sampling from an isotropic distribution and is close to the origin (i.e. $w_0^{(k)} \sim 0.01 \cdot N(0, I_d/d)$). We apply gradient descent with the same step size $\eta = 0.001$ to train different sizes of neural networks. Each curve represents the progress of gradient descent for different K. Subfigure (a) and (b): Objective value (2) vs. iteration t on training data (testing data, respectively). For subfigure (c) and (d), please see Section 3 for the precise definition and details.

e.g. (Candés et al., 2013), (Fannjiang & Strohmer, 2020), (Shechtman et al., 2015)). There are also some specialized algorithms designed for achieving a better sample complexity or computational complexity to recover w_* modulo the unrecoverable sign (e.g. (Candés & Li, 2014; Candés et al., 2015; 2013; Chen & Candés, 2017; Ma et al., 2017)). We choose this problem as a starting point of understanding acceleration due to over-parametrization. Specifically, we consider the following way to over-parametrize the original objective (1),

$$\min_{W \in \mathbb{R}^{d \times K}} f(W) := \frac{1}{4n} \sum_{i=1}^{n} \left((x_i^\top w^{(1)})^2 + (x_i^\top w^{(2)})^2 + ... + (x_i^\top w^{(K)})^2 - y_i \right)^2, \quad (2)$$

where $w^{(j)}$ denote the j_{th} column of the weight matrix $W \in \mathbb{R}^{d \times K}$. Optimizing the objective function (2) can be viewed as training a student network with K student neurons. While a d dimensional model exists which perfectly predicts the labels generated by the teacher neuron (i.e. $\pm w_*$), one can consider training a much larger model instead (i.e. $d \times K$ number of parameters). Note that if $K = 1$, objective (2) reduces to the original objective (1). On the other hand, a larger $k > 1$ means a higher degree of over-parametrization.

Let us now establish, empirically, the clear advantage of over-parametrization for accelerating the learning and optimization process. We tried $K = \{1, 3, 10\}$ number of student neurons in (2), which represent different degrees of over-parametrization, and we applied gradient descent to train the networks. Figure 1 shows the results. The empirical findings displayed on the figure are quite stark. First, we see that over-parametrization not only helps to decrease the training error faster but also decrease the testing error faster (c.f. subfigure (a) and (b)). Furthermore, the generalization error is very small since both training error and testing error approach zero. Second, regardless of the size K, a common pattern is that the dynamics of gradient descent can be divided into two stages. In the first stage, gradient descent makes little progress on decreasing the function value; while in the second stage, the iterate generated by gradient descent exhibits a linear convergence to a global solution. Specifically, by over-parametrization, gradient descent spends fewer iterations in the first stage and enters the linear convergence regime that makes the fast progress more quickly. We provide more results in Appendix F. Specifically, we also tried different values of the step size η and we observed similar patterns as Figure 1 shows. Even when gradient descent uses the best step size for each model with different neurons K, we still observe that gradient descent enters the linear convergence regime faster for a larger model. Thus, the acceleration due to over-parametrization *cannot* be simply explained by that gradient descent uses a larger *effective* step size, as the effect due to parameters η and K is complementary in the experiment.

In the later sections, we will answer why an over-parametrized network trained by gradient descent can still generalize well and why over-parametrization helps gradient descent to enter the linear convergence regime faster. We will also point out an issue regrading the necessity of over-parametrization in the end.

2 RELATED WORKS

Over-parametrization: Though our work focuses on understanding why over-parametrization leads to acceleration in optimization (i.e. improving the convergence time), we also want to acknowledge some related works of understanding over-parametrization in different aspects (e.g. (Arora et al., 2019; Brutzkus & Globerson, 2019; Emschwiller et al., 2020; Goldt et al., 2019; Tian, 2020) and have a brief review in Appendix A. There is also a trend of works studying how over-parametrization changes the optimization landscape of empirical risk minimization for neural nets with quadratic activation. (e.g. (Du & Lee, 2018; Gamarnik et al., 2019; Ge et al., 2019; Kazemipour et al., 2019; Soltanolkotabi et al., 2018; Nguyen & Hein, 2017; Venturi et al., 2019; Mannelia et al., 2020)). The goals of these works are different from ours. We provide more details in Appendix A.

Quadratic activation and matrix sensing: The optimization landscape of problem (1) (i.e. phase retrieval) and its variants has been studied by (Davis et al., 2018; Soltanolkotabi, 2014; Sun et al., 2016; White et al., 2016), which shows that as long as the number of samples is sufficiently large, it has no spurious local optima and all the local optima are globally optimal. Chen et al. (2019) provably show that applying gradient descent with an isotropic random initialization for solving (1) leads to an optimal solution that recovers the teacher neuron w_* modulo the unrecoverable sign. In this work we show that an over-parametrized student network trained by gradient descent takes even fewer iterations to recover w_*. We also note that the optimization problem (2) can be rewritten as the form of matrix sensing (see e.g. Gunasekar et al. (2017); Li et al. (2019; 2018); Gidel et al. (2019)). In Appendix A, we provide a brief review of matrix sensing.

Learning a single neuron: Studying learning a single neuron in non-convex optimization is not new (e.g. Goel et al. (2019); Yehudai & Shamir (2020); Kakade et al. (2011); Goel et al. (2017); Kalan et al. (2019); Soltanolkotabi (2017); Mei et al. (2018); Frei et al. (2020)). However, those works are not for showing faster convergence by over-parametrization. The goals are different.

3 PRELIMINARIES

Notations and assumptions: We use the notation $W^{\#K} := [w^{(1)}_{\#K}, \dots, w^{(K)}_{\#K}] \in \mathbb{R}^{d\times K}$ to represent the weights of a student network with K number of neurons. Each column k of the matrix $W^{\#K}$ is the weight vector $w^{(k)}_{\#K} \in \mathbb{R}^d$ that corresponds to the neuron k of the student network. Thus, $W^{\#1} := [w^{(1)}_{\#1}] \in \mathbb{R}^d$ is the network consists of a single neuron; while for $K > 1$, the notation represents the weights of an over-parametrized network. In this paper, without loss of generality, we assume that $w_* = \|w_*\|e_1 \in \mathbb{R}^d$ with e_1 being the unit vector whose first element is 1. We define the parallel component $w^{(k),\|}_{\#K,t}$ and the perpendicular component $w^{(k),\perp}_{\#K,t}$ in iteration t as follows,

$$\begin{aligned} w^{(k),\|}_{\#K,t} &:= \langle w^{(k)}_{\#K,t}, w_*\rangle = \|w_*\| w^{(k)}_{\#K,t}[1] \\ w^{(k),\perp}_{\#K,t} &:= [w^{(k)}_{\#K,t}[2], \dots, w^{(k)}_{\#K,t}[d]]^\top. \end{aligned} \tag{3}$$

Namely, the parallel component $w^{(k),\|}_{\#K,t}$ is the projection of a student neuron k learned in iteration t on the teacher neuron w_*; while the perpendicular component $w^{(k),\perp}_{\#K,t}$ is the d-1 dimensional sub-vector of $w^{(k)}_{\#K,t}$ excluding the first element. We assume that each neuron k of each network, $w^{(k)}_{\#K,0} \in \mathbb{R}^d$, is initialized i.i.d. randomly from an isotropic distribution (e.g. gaussian distribution).

Metric of the progress in optimization: The first challenge to show that over-parametrization leads to acceleration in optimization is the design of the metric for the comparison. Since different K's corresponds to different optimization problems, the notion of *acceleration* here is non-standard in optimization literature. In the optimization literature, acceleration usually means that an algorithm takes a fewer iterations than other algorithms for the *same* optimization problem. Fortunately, by exploiting the problem structure, we can have a natural metric of the progress as follows. For any size of the student network $W \in \mathbb{R}^{d\times K}$, we consider

$$\text{dist}(W, w_*) := \min_{q\in\mathbb{R}^K:\|q\|_2\le 1} \|W - w_* q^\top\|. \tag{4}$$

This is due to the observation that for any K, the global optimal solutions of (2) that achieve zero testing error are $w_* q^\top \in \mathbb{R}^{d\times K}$ for any $q \in \mathbb{R}^K$ such that $\|q\|_2 = 1$. To see this, substitute

$W = w_* q^\top \in \mathbb{R}^{d\times K}$ into (2). We have that for any $x_i \in \mathbb{R}^d$ it holds that $(x_i^\top w^{(1)})^2 + (x_i^\top w^{(2)})^2 + ... + (x_i^\top w^{(K)})^2 - y_i = \|x_i^\top W\|_F^2 - (x_i^\top w_*)^2 = \mathrm{tr}\big((x_i^\top w_* q^\top)^\top (x_i^\top w_* q^\top)\big) - (x_i^\top w_*)^2 = 0$. Therefore, the metric $\mathrm{dist}(W, w_*)$ as be viewed as a surrogate of the testing error. In particular, $\mathrm{dist}(W_t, w_*)$ represents the distance of the current iterate W_t and its closest global optimal solution to the over-parametrized objective (2) that achieves zero testing error. Note that the argmin of (55) is $q_* := \frac{W^\top w_*}{\|W^\top w_*\|_2} = \arg\min_{q\in\mathbb{R}^K:\|q\|_2\le 1} \|W - w_* q^\top\|$. On sub-figure (c) of Figure 1, we plot the distance of the iterates generated by gradient descent and its closet global optimal solution for different sizes K of neural nets. We see that over-parametrization enables shrinking the distance $\mathrm{dist}(W_t^{\#K}, w_*)$ faster.

Gradient descent dynamics: For the notation brevity, we will suppress the symbol $\#K$ when it is clear in the context. The gradient of a student neuron $w^{(k)}$ for the over-parametrized problem (2) is

$$\nabla_{w^{(k)}} f(W) := \tfrac{1}{n}\textstyle\sum_{i=1}^n \big((x_i^\top w^{(1)})^2 + (x_i^\top w^{(2)})^2 + \cdots + (x_i^\top w^{(K)})^2 - y_i\big)(x_i^\top w^{(k)})x_i. \tag{5}$$

Its expectation, which is the population gradient of a student neuron $w^{(k)}$ (i.e. gradient when the number of samples n is infinite), is

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim N(0,I_d)}[\nabla_{w^{(k)}} f(W)] =& \big(3\|w^{(k)}\|^2 - \|w_*\|^2\big)w^{(k)} - 2(w_*^\top w^{(k)})w_* \\ &+ \textstyle\sum_{j\neq k}^K 2((w^{(j)})^\top w^{(k)})w^{(j)} + \|w^{(j)}\|^2 w^{(k)},\end{aligned} \tag{6}$$

where we use the fact that for any vector $u, v \in \mathbb{R}^d$, $\mathbb{E}_{x\sim N(0,I_d)}[(x^\top u)^3 x] = 3\|u\|^2 u$ and $\mathbb{E}_{x\sim N(0,I_d)}[(x^\top u)^2 (x^\top v)x] = 2(u^\top v)u + \|u\|^2 v$. For $K = 1$, (6) becomes $\mathbb{E}[\nabla_{w^{(1)}} f(W^{\#1})] = \big(3\|w^{(1)}\|^2 - \|w_*\|^2\big)w^{(1)} - 2(w_*^\top w^{(1)})w_*$. If gradient descent uses the population gradient for the update (i.e. $w_{\#1,t+1}^{(1)} = w_{\#1,t}^{(1)} - \eta\mathbb{E}[\nabla_{w^{(1)}} f(W_t^{\#1})]$), then the dynamics of the student network consists of a single neuron (i.e. $K = 1$) evolves as follows,

$$\begin{aligned} w_{t+1}^{(1),\|} &= w_t^{(1),\|}\big(1 + \eta(3\|w_*\|^2 - 3\|w_t^{(1)}\|^2)\big) \\ w_{t+1}^{(1),\perp} &= w_t^{(1),\perp}\big(1 + \eta(\|w_*\|^2 - 3\|w_t^{(1)}\|^2)\big). \end{aligned} \tag{7}$$

On the other hand, if a student network has $K > 1$ neurons, the dynamics of each neuron k of the student network evolves as follows,

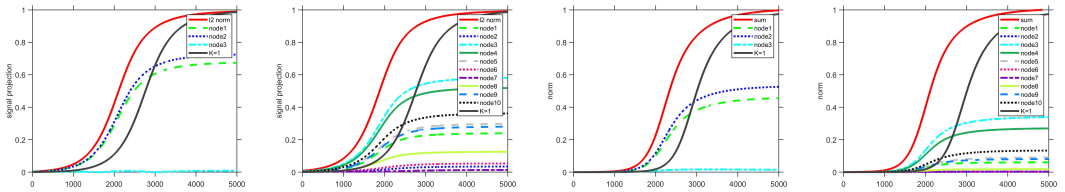
$$\begin{aligned} w_{t+1}^{(k),\|} &= \underbrace{w_t^{(k),\|}\big(1 + 3\eta(\|w_*\|^2 - \|w_t^{(k)}\|^2)\big)}_{\text{component A}} - \underbrace{\big(2\eta\sum_{j\neq k}^K \langle w_t^{(j)}, w_t^{(k)}\rangle w_t^{(j),\|} + \eta w_t^{(k),\|}\sum_{j\neq k}^K \|w_t^{(j)}\|^2\big)}_{\text{component B}} \\ w_{t+1}^{(k),\perp} &= \underbrace{w_t^{(k),\perp}\big(1 + \eta(\|w_*\|^2 - 3\|w_t^{(k)}\|^2)\big)}_{\text{component C}} - \underbrace{\big(2\eta\sum_{j\neq k}^K \langle w_t^{(j)}, w_t^{(k)}\rangle w_t^{(j),\perp} + \eta w_t^{(k),\perp}\sum_{j\neq k}^K \|w_t^{(j)}\|^2\big)}_{\text{component D}} \end{aligned} \tag{8}$$

where both the component B of $w_{t+1}^{(k),\|}$ and the component D of $w_{t+1}^{(k),\perp}$ can be viewed as the terms due to the interaction of student neuron k and the other student neurons.

More observations: On Subfigure (d) of Figure 1, we plot a quantity over iterations, which is

$$\mathrm{vec}(w_* q_t^\top - W_t^{\#K})^\top \nabla^2 f(W_t^{\#K})\mathrm{vec}(w_* q_t^\top - W_t^{\#K}), \tag{9}$$

where $\nabla^2 f(W_t^{\#K}) \in \mathbb{R}^{dK\times dK}$ is the Hessian and $w_* q_t^\top$ is the closet global optimal solution to $W_t^{\#K}$ and the notation $\mathrm{vec}(\cdot)$ represents the vectorization operation of its matrix argument. The quantity can be viewed as a measure of the strong convexity. Specifically, if the quantity is larger than 0, then it suggests that the current optimization landscape is strongly convex with respect to $w_* q_t^\top$. Hence, the observation suggests that gradient descent enters a benign region faster for a larger student network. In the following section, We will answer why over-parametrization helps gradient descent to enter a region that has the benign optimization landscape faster.



(a) $|w_{\#K,t}^{(k),\|}|$ $(K=3)$. (b) $|w_{\#K,t}^{(k),\|}|$ $(K=10)$. (c) $\|w_{\#K,t}^{(k)}\|^2$ $(K=3)$. (d) $\|w_{\#K,t}^{(k)}\|^2$ $(K=10)$.

Figure 2: Subfigure (a) and (b): parallel component $|w_{\#K,t}^{(k),\|}|$ of each neuron k vs. iteration t for student networks with size $K=\{3,10\}$ trained by gradient descent. The curve "l2 norm" represents $\sqrt{\sum_{k=1}^{K}|w_{\#K,t}^{(k),\|}|^2}$, which can be viewed as the aggregate projection of a student network $W\in\mathbb{R}^{d\times K}$ on the teacher neuron $w_*\in\mathbb{R}^d$. For comparison, we also plot the curve for $|w_{\#1,t}^{(1),\|}|$ labeled by $K=1$ on the same subfigures. Subfigure (c) and (d): the square norm $\|w_{\#K,t}^{(k)}\|^2$ of each neuron k for different K's. The curve "sum" represents $\|W^{\#K}\|_F^2$. For comparison, we also plot the curve for $\|w_{\#1,t}^{(1)}\|^2$ labeled by $K=1$ on the subfigures.

4 Analysis

In this section, we answer why over-parametrization leads to the acceleration. We first show that gradient descent (GD) exhibits linear convergence to a global optimal solution when $\text{dist}(W^{\#K},w_*)$ is small. We then answer why over-parametrization helps the iterate to enter the neighborhood of a global optimal solution faster. For the ease of analysis, we assume that gradient descent uses population gradient (6) for the update and we denote $\nabla F(W):=\mathbb{E}_x[\nabla f(W)]$ and $\nabla^2 F(W):=\mathbb{E}_x[\nabla^2 f(W)]$ accordingly.

4.1 When does the iterate generated by gradient descent enter a benign region?

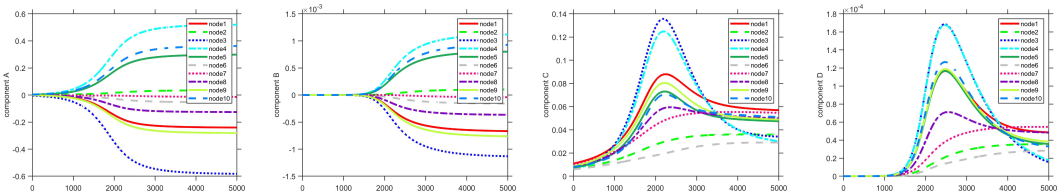
We first introduce a key lemma. The lemma shows that whenever the iterate is in the neighborhood of a global optimal solution, gradient descent has a linear convergence rate.

Lemma 1. *(locally linear convergence) Suppose that at time t_0, $\text{dist}(W_{t_0},w_*):=\|W_{t_0}-w_*q_{t_0}^\top\|\le\nu\|w_*\|$ where $W_{t_0}\in\mathbb{R}^{d\times K}$, $\nu>0$ satisfies $2-14\nu-2\nu^2>0$, and $q_{t_0}:=\arg\min_{q\in\mathbb{R}^K:\|q\|_2\le 1}\|W_{t_0}-w_*q^\top\|$. Then, gradient descent with the step size $\eta\le\frac{2-14\nu-2\nu^2}{(13+16\nu^2)^2\|w_*\|^4}$ generates iterates $\{W_t\}_{t\ge t_0}$ satisfying $\text{dist}^2(W_{t+1},w_*)\le\big(1-\eta(2-14\nu-2\nu^2)\big)\text{dist}^2(W_t,w_*)$.*

The proof is available in Appendix B. Note that the lemma holds for *any* size K of neural nets, as long as the condition, $\text{dist}(W,w_*)\le\nu\|w_*\|_2$ with the required ν, is satisfied. The condition ensures the locally linear convergence of gradient descent and might be easily satisfied by having a larger number of student neurons $K>1$. Specifically, each neuron $w^{(k)}\in\mathbb{R}^d$ of $W\in\mathbb{R}^{d\times K}$ only needs to have a smaller component on the direction of w_* in order to have W be sufficiently close to the teacher neuron w_* up to a transform $q^\top$, compared to the case when one only has a single student neuron $(K=1)$. To support this argument, we plot the quantities $|w_{\#K,t}^{(k),\|}|=|\langle w_{\#K,t}^{(k)},w_*\rangle|$ for each $k\in[K]$ of different student networks trained by gradient descent on Figure 2. The figure shows that with more student neurons K, each $w^{(k)}$ only needs a smaller projection on w_* for the aggregate projection $\sqrt{\sum_{k=1}^{K}|w_{\#K,t}^{(k),\|}|^2}$ to achieve certain level. In the later subsections, we will provide a formal analysis.

4.2 How does gradient descent work for the student network consists of a single neuron?

In this subsection, we analyze the case that the student network only has a single neuron. For brevity, we suppress the notation $\#1$ in the following. Define $T_\gamma:=\min\{t:||w_t[1]|-\|w_*\||\le\gamma$ and $\|w_t^\perp\|\le\gamma\}$. Note that if $||w_t[1]|-\|w_*\||\le\gamma$ and that $\|w_t^\perp\|\le\gamma$, then $\text{dist}^2(w_t,w_*)=$



(a) Component A vs. t. (b) Component B vs. t. (c) Component C vs. t. (d) Component D vs. t.

Figure 3: Subfigure (a) shows $w_t^{(k),\|}\big(1+\eta(3\|w_*\|^2-3\|w_t^{(k)}\|^2)\big)$ (i.e. component A of $w_{\#K,t}^{(k),\|}$) versus iteration t for each neuron k. Subfigure (b) plots $2\eta\sum_{j\neq k}^{K}((w_t^{(j)})^\top w_t^{(k)})w_t^{(j),\|}+\eta w_t^{(k),\|}\sum_{j\neq k}^{K}\|w_t^{(j)}\|^2$ (i.e. component B of $w_{\#K,t}^{(k),\|}$) versus iteration t for each neuron k. Subfigure (c) plots the norm of component C of $w_{\#K,t}^{(k),\perp}$, while subfigure (d) plots the norm of component D of $w_{\#K,t}^{(k),\perp}$ for each neuron k. The empirical findings show that the components due to interaction of the other neurons (i.e. component B and D) are small (notice that the scale of the vertical axis of (a) and (b), (c) and (d) are different) compared to their counterparts (i.e. component A and C respectively), which suggests that $\theta, \vartheta \approxeq 10^{-4}$ on (10) empirically. Similar patterns exhibit in training under different K's (Appendix F).

$||w_t[1]|-\|w_*\||^2+\|w_t^\perp\|_2^2 \leq 2\gamma^2$. Let us assume that (C1) $\|w_*\| > 1.1\gamma$ (strong signal) and (C2) $\gamma \geq 10\|w_0\|$ (small initialization). Note that to invoke Lemma 1 for showing locally linear convergence after the iterate gets into a benign region, we will set $2\gamma^2=\nu^2\|w_*\|^2$ with ν satisfying $2-14\nu-2\nu^2>0$ (i.e. $\nu\leq 0.141$); consequently, (C1) is trivially satisfied.

Theorem 1. *Suppose that the conditions (C1-C2) hold. Assume that the step size satisfies $\eta \leq c/\|w_*\|^2$ for some sufficiently small constant $c>0$. Then gradient descent for problem (1) (i.e. $K=1$) has $T_\gamma \leq \frac{\log(\frac{\|w_*\|-\gamma}{|w_0[1]|})}{\log(1+\eta\Delta)}$, where $\Delta := 6\gamma(\|w_*\|-\gamma)>0$. Furthermore, for $0\leq t\leq T_\gamma$, we have that $|w_t^{\|}| \geq (1+\eta\Delta)^t|w_0^{\|}|$ and $\|w_t^\perp\|\leq\gamma$.*

Theorem 1 states that to achieve $\text{dist}^2(w_t,w_*) \leq 2\gamma^2$, gradient descent only needs at most $\log(\frac{\|w_*\|-\gamma}{|w_0[1]|})/\log(1+\eta\Delta)$ number of iterations. Furthermore, on the signal direction, $|w_t[1]|$ (and hence $w_t^{\|}$) grows exponentially at a rate at least $1+\eta\Delta$ before reaching at $\|w_*\|-\gamma$. On the other hand, by a close-to-zero initialization (C2), the perpendicular component $\|w^\perp\|$ remains small. Consequently, we have that $\text{dist}^2(W_t^{\#1},w_*)\leq\max(\gamma^2,\big(|w_*|-|w_0[1]|(1+\eta\Delta)^t\big)^2)+\gamma^2$, for $0\leq t\leq T_\gamma$ before gradient descent enters the linear convergence regime. The proof of Theorem 1 is available in Appendix C. A similar result as Theorem 1 was shown in (Chen et al., 2019).

4.3 How does over-parametrization help entering a benign region?

Let us begin by providing a more detailed observation regarding the dynamics of gradient descent. Figure 3 plots each component of $w_{\#K,t}^{(k),\|}$ and $w_{\#K,t}^{(k),\perp}$ in the gradient descent dynamics (8) for $K=10$. The empirical findings show that the component due to the interaction (component B, or component D respectively) is negligible compared to the component that is without the dependency on the other neurons (component A, or component C respectively). Based on the observation, we can re-write the population dynamics (8) in the early stage (i.e. before gradient descent enters the linear convergence regime) as follows,

$$\begin{aligned}|w_{\#K,t+1}^{(k),\|}| &\geq (1-\theta)|w_{\#K,t}^{(k),\|}|\big(1+\eta(3\|w_*\|^2-3\|w_{\#K,t}^{(k)}\|^2)\big)\\ \|w_{\#K,t+1}^{(k),\perp}\| &\leq (1+\vartheta)\|w_{\#K,t}^{(k),\perp}\|\big(1+\eta(\|w_*\|^2-3\|w_{\#K,t}^{(k)}\|^2)\big).\end{aligned} \tag{10}$$

for some small numbers $\theta,\vartheta \ll 1$ (empirically $\approxeq 10^{-4}$ as Figure 3 shows). That is, in the early stage, the approximated dynamics (10) of each student neuron of an over-parametrized network does not deviate too much from the original dynamics without over-parametrization (7). The approximated dynamics (10) will be only used for analyzing the stage before the iterate enters the benign region, i.e. used only before entering a linear convergence regime, though the emprical findings (Figure 3) suggest that the approximated dynamics hold all the time.

On the other hand, by comparing subfigure (a) and subfigure (b) of Figure 3, we see that component A and B of each neuron k are with the same sign during the execution. This observation together with the dynamics (8) tend to implies that $|w_{\#K,t+1}^{(k),\|}| \leq |w_{\#K,t}^{(k),\|}|\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#K,t}^{(k)}\|^2)\big)$ when $K > 1$. On the other hand, the dynamics (7) has $|w_{\#1,t+1}^{(1),\|}| = |w_{\#1,t}^{(1),\|}|\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#1,t}^{(1)}\|^2)\big)$. Also, as the case of single neuron, the perpendicular component of each neuron k of K (i.e. $\|w_{\#K,t}^{(k),\perp}\|$) remains small (figures in Appendix F). Consequently, we could write

$$|w_{\#1,t}^{(1),\|}| \gtrsim |w_{\#K,t}^{(k),\|}| \text{ and } \|w_{\#1,t}^{(1)}\| \gtrsim \|w_{\#K,t}^{(k)}\|, \tag{11}$$

where the approximation $\gtrsim$ accounts for the fact that the size of initial points due to the random initialization may be different and that the small interaction components are present in the dynamics for $K > 1$. Figure 2 confirms that the relation (11) generally holds on average empirically.

Lemma 2. *Suppose that the approximated dynamics (10) and (11) hold from iteration* 0 *to iteration* t*. Then, the network with a single neuron and an over-parametrized network trained by GD with the same step size* η *has* $\sqrt{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2} \gtrsim \sqrt{(1-2t\theta)}\sqrt{K}|w_{\#1,t}^{(1),\|}|$.

Lemma 2 states that if the single neuron of the non-overparametrized network has a certain projection on w_* at time t, then the over-parametrized network with K neurons will have approximately $\sqrt{K}$ times larger projection on w_*, modulo the $\sqrt{1-2t\theta}$ factor which is close to 1 if the product $t\theta$ is small (as Figure 3 shows). This demonstrates the advantage of over-parametrization — over-parametrization helps to make more progress on growing the model's projection on w_*.

Lemma 3. *Suppose that* $\eta \leq \frac{1}{3\|w_*\|^2}$*. By following the conditions as Lemma 2, we have that* $\|w_{\#K,t}^{(k),\perp}\| \lesssim \frac{|w_{\#K,t}^{(k),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t} \lesssim \frac{|w_{\#1,t}^{(1),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t}$*, where* $\psi := \big(1-\theta-\vartheta-\theta\vartheta\big)\big(1+\eta\|w_*\|^2\big)$.

Lemma 3 states that the ratio of the perpendicular component $\|w_{\#K,t}^{(k),\perp}\|$ to the parallel component $|w_{\#K,t}^{(k),\|}|$ of each neuron decays exponentially if $\psi > 1$ (which holds if $\eta\|w_*\|^2 \gtrsim \frac{\theta+\vartheta}{1-\theta-\vartheta}$). By combining Lemma 2 and 3, we have the following theorem, which characterizes the difference of the distances to w_* at iteration t.

Theorem 2. *(Snapshot at* t*) Suppose that the approximated dynamics (10) and (11) hold from* 0 *to* t *and that at iteration* t*, the student network with a single neuron trained by GD with the step size* η *has* $|w_{\#1,t}^{(1),\|}| = c_{1,t}\|w_*\|^2$ *for some number* $c_{1,t}$ *satisfying* $1 > c_{1,t} > 0$*. Denote* $c_{2,t}$ *a number that satisfies* $\frac{c_{1,t}\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k)}[1]|}\frac{1}{\psi^t} \leq c_{2,t}$ *for each* $k \in [K]$*. Suppose that the step size* η *also satisfies* $\eta \leq \frac{1}{3\|w_*\|^2}$ *and makes* $\psi := \big(1-\theta-\vartheta-\theta\vartheta\big)\big(1+\eta\|w_*\|^2\big) > 1$*. Then, an over-parametrized network* $W_t^{\#K}$ *trained by GD with the same* η *has*

$$\text{dist}^2(W_t^{\#1}, w_*) - \text{dist}^2(W_t^{\#K}, w_*) \gtrsim \|w_*\|^2\big(2c_{1,t}(\sqrt{(1-2t\theta)}\sqrt{K}-1) - K(c_{1,t}^2+c_{2,t}^2) + c_{1,t}^2\big).$$

Recall that in Theorem 1, we upper-bound $\text{dist}^2(W_t^{\#1}, w_*)$. Simply combining Theorem 1 and 2 leads to a distance upper-bound of $\text{dist}^2(W_t^{\#K}, w_*)$ at certain iteration t. The lower bound of the difference of the distances in Theorem 2 shows a strict improvement due to over-parametrization when it is positive, which answers why over-parametrization helps gradient descent to enter the linear convergence regime faster — gradient descent for an over-parametrized network shrinks the distance to w_* faster in the early stage. Note that the lower bound is a quadratic function of $\sqrt{K}$ and is increasing for $1 \leq \sqrt{K} \leq \frac{c_{1,t}\sqrt{1-2t\theta}}{c_{1,t}^2+c_{2,t}^2}$, which means that up to a certain threshold of K, more over-parametrization could lead to more improvements. Moreover, if $c_{1,t}$ and $c_{2,t}$ further satisfy $(*)$ $2c_{1,t}\big(\sqrt{2}\sqrt{(1-2t\theta)}-1\big) - c_{1,t}^2 - 2c_{2,t}^2 > 0$, then the lower bound of the difference for $K = 2$ neurons is strictly positive and keeps being positive up to $\sqrt{K} \leq \lfloor \frac{c_{1,t}\sqrt{1-2t\theta}+\sqrt{c_{1,t}^2(1-2t\theta)-(c_{1,t}^2+c_{2,t}^2)(2c_{1,t}-c_{1,t}^2)}}{c_{1,t}^2+c_{2,t}^2} \rfloor$, which gives an upper limit of the degree of over-parametrization that allows acceleration. The condition (*) is easily satisfied when (1) $t\theta \ll 1$ so that $\sqrt{(1-2t\theta)} \approxeq 1$ and (2) $c_{2,t} \ll 1$, which happens when the approximated dynamics (8) holds for a small θ and that the ratio of the perpendicular component to the parallel component of neuron k decays sufficiently fast (Lemma 3).

4.4 IS OVER-PARAMETRIZATION NECESSARY FOR ACCELERATION?

In this subsection, let us consider a new objective,

$$\min_{w\in\mathbb{R}^d} f(w) := \frac{1}{4n}\sum_{i=1}^n \left(C(x_i^\top w)^2 - y_i\right)^2, \tag{12}$$

where $C \geq 1$ is an user-defined parameter. This modification can be viewed as setting the weight of the second layer to C instead of 1. The population dynamics due to gradient descent becomes $w_{t+1}^{(1),\|} = w_t^{(1),\|}\big(1 + \eta(3C\|w_*\|^2 - 3C^2\|w_t^{(1)}\|^2)\big)$ and $w_{t+1}^{(1),\perp} = w_t^{(1),\perp}\big(1 + \eta(C\|w_*\|^2 - 3C^2\|w_t^{(1)}\|^2)\big)$. Let us also consider an over-parametrized version of (12),

$$\min_{W\in\mathbb{R}^{d\times K}} f(W) := \frac{1}{4n}\sum_{i=1}^n \left(C(x_i^\top w^{(1)})^2 + C(x_i^\top w^{(2)})^2 + ... + C(x_i^\top w^{(K)})^2 - y_i\right)^2. \tag{13}$$

On Figure 4, we report gradient descent with the same step size η for solving (13) under different C's and K's. The result is interesting; it shows that simply scaling the weight of the single neuron (i.e. $C > 1$) without over-parametrization (i.e. $K = 1$) can achieve a similar acceleration and that the effect seems to be more significant. The observation questions the necessity of over-parametrization, as the figure suggests that one can avoid the computational overhead due to over-parametrization by simply scaling up the weight of the output node while enjoys acceleration.

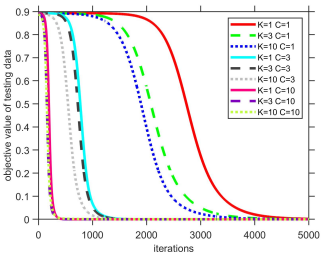


Figure 4: Gradient descent for (13) under different K and C

Let us provide an analysis of the acceleration observed here. Since $w = \pm w_*/\sqrt{C}$ are the global solutions of (12), we define $T_{\gamma,C} := \min\{t : ||w_t[1]| - \|w_*\|/\sqrt{C}| \leq \gamma \text{ and } \|w_t^\perp\| \leq \gamma\}$. For γ being sufficiently small, we can show that gradient descent enters the linear convergence regime at $T_{\gamma,C}$ (see Subsection E.1 in Appendix E, where we provide a counterpart of Lemma 1 for general C). Therefore, the question is how the scaling of the output weight helps to reduce the number of iterations spent in the early stage. We have the following theorem. The proof is in Appendix E.

Theorem 3. *Suppose that $\|w_*\| > 1.1\sqrt{C}\gamma$ and $\gamma \geq 10\|w_0\|$. Assume that the step size satisfies $\eta \leq c/\|w_*\|^2$ for some sufficiently small constant $c > 0$. Then gradient descent for problem (12) has $T_{\gamma,C} \leq \frac{\log(\frac{\|w_*\|/\sqrt{C}-\gamma}{|w_0[1]|})}{\log(1+\eta\Delta_C)}$, where $\Delta_C := 6\gamma C(\sqrt{C}\|w_*\| - C\gamma) > 0$. Furthermore, for $0 \leq t \leq T_{\gamma,C}$, we have that $|w_t^{\|}| \geq (1+\eta\Delta_C)^t|w_0^{\|}|$ and $\|w_t^\perp\| \leq \gamma$.*

As Theorem 1, the condition $\|w_*\| > 1.1\sqrt{C}\gamma$ (strong signal) in the theorem is trivially satisfied when we set $2\gamma^2 = \nu^2\|w_*\|^2$ and makes $\nu > 0$ a small number (see Appendix E.1 for details). Now let us conclude by making two important remarks regarding the acceleration effect due to scaling up the output weight. First, the single neuron only needs to be $w_*/\sqrt{C}$ for achieving zero testing error. Since we have a close-to-zero random initialization here, it means that $|w_0^{(1)}|$ only needs to grow from a close-to-zero number to $\|w_*\|/\sqrt{C}$ instead of $\|w_*\|$. This implies that the modification shortens the *effective distance* to a global optimal solution. Second, up to a certain threshold C, the growth rate of the parallel component, $(1+\eta\Delta_C)$, is larger for a larger $C \geq 1$. To see this, fix a $\|w_*\|$ and γ, one can show that the derivative of Δ_C w.r.t. C is non-negative as long as $3\|w_*\|/(4\gamma) \geq \sqrt{C}$. The means that the *speed* to get into the neighborhood of a global optimal solution is faster as C increases. These two mechanisms explain why scaling up the weight of the output node helps reducing the number of iterations $T_{\gamma,C}$ for gradient descent entering the linear rate regime. Our findings thus show that the scaling of the output layer has a great impact on the speed of convergence. Nevertheless, we notice that some works in the literature suggest leaving the output layer (the last layer) untrained (e.g. (Hoffer et al., 2018)) and that most of the theory works regarding convergence results of GD in deep learning assume that the output layer is fixed (e.g. (Allen-Zhu et al., 2019; Du et al., 2019)). Hence, our results here might raise another important issue in practice.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, , and Zhao Song. A convergence theory for deep learning via overparameterization. *ICML*, 2019.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *ICML*, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *NerurIPS*, 2019.

Alon Brutzkus and Amir Globerson. Why do larger models generalize better? a theoretical perspective via the xor problem. *ICML*, 2019.

Emmanuel J. Candés and Xiaodong Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 2014.

Emmanuel J. Candés, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 2013.

Emmanuel J. Candés, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 2015.

Yuxin Chen and Emmanuel J. Candés. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 2017.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 2019.

Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal on Numerical Analysis*, 2018.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. *ICML*, 2018.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *ICLR*, 2019.

Matt Emschwiller, David Gamarnik, Eren C. Kızıldag, and Ilias Zadik. Neural networks and polynomial regression. demystifying the overparametrization phenomena. *arXiv:2003.10523*, 2020.

Albert Fannjiang and Thomas Strohmer. The numerics of phase retrieval. *Acta Numerica*, 2020.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. *arXiv:2005.14426*, 2020.

David Gamarnik, Eren C. Kızıldag, and Ilias Zadik. Stationary points of shallow neural networks with quadratic activation function. *arXiv:1912.01599*, 2019.

Rong Ge, Runzhe Wang, and Haoyu Zhao. Mildly overparametrized neural nets can memorize training data efficiently. *arXiv:1909.11837*, 2019.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *NeurIPS*, 2019.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. *COLT*, 2017.

Surbhi Goel, Adam Klivans, and Raghu Meka. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. *NeurIPS*, 2019.

Sebastian Goldt, Madhu Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborova. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *NeurIPS*, 2019.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. *NIPS*, 2017.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *NIPS*, 2017.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. *ICLR*, 2018.

Sham M. Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression? *NeurIPS*, 2011.

Seyed Mohammadreza Mousavi Kalan, Mahdi Soltanolkotabi, and A. Salman Avestimehr. Fitting relus via sgd and quantized sgd. *ISIT*, 2019.

Abbas Kazemipour, Brett Larsen, and Shaul Druckmann. No spurious local minima in deep quadratic networks. *arXiv:2001.00098*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. *AISTATS*, 2019.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *NeurIPS*, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *COLT*, 2018.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. *NeurIPS*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. Aggregated momentum: Stability through passive damping. *ICLR*, 2019.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 2017.

Stefano Sarao Mannelia, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. *arXiv:2006.15459*, 2020.

Stefano Sarao Mannellia, Giulio Birolib, Chiara Cammarotac, Florent Krzakalab, Pierfrancesco Urbania, and Lenka Zdeborová. Complex dynamics in simple neural networks: Understanding gradient flow in phase retrieval. *arXiv:2006.06997*, 2020.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *PNAS*, 2018.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *ICML*, 2017.

Itay Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. *arXiv:2006.01005*, 2020.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 2015.

Mahdi Soltanolkotabi. Algorithms and theory for clustering and nonconvex quadratic programming. *Stanford University Ph. D. Dissertation*, 2014.

Mahdi Soltanolkotabi. Learning relus via gradient descent. *NeurIPS*, 2017.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *IEEE ISIT*, 2016.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. *ICML*, 2013.

Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized kaczmarz: Theoretical guarantees. *Information and Inference*, 2018.

Yuandong Tian. Student specialization in deep rectified networks with finite width and input dimension. *ICML*, 2020.

Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *JMLR*, 2019.

Chris D. White, Sujay Sanghavi, and Rachel Ward. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 2016.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. *arXiv:2001.05205*, 2020.

Chong You, Zhihui Zhu, Qing Qu, and Yi Ma. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. *arXiv:2006.08857*, 2020.

A Related works

Over-parametrization: Though our work focus on understanding why over-parametrization leads to acceleration in optimization (i.e. improving the convergence time), we also want to acknowledge some related works of understanding over-parametrization in different aspects (e.g. (Arora et al., 2019; Goldt et al., 2019; Brutzkus & Globerson, 2019; Emschwiller et al., 2020; Tian, 2020; Safran et al., 2020). Arora et al. (2019) study implicit regularization of gradient descent in deep linear neural network for matrix factorization. They consider increasing the *depth* of the linear network and analyze how the singular values of the learned solution evolves during gradient flow due to increasing the depth. Goldt et al. (2019) use ordinary different equation as a tools for analyzing the asymptotic generalization error of SGD when the dimension of input data goes to infinity, for the case that the number of hidden nodes of a student network is larger than that of a teacher network. Brutzkus & Globerson (2019) prove that for an over-parameterized convolutional networks, gradient descent converges to a global solution with better generalization performance compared to global minima of smaller networks. Emschwiller et al. (2020) show that for polynomial activation function, if a trained student network interpolates sufficient number of training data, then the generalization error can be arbitrarily small, regardless of the size of the student network. Tian (2020) identify conditions of which each neuron of a teacher network can be *explained* by some neurons of a larger student network. Safran et al. (2020) study the effects of over-parameterization on the optimization landscape of a network with ReLU activation and show that there is a significant set of points around a global solution that satisfies a notion of one-point strong convexity for an over-parametrized model.

On the other hand, there is a trend of works studying how over-parametrization changes the optimization landscape of empirical risk minimization for neural nets with quadratic activation. (e.g. (Du & Lee, 2018; Gamarnik et al., 2019; Ge et al., 2019; Kazemipour et al., 2019; Soltanolkotabi et al., 2018; Nguyen & Hein, 2017; Venturi et al., 2019; Mannelia et al., 2020)). Du & Lee (2018) show that when the number of neurons is sufficiently larger than the squared root of the number of samples, then there is no spurious minima. Mannelia et al. (2020) study how the number of samples and the width of the teacher network affect the optimization landscape of empirical risk minimization. They also identify some conditions so that gradient flow can have small generalization error.

Quadratic activation: The optimization landscape of problem (1) (i.e. phase retrieval) and its variants has been studied by (Davis et al., 2018; Soltanolkotabi, 2014; Sun et al., 2016; White et al., 2016), which shows that as long as the number of samples is sufficiently large, it has no spurious

local optimum and all the local optima are globally optimal. In addition to the landscape analysis, there are some convergence results of GD in the literature (e.g. (Chen et al., 2019; Ge et al., 2019)). Ge et al. (2019) show that gradient descent can train a quadratic network to memorize training data perfectly in a mildly over-parametrized regime. Chen et al. (2019) provably show that applying gradient descent with an isotropic random initialization to solving (1) leads to a solution that recovers the teacher neuron w_* modulo the unrecoverable sign. Other relate works include (Tan & Vershynin, 2018) and (Mannellia et al., 2020). Tan & Vershynin (2018) study online gradient descent while Mannellia et al. (2020) study gradient flow for solving phase retrieval. In this work, we show that gradient descent on a over-parametrized student network (namely, (2)) takes even fewer iterations to recover w_*. Furthermore, we show that up to a certain threshold, the larger the number of student neuron K, the faster the global convergence.

Matrix Sensing: In matrix sensing, the setup is that there is an unknown rank-r PSD matrix $W_* \in \mathbb{R}^{d\times d}$ such that $y_i = \langle A_i, W_*\rangle = \text{tr}(A_i^\top W_*)$, where $A_i \in \mathbb{R}^{d\times d}$ is a symmetric measurement matrix. Li et al. (2018) shows that gradient descent on $\min_{U\in\mathbb{R}^{d\times d}} \frac{1}{4n}\sum_{i=1}^n \left(y_i - \langle A_i, UU^\top\rangle\right)^2$ with a close-to-zero random initialization recovers W_*, when the matrices $\{A_i\}$ satisfy restricted isometry property (RIP). Their work shows an implicit regularization of gradient descent under over-parametrization, since the result indicates that gradient descent recovers W_* without knowing the matrix rank beforehand. Li et al. (2019), another group of authors, consider the case that $A_i = x_i x_i^\top$, $x_i \sim N(0, I_d)$, and W_* is an unknown rank-r PSD matrix. They show that with a specialized initialization called spectral initialization, gradient descent solves $\min_{U\in\mathbb{R}^{d\times r}} \frac{1}{4n}\sum_{i=1}^n \left(y_i - \langle A_i, UU^\top\rangle\right)^2$ and recovers W_* up to some orthonormal transform. Note that one can rewrite (2) in the form of matrix sensing, namely, $\min_{W\in\mathbb{R}^{d\times K}} \frac{1}{4n}\sum_{i=1}^n \left(y_i - \langle x_i x_i^\top, WW^\top\rangle\right)^2$, $W_* = w_* w_*^\top$, and $y_i = \langle x_i x_i^\top, W_*\rangle = (w_*^\top x)^2$. Our work does not assume RIP property nor require the spectral initialization. Furthermore, the goal of our work is different from these works; we focus on understanding the acceleration due to over-parametrization. Lastly, we note that matrix sensing has been a subject of understanding implicit regularization of gradient descent (see e.g. Gunasekar et al. (2017); Gidel et al. (2019); You et al. (2020)), where the goal is to understand which solution gradient descent will converge to.

Two-phase dynamics of gradient descent: The behavior of gradient descent in our problem can be divided into two phases. We note that the multi-phase of dynamics of gradient descent is also present for solving other problems (see e.g. (Li & Yuan, 2017)).

B Proof of Lemma 1

Recall the optimization problem is

$$\min_{W\in\mathbb{R}^{d\times K}} f(W) := \frac{1}{4n}\sum_{i=1}^n \left((x_i^\top w^{(1)})^2 + (x_i^\top w^{(2)})^2 + ... + (x_i^\top w^{(K)})^2 - y_i\right)^2,$$

and the notation that $\nabla F(W) := \mathbb{E}_x[\nabla f(W)]$ and $\nabla^2 F(W) := \mathbb{E}_x[\nabla^2 f(W)]$. We first introduce some key lemmas. Lemma 4 below says that for any $W \in \mathbb{R}^{d\times K}$, if the iterate is sufficiently close to a global solution $w_* q^\top$, then the landscape is essentially strongly convex. Lemma 5, on the other hand, shows the smoothness of the landscape.

Lemma 4. *(locally strong convexity) Assume that* $\text{dist}(W_{t_0}, w_*) := \|W_{t_0} - w_* q_{t_0}^\top\|_F \le \nu\|w_*\|_2$, *where* $W_{t_0} \in \mathbb{R}^{d\times K}$, $\nu > 0$, $q_{t_0} := \arg\min_{q\in\mathbb{R}^K:\|q\|_2\le 1} \|W_{t_0} - w_* q^\top\|_F$, *and* $w_* \in \mathbb{R}^d$ *being the teacher neuron. Then*

$$vec(V)^\top \nabla^2 F(W_{t_0}) vec(V) \ge 2tr(q_{t_0} w_*^\top V q_{t_0} w_*^\top V) + (2 - 14\nu - 2\nu^2)\|w_*\|_2^2 \|V\|_F^2$$

for any $V \in \mathbb{R}^{d\times K}$, *where* $tr(\cdot)$ *denotes the matrix trace.*

Proof. The proof is a modification of the proof of Lemma 14 in Li et al. (2019). For notation brevity, we will suppress script t_0 in the following (i.e. $W \leftarrow W_{t_0}$, $q \leftarrow q_{t_0}$). We have that

$$\begin{aligned}
&\mathrm{vec}(V)^\top \nabla^2 f(W) \mathrm{vec}(V) \\
&= \frac{1}{m}\sum_{i=1}^m \mathrm{vec}(V)^\top \big[\big((\|x_i^\top W\|_2^2 - y_i) I_K + 2W^\top x_i x_i^\top W\big) \otimes x_i x_i^\top\big] \mathrm{vec}(V) \\
&= \frac{1}{m}\sum_{i=1}^m \big(\|x_i^\top W\|_2^2 - y_i\big) \mathrm{vec}(V)^\top \mathrm{vec}(x_i x_i^\top V) + \frac{1}{m}\sum_{i=1}^m \mathrm{vec}(V)^\top \mathrm{vec}(2 x_i x_i^\top V W^\top x_i x_i^\top W). \\
&= \frac{1}{m}\sum_{i=1}^m \big[\big(\|x_i^\top W\|_2^2 - \|x_i^\top w_* q^\top\|_2^2\big)\|x_i^\top V\|_2^2 + 2\big(x_i^\top W V^\top x_i\big)^2\big].
\end{aligned}$$

By taking the expectation above over $x_i \sim N(0, I_d)$ and using Lemma 6, we have that

$$\begin{aligned}
\mathbb{E}[\mathrm{vec}(V)^\top \nabla^2 f(W) \mathrm{vec}(V)] = \|W\|_F^2\|V\|_F^2 + 2\|V^\top W\|_F^2 - \big(\|w_* q^\top\|_F^2\|V\|_F^2 + 2\|V^\top w_* q^\top\|_F^2\big) \\
+ 2\big(\mathrm{tr}(W^\top V)^2 + \mathrm{tr}(W^\top V W^\top V) + \|W V^\top\|_F^2\big).
\end{aligned} \tag{14}$$

Now let us set $W = w_* q^\top + \theta H$ with a H satisfying $\|H\|_F = 1$ and $\theta := \nu\|w_*\|$ for some number $\nu > 0$.

$$\begin{aligned}
\|W\|_F^2 &= \|w_* q^\top\|_F^2 + \theta^2\|H\|_F^2 + 2\theta\langle w_* q^\top, H\rangle \\
&\geq \|w_*\|^2 - 2\theta\|w_* q^\top\|_F\|H\|_F \\
&\geq \|w_*\|^2 - 2\theta\|w_*\|\|H\|_F \\
\|V^\top W\|_F^2 &= \|V^\top w_* q^\top\|_F^2 + \theta^2\|V^\top H\|_F^2 + 2\theta\mathrm{tr}(V^\top w_* q^\top H^\top V) \\
&\geq \|V^\top w_* q^\top\|_F^2 - 2\theta\|w_* q^\top\|\|H\|_F\|V\|_F^2 \\
\|V W^\top\|_F^2 &= \|V q w_*^\top\|_F^2 + \theta^2\|V H^\top\|_F^2 + 2\theta\mathrm{tr}(V w_*^\top q H V^\top) \\
&\geq \|V q w_*^\top\|_F^2 - 2\theta\|w_*^\top q\|\|H\|_F\|V\|_F^2 \\
&= \|V q w_*^\top\|_F^2 - 2\theta\|w_*\|\|H\|_F\|V\|_F^2 \\
\mathrm{tr}(W^\top V W^\top V) &= \mathrm{tr}(q w_*^\top V q w_*^\top V) + 2\theta\mathrm{tr}(H^\top V q w_*^\top V) + \theta^2\mathrm{tr}(H^\top V H^\top V) \\
&= \mathrm{tr}(q w_*^\top V q w_*^\top V) - 2\theta\|w_* q^\top\|\|H\|_F\|V\|_F^2 - \theta^2\|H\|_F^2\|V\|_F^2.
\end{aligned} \tag{15}$$

Combining (14) and (15), together with the bilinear property of the expectation so that $\mathrm{vec}(V)^\top \nabla^2 F(w) \mathrm{vec}(V) = \mathbb{E}[\mathrm{vec}(V)^\top \nabla^2 f(W) \mathrm{vec}(V)]$, we have that

$$\begin{aligned}
\mathrm{vec}(V)^\top \nabla^2 F(W) \mathrm{vec}(V) \geq& 2\mathrm{tr}(q w_*^\top V q w_*^\top V) + 2\|V q w_*^\top\|_F^2 \\
&- 14\theta\|w_*\|\|H\|_F\|V\|_F^2 - 2\theta^2\|H\|_F^2\|V\|_F^2 \\
\overset{(a)}{=}& 2\mathrm{tr}(q w_*^\top V q w_*^\top V) + 2\|w_*\|^2\|V\|_F^2 \\
&- 14\theta\|w_*\|\|H\|_F\|V\|_F^2 - 2\theta^2\|H\|_F^2\|V\|_F^2 \\
=& 2\mathrm{tr}(q w_*^\top V q w_*^\top V) + (2 - 14\nu - 2\nu^2)\|w_*\|^2\|V\|_F^2,
\end{aligned} \tag{16}$$

where (a) is due to that $\|AB\|_F^2 \geq \sigma_r^2(A)\|B\|_F^2$ with r being the rank of A, which in our case $A := w_* q^\top$ is a rank one matrix and $B := V^\top$; consequently $\sigma_1^2(A) = \|w_* q^\top\|_F^2 = \|w_*\|^2$. □

Lemma 5. *(smoothness) Assume that* $\mathrm{dist}(W_{t_0}, w_*) := \|W_{t_0} - w_* q_{t_0}^\top\|_F \leq \nu\|w_*\|_2$ *where* $W_{t_0} \in \mathbb{R}^{d\times K}$, $\nu > 0$, $q_{t_0} := \arg\min_{q\in\mathbb{R}^K:\|q\|_2\leq 1}\|W - w_* q^\top\|_F$, *and* $w_* \in \mathbb{R}^d$ *being the teacher neuron. Then,* $\|\nabla^2 F(W)\|_2 \leq (15 + 16\nu^2)\|w_*\|^2$.

Proof. For notation brevity, we will suppress script t_0 in the following (i.e. $W \leftarrow W_{t_0}$, $q \leftarrow q_{t_0}$). We have that

$$\begin{aligned}
\|\nabla^2 F(W)\|_2 &= \|\mathbb{E}\big[[\big(\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2\big) I_K + 2W^\top x x^\top W] \otimes x x^\top\big]\|_2 \\
&\le \|\mathbb{E}\big[[\big|\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2\big| I_K + 2\|x^\top W\|_2^2 I_K] \otimes x x^\top\big]\|_2 \\
&\overset{(a)}{\le} \|\mathbb{E}\big[\big|\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2\big| x x^\top\big]\|_2 + 2\|\mathbb{E}\big[\|x^\top W\|_2^2 x x^\top\big]\|_2 \\
&\overset{(b)}{=} \|\mathbb{E}\big[(\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2) x x^\top\big]\|_2 + 2\|\|W\|_F^2 I_d + 2WW^\top\|_2 \\
&\overset{(c)}{=} \|(\|W\|_F^2 - \|w_* q^\top\|_F^2) I_d + 2(WW^\top - w_* w_*^\top)\|_2 + 2\|\|W\|_F^2 I_d + 2WW^\top\|_2 \\
&\le 7\|W\|_F^2 + \|w_* q^\top\|_F^2 + 2\|WW^\top - w_* w_*^\top\|_2 \\
&\le (15 + 16\nu^2)\|w_*\|^2,
\end{aligned} \tag{17}$$

where (a) is due to $\|I \otimes A\|_2 \le \|I\|_2 \|A\|_2 = \|A\|_2$, (b,c) is due to Lemma 6, and the last inequality is by setting $W = w_* q^\top + \theta H$ with a H satisfying $\|H\|_F = 1$ and $\theta := \nu\|w_*\|$.

□

Lemma 4 and Lemma 5 together implies that when the the iterate is in a neighborhood of a global optimal solution, then gradient descent has a linear convergence rate.

Lemma 1: *(locally linear convergence) Suppose that at time t_0, $\text{dist}(W_{t_0}, w_*) := \|W_{t_0} - w_* q_{t_0}^\top\| \le \nu\|w_*\|$ where $W_{t_0} \in \mathbb{R}^{d\times K}$, $\nu > 0$ satisfies $2 - 14\nu - 2\nu^2 > 0$, and $q_{t_0} := \arg\min_{q\in\mathbb{R}^K : \|q\|_2 \le 1} \|W_{t_0} - w_* q^\top\|$. Then, gradient descent with the step size $\eta \le \frac{2-14\nu-2\nu^2}{(15+16\nu^2)^2\|w_*\|^4}$ generates iterates $\{W_t\}_{t\ge t_0}$ satisfying*

$$\text{dist}^2(W_{t+1}, w_*) \le (1 - \eta(2 - 14\nu - 2\nu^2))\text{dist}^2(W_t, w_*).$$

Proof. We have that

$$\begin{aligned}
\text{dist}^2(W_{t+1}) &:= \|W_{t+1} - w_* q_{t+1}^\top\|_F^2 \\
&\le \|W_{t+1} - w_* q_t^\top\|_F^2 \\
&= \|W_t - \eta\nabla F(W_t) - w_* q_t^\top\|_F^2 \\
&\overset{(a)}{=} \|w_t - \bar{w}_* - \eta\text{vec}\big(\nabla F(W_t) - \nabla F(w_* q_t^\top)\big)\|_F^2 \\
&\overset{(b)}{=} \|w_t - \bar{w}_* - \eta\Big(\int_0^1 \nabla^2 F(W_t(\tau))d\tau\Big)(w_t - \bar{w}_*)\|_F^2 \\
&= (w_t - \bar{w}_*)^\top \big(I_{dK} - \eta\int_0^1 \nabla^2 F(W_t(\tau))d\tau\big)^2 (w_t - \bar{w}_*) \\
&\le \|w_t - \bar{w}_*\|_2^2 - 2\eta(w_t - \bar{w}_*)^\top\Big(\int_0^1 \nabla^2 F(W_t(\tau))d\tau\Big)(w_t - \bar{w}_*) \\
&\quad + \eta^2\|\int_0^1 \nabla^2 F(W_t(\tau))d\tau\|_2^2 \|w_t - \bar{w}_*\|_2^2.
\end{aligned} \tag{18}$$

where (a) we use the notations that $w_t := \text{vec}(W_t)$ and $\bar{w}_* := \text{vec}(w_* q_t^\top)$ and that $\nabla F(w_* q_t^\top) = 0$ and (b) we denote $W_t(\tau) := w_* q_t^\top + \tau\big(W_t - w_* q_t^\top\big)$.

Notice that $\text{dist}(W_t(\tau), w_*) \le \|W_t(\tau) - w_* q_t^\top\| = \tau\|W_t - w_* q_t^\top\| \le \nu\|w_*\|$. So we can invoke Lemma 4 to obtain that

$$\begin{aligned}
&(w_t - \bar{w}_*)^\top\Big(\int_0^1 \nabla^2 F(W_t(\tau))d\tau\Big)(w_t - \bar{w}_*) \\
&\ge 2\text{tr}(q_t w_*^\top (W_t - w_* q_t^\top) q_t w_*^\top (W_t - w_* q_t^\top)) + (2 - 14\nu - 2\nu^2)\|w_*\|^2\|W_t - w_* q_t^\top\|_F^2 \\
&\ge (2 - 14\nu - 2\nu^2)\|w_*\|^2\|W_t - w_* q_t^\top\|_F^2.
\end{aligned} \tag{19}$$

where the last inequality is due to that $q_t w_*^\top (W_t - w_* q_t^\top)$ is a symmetric matrix since $q_t = \frac{W_t^\top w_*}{\|W_t^\top w_*\|}$ so that $\text{tr}(q_t w_*^\top (W_t - w_* q_t^\top) q_t w_*^\top (W_t - w_* q_t^\top)) = \|q_t w_*^\top (W_t - w_* q_t^\top)\|_F^2 \geq 0$. Furthermore, by Lemma 5, we have that

$$\|\nabla^2 F(W)\|_2 \leq (15 + 16\nu^2)\|w_*\|^2. \tag{20}$$

So

$$\eta^2 \| \int_0^1 \nabla^2 F(W_t(\tau)) d\tau \|_2^2 \|w_t - \bar{w}_*\|_2^2 \leq \eta^2 (15 + 16\nu^2)^2 \|w_*\|^4 \|w_t - \bar{w}_*\|_2^2. \tag{21}$$

Combining (18), (19), and (21), we get

$$\begin{aligned}
\text{dist}^2(W_{t+1}, w_*) \leq & \|w_t - \bar{w}_*\|_2^2 - 2\eta (w_t - \bar{w}_*)^\top \big(\int_0^1 \nabla^2 F(W_t(\tau)) d\tau \big) (w_t - \bar{w}_*) \\
& + \eta^2 \| \int_0^1 \nabla^2 F(W_t(\tau)) d\tau \|_2^2 \|w_t - \bar{w}_*\|_2^2 \\
\leq & \|W_t - w_* q_t^\top\|_F^2 - 2\eta(2 - 14\nu - 2\nu^2)\|w_*\|^2 \|W_t - w_* q_t^\top\|_F^2 \\
& + \eta^2 (15 + 16\nu^2)^2 \|w_*\|^4 \|W_t - w_* q_t^\top\|_F^2 \\
\leq & (1 - \eta(2 - 14\nu - 2\nu^2)) \text{dist}^2(W_t, w_*),
\end{aligned} \tag{22}$$

where the inequality we use that $\eta \leq \frac{2 - 14\nu - 2\nu^2}{(15 + 16\nu^2)^2 \|w_*\|^4}$.

□

Lemma 6. *(Lemma 12 in Li et al. (2019)) Suppose $x \sim N(0, I_d)$. Then for any fixed matrices $W, V \in \mathbb{R}^{d \times r}$. We have that*

$$\begin{aligned}
\mathbb{E}[\|x^\top V\|_2^2 \|x^\top W\|_2^2] &= \|V\|_F^2 \|W\|_F^2 + 2\|V^\top W\|_F^2 \\
\mathbb{E}[(x^\top W V^\top x)^2] &= \big(tr(W^\top V)\big)^2 + tr(W^\top V W^\top V) + \|W V^\top\|_F^2.
\end{aligned}$$

C Proof of Theorem 1

Theorem: 1 *Suppose that the conditions (C1-C2) hold. Assume that the step size satisfies $\eta \leq c/\|w_*\|^2$ for some sufficiently small constant $c > 0$. Then gradient descent for problem (1) (i.e. $K = 1$) has $T_\gamma \leq \frac{\log(\frac{\|w_*\| - \gamma}{|w_0[1]|})}{\log(1 + \eta\Delta)}$, where $\Delta := 6\gamma(\|w_*\| - \gamma) > 0$. Furthermore, for $0 \leq t \leq T_\gamma$, we have that $|w_t^{\|}| \geq (1 + \eta\Delta)^t |w_0^{\|}|$ and $\|w_t^\perp\| \leq \gamma$.*

Proof. Since it is clear about the number of the student neuron (which is 1), in the following, we suppress the subscript $\#1$ and the superscript (1) for the brevity of notations. Recall the dynamics,

$$\begin{aligned}
w_{t+1}^{\|} &= w_t^{\|} \big(1 + \eta(3\|w_*\|^2 - 3\|w_t\|^2)\big) \\
w_{t+1}^{\perp} &= w_t^{\perp} \big(1 + \eta(\|w_*\|^2 - 3\|w_t\|^2)\big).
\end{aligned} \tag{23}$$

and note that $w_t^{\|} = w_t[1]\|w_*\|$ so we have that

$$w_{t+1}[1] = w_t[1]\big(1 + \eta(3\|w_*\|^2 - 3\|w_t\|^2)\big) \tag{24}$$

For a number $\gamma > 0$, define $T_\gamma := \min\{t : ||w_t[1]| - \|w_*\|| \leq \gamma \text{ and } \|w_t^\perp\| \leq \gamma\}$. We can decompose the square of the distance term as follows,

$$\text{dist}^2(W_t^{\#1}, w_*) = ||w_t[1]| - \|w_*\||^2 + \|w_t^\perp\|^2.$$

In the latter part of this proof, we will show that $\|w_t^\perp\| \leq \gamma$ for all $t \leq T_\gamma$. Let us upper-bound the norm of $\|w_t\|^2$ for $t \leq T_\gamma$ as follows.

$$\begin{aligned}
\|w_t\|^2 = w_t[1]^2 + \|w_t^\perp\|^2 &\leq (\|w_*\| - \gamma)^2 + \gamma^2 \\
&= \|w_*\|^2 - 2\gamma\|w_*\| + 2\gamma^2,
\end{aligned} \tag{25}$$

where the inequality is because $w_t[1]^2 \leq (\|w_*\| - \gamma)^2$ for $t \leq T_\gamma$ by the definition and that $\|w_t^\perp\| \leq \gamma$ for all $t \leq T_\gamma$ proved in the latter part. Hence, we have that

$$\begin{aligned}|w_{t+1}[1]| &= |w_t[1]|\big(1 + \eta(3\|w_*\|^2 - 3\|w_t\|^2)\big) \geq |w_t[1]|\big(1 + \eta(3\|w_*\|^2 - 3(\|w_*\|^2 - 2\gamma\|w_*\| + 2\gamma^2))\big) \\ &:= |w_t[1]|\big(1 + \eta\Delta\big) = |w_0[1]|\big(1 + \eta\Delta\big)^{t+1},\end{aligned} \tag{26}$$

with

$$\Delta := 6\gamma(\|w_*\| - \gamma) > 0.$$

Based on (26), it takes at most number of iterations

$$T_\gamma \leq \frac{\log(\frac{\|w_*\| - \gamma}{|w_0[1]|})}{\log(1 + \eta\Delta)}$$

for $w_t[1]$ to rises above $\|w_*\| - \gamma$ if $w_0[1] > 0$. Similarly, if $w_0[1] < 0$, it takes T_γ iterations for $w_0[1]$ to satisfy $w_t[1] \leq -\|w_*\| + \gamma$.

Now we switch to show that for $0 \leq t \leq T_\gamma$, $\gamma \geq \|w_t^\perp\|$. We are going show that the perpendicular component $\|w_t^\perp\|$ start decaying before it could have increased above γ. From the dynamics $\|w_{t+1}^\perp\| = \|w_t^\perp\|\big(1 + \eta(\|w_*\|^2 - 3\|w_t\|^2)\big)$, we see that once $\|w_t\|^2 \geq \frac{1}{3}\|w_*\|^2$, the size of the perpendicular component $\|w_t^\perp\|$ starts decaying. On the other hand, $\|w_t^\perp\|$ is increasing when $\|w_t\|^2 \leq \frac{1}{3}\|w_*\|^2$, which leads to the following before the perpendicular component starts decaying,

$$\begin{aligned}\|w_t\|^2 \geq w_t^2[1] = \frac{1}{\|w_*\|^2}|w_t^\parallel|^2 &= \frac{1}{\|w_*\|^2}|w_{t-1}^\parallel|^2\big(1 + \eta(3\|w_*\|^2 - 3\|w_{t-1}\|^2)\big)^2 \\ &\geq \frac{1}{\|w_*\|^2}|w_{t-1}^\parallel|^2\big(1 + 2\eta\|w_*\|^2\big)^2 \\ &\geq \frac{1}{\|w_*\|^2}|w_0^\parallel|^2\big(1 + 2\eta\|w_*\|^2\big)^{2t} = |w_0[1]|^2\big(1 + 2\eta\|w_*\|^2\big)^{2t},\end{aligned} \tag{27}$$

which means that that the size of $\|w_t\|^2$ grows at the rate at least $\big(1 + 2\eta\|w_*\|^2\big)^2$ before the perpendicular component $\|w_t^\perp\|$ starts decaying.

The inequality (27) also implies that the number of iterations such that $\|w_t\|^2 \leq \frac{1}{3}\|w_*\|^2$ is at most

$$t^* \leq \frac{1}{2}\frac{\log(\frac{\|w_*\|^2}{3|w_0[1]|^2})}{\log\big((1 + 2\eta\|w_*\|^2)^2\big)}. \tag{28}$$

After t^*, we have that $\|w_t^\perp\|$ is decaying. So we only have to show that for $0 \leq t \leq t^*$ the perpendicular component never rise above γ. It suffices to show that an upper bound of $\|w_t^\perp\|$ for $0 \leq t \leq t^*$ is not greater than γ,

$$\begin{aligned}\|w_t^\perp\| &= \|w_{t-1}^\perp\|\big(1 + \eta(\|w_*\|^2 - 3\|w_{t-1}\|^2)\big) \\ &\leq \|w_{t-1}^\perp\|\big(1 + \eta(\|w_*\|^2 - 3w_{t-1}^2[1])\big) \\ &\leq \|w_0^\perp\| \cdot \Pi_{s=0}^{t^*-1}\big(1 + \eta(\|w_*\|^2 - 3w_s^2[1])\big) \\ &\overset{(a)}{\leq} \|w_0^\perp\| \cdot \Pi_{s=0}^{t^*-1}\big(1 + \eta(\|w_*\|^2 - 3w_0^2[1](1 + \eta\Delta)^{2s})\big) \\ &\overset{?}{\leq} \gamma,\end{aligned} \tag{29}$$

where (a) we use (26). By taking logarithm on the both sides of $\overset{?}{\leq}$, it suffices to show that

$$\log\|w_0^\perp\| + \sum_{s=0}^{t^*-1}\log\big(1 + \eta(\|w_*\|^2 - 3w_0^2[1](1 + \eta\Delta)^{2s})\big) \overset{?}{\leq} \log\gamma. \tag{30}$$

Using the fact that $\log x \geq 1 - \frac{1}{x}$ and that $\log(1+x) \leq x$ for $x > -1$, it suffices to show that

$$\sum_{s=0}^{t^*-1} \eta(\|w_*\|^2 - 3w_0^2[1](1+\eta\Delta)^{2s}) \overset{?}{\leq} 1 - \frac{\|w_0^\perp\|}{\gamma}.$$

which can be guaranteed if

$$\eta t^* \|w_*\|^2 \leq 3\eta w_0^2[1] \frac{(1+\eta\Delta)^{2t^*}-1}{(1+\eta\Delta)^2-1} + \frac{9}{10}, \tag{31}$$

where we use that $\gamma \geq 10\|w_0^\perp\|$. So it suffices to have the step size satisfy $\eta = c/\|w_*\|^2$ for some sufficiently small constant $c > 0$.

□

D Proof of Lemma 2, Lemma 3, and Theorem 2

Lemma 2: *Suppose that the approximated dynamics (10) and (11) hold from iteration* 0 *to iteration* t*. Then, the network with a single neuron and an over-parametrized network trained by GD with the same step size* η *has* $\sqrt{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2} \gtrsim \sqrt{(1-2t\theta)}\sqrt{K}|w_{\#1,t}^{(1),\|}|$.

Proof. From the dynamics (10), we have that

$$\begin{aligned} |w_{\#K,t}^{(k),\|}| \geq& (1-\theta)|w_{\#K,t-1}^{(k),\|}|\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#K,t-1}^{(k)}\|^2)\big) \\ \geq& (1-\theta)^t |w_{\#K,0}^{(k),\|}| \cdot \Pi_{s=0}^{t-1}\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#K,s}^{(k)}\|^2)\big). \end{aligned} \tag{32}$$

Therefore,

$$\begin{aligned} \frac{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2}{|w_{\#1,t}^{(1),\|}|^2} \overset{(a)}{\geq}& \frac{\sum_{k=1}^K (1-\theta)^{2t}|w_{\#K,0}^{(k),\|}|^2 \cdot \Pi_{s=0}^{t-1}\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#K,s}^{(k)}\|^2)\big)^2}{|w_{\#1,0}^{(1),\|}|^2 \cdot \Pi_{s=0}^{t-1}\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#1,s}^{(1)}\|^2)\big)^2}. \\ \overset{(b)}{\gtrsim} \sum_{k=1}^K (1-\theta)^{2t} \frac{|w_{\#K,0}^{(k),\|}|^2}{|w_{\#1,0}^{(1),\|}|^2} \overset{(c)}{\geq}& (1-2t\theta)\sum_{k=1}^K \frac{|w_{\#K,0}^{(k),\|}|^2}{|w_{\#1,0}^{(1),\|}|^2} \overset{(d)}{\gtrsim} (1-2t\theta)K, \end{aligned} \tag{33}$$

where (a) is due to (32) and the recursive expansion, (b) is by using that $\|w_{\#1,s}^{(1)}\|^2 \gtrsim \|w_{\#K,s}^{(k)}\|^2$ so that $\Pi_{s=0}^{t-1}\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#K,s}^{(k)}\|^2)\big)^2 \gtrsim \Pi_{s=0}^{t-1}\big(1+\eta(3\|w_*\|^2 - 3\|w_{\#1,s}^{(1)}\|^2)\big)^2$, (c) is by $(1+x)^r \geq 1+rx$ for $x > -1$ and $r \in R \setminus (0,1)$, and (d) is by the initialization such that each node is i.i.d. initialized from the gaussian distribution.

□

Lemma: 3 *Suppose that* $\eta \leq \frac{1}{3\|w_*\|^2}$*. By following the conditions as Lemma 2, we have that* $\|w_{\#K,t}^{(k),\perp}\| \lesssim \frac{|w_{\#K,t}^{(k),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t} \lesssim \frac{|w_{\#1,t}^{(1),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t}$*, where* $\psi := (1-\theta-\vartheta-\theta\vartheta)\big(1+\eta\|w_*\|^2\big)$.

Proof. For brevity, let us suppress the subscript $\#K$ for the moment. By the approximated dynamics (10), we have that

$$\begin{aligned} \frac{|w_{t+1}^{(k),\|}|}{\|w_{t+1}^{(k),\perp}\|} \geq& \frac{(1-\theta)(1+\eta(3\|w_*\|^2 - \|w_t^{(k)}\|^2))}{(1+\vartheta)(1+\eta(\|w_*\|^2 - \|w_t^{(k)}\|^2))}\frac{|w_t^{(k),\|}|}{\|w_t^{(k),\perp}\|} \\ \geq& (1-\theta-\vartheta-\theta\vartheta)\big(1+2\eta\|w_*\|^2 - \eta^2(3\|w_*\|^2 - \|w_t^{(k)}\|^2)(\|w_*\|^2 - \|w_t^{(k)}\|^2)\big)\frac{|w_t^{(k),\|}|}{\|w_t^{(k),\perp}\|} \\ \gtrsim& \psi\frac{|w_t^{(k),\|}|}{\|w_t^{(k),\perp}\|}, \end{aligned} \tag{34}$$

where the second inequality we use $\frac{1+a}{1+b} \geq 1+a-b-ab$ for $b \geq -1$ and the last inequality is because $(1-\theta-\vartheta-\theta\vartheta)\big(1+2\eta\|w_*\|^2-\eta^2(3\|w_*\|^2-\|w_t^{(k)}\|^2)(\|w_*\|^2-\|w_t^{(k)}\|^2)\big) \gtrsim (1-\theta-\vartheta-\theta\vartheta)\big(1+\eta\|w_*\|^2\big) := \psi$, as $\|w_{\#K,t}^{(k)}\|^2 \lesssim \|w_{\#1,t}^{(1)}\|^2 < \|w_*\|^2$ and that $\eta \leq \frac{1}{3\|w_*\|^2}$.

Consequently, $\frac{|w_t^{(k),\|}|}{\|w_t^{(k),\perp}\|} \gtrsim \psi^t \frac{|w_0^{(k),\|}|}{\|w_0^{(k),\perp}\|}$. Namely,

$$\|w_{\#K,t}^{(k),\perp}\| \lesssim \frac{|w_{\#K,t}^{(k),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t}. \tag{35}$$

Moreover, by the condition that $|w_{\#1,t}^{(1),\|}| \gtrsim |w_{\#K,t}^{(k),\|}|$, we have that

$$\|w_{\#K,t}^{(k),\perp}\| \lesssim \frac{|w_{\#K,t}^{(k),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t} \lesssim \frac{|w_{\#1,t}^{(1),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t}. \tag{36}$$

□

Theorem: 2 *(Snapshot at t) Suppose that the approximated dynamics (10) and (11) hold from 0 to t and that at iteration t, the student network with a single neuron trained by GD with the step size η has $|w_{\#1,t}^{(1),\|}| = c_{1,t}\|w_*\|^2$ for some number $c_{1,t}$ satisfying $1 > c_{1,t} > 0$. Denote $c_{2,t}$ a number that satisfies $\frac{c_{1,t}\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k)}[1]|}\frac{1}{\psi^t} \leq c_{2,t}$ for each $k \in [K]$. Suppose that the step size η also satisfies $\eta \leq \frac{1}{3\|w_*\|^2}$ and makes $\psi := (1-\theta-\vartheta-\theta\vartheta)\big(1+\eta\|w_*\|^2\big) > 1$. Then, an over-parametrized network $W_t^{\#K}$ trained by GD with the same η has*
$\text{dist}^2(W_t^{\#1}, w_*) - \text{dist}^2(W_t^{\#K}, w_*) \gtrsim \|w_*\|^2\big(2c_{1,t}(\sqrt{(1-2t\theta)}\sqrt{K}-1) - K(c_{1,t}^2+c_{2,t}^2) + c_{1,t}^2\big)$.

Proof. Let us compute $\text{dist}(W_t^{\#1}, w_*)$ and $\text{dist}(W_t^{\#k}, w_*)$. For $\text{dist}(W_t^{\#1}, w_*)^2$, we have that

$$\begin{aligned}\text{dist}(W_t^{\#1}, w_*)^2 &= \min_{q\in\{-1,+1\}} \|W_t^{\#1} - w_* q^\top\|^2 = \|w_{\#1,t}^{(1)}\|^2 - 2|w_{\#1,t}^{(1),\|}| + \|w_*\|^2 \\ &= \|w_{\#1,t}^{(1)}\|^2 + \|w_*\|^2 - 2\sqrt{\|w_{\#1,t}^{(1),\|}\|^2}.\end{aligned} \tag{37}$$

On the other hand, for $\text{dist}^2(W_t^{\#K}, w_*)$, we have that

$$\begin{aligned}\text{dist}^2(W_t^{\#K}, w_*) &= \min_{q\in\mathbb{R}^K:\|q\|_2\leq 1} \|W_t^{\#K} - w_* q^\top\|_F^2 \\ &= \min_{q\in\mathbb{R}^K:\|q\|_2\leq 1} \text{tr}\big((W_t^{\#K} - w_* q^\top)^\top (W_t^{\#K} - w_* q^\top)\big) \\ &= \|W_t^{\#K}\|_F^2 + \|w_*\|^2 - 2\max_{q\in\mathbb{R}^K:\|q\|_2\leq 1} \text{tr}\big((W_t^{\#K})^\top w_* q^\top\big) \\ &= \|W_t^{\#K}\|_F^2 + \|w_*\|^2 - 2\sqrt{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2},\end{aligned} \tag{38}$$

where the last inequality is due to that

$$\max_{q\in\mathbb{R}^K:\|q\|_2\leq 1} \text{tr}\big((W_t^{\#K})^\top w_* q^\top\big) \overset{(a)}{=} \max_{q\in\mathbb{R}^K:\|q\|_2\leq 1} \text{tr}\big(\bar{v} q^\top\big) = \max_{q\in\mathbb{R}^K:\|q\|_2\leq 1} \text{tr}\big(q^\top \bar{v}\big) \overset{(b)}{=} \|\bar{v}\|. \tag{39}$$

where (a) we denote $\bar{v} := [w_{\#K,t}^{(1),\|}, w_{\#K,t}^{(2),\|}, \ldots, w_{\#K,t}^{(K),\|}]^\top \in \mathbb{R}^K$ and (b) is because $q = \bar{v}/\|\bar{v}\|$. Combining (37) and (38), we have that

$$\begin{aligned}&\text{dist}^2(W_t^{\#1}, w_*) - \text{dist}^2(W_t^{\#K}, w_*) \\ &= 2\Big(\sqrt{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2} - \sqrt{\|w_{\#1,t}^{(1),\|}\|^2}\Big) - \big(\|W_t^{\#K}\|_F^2 - \|W_t^{\#1}\|_F^2\big).\end{aligned} \tag{40}$$

To continue, we need the lower bound of $\sqrt{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2} - \sqrt{\|w_{\#1,t}^{(1),\|}\|^2}$ and the upper bound of $\|W_t^{\#K}\|_F^2 - \|W_t^{\#1}\|_F^2$. By Lemma 2, we have that

$$\sqrt{\sum_{k=1}^K |w_{\#K,t}^{(k),\|}|^2} - \sqrt{\|w_{\#1,t}^{(1),\|}\|^2} \gtrsim (\sqrt{(1-2t\theta)}\sqrt{K}-1)\|w_{\#1,t}^{(1),\|}\| = (\sqrt{(1-2t\theta)}\sqrt{K}-1)c_{1,t}\|w_*\|^2. \tag{41}$$

On the other hand, for the difference of the norms, we have that $\|W_t^{\#1}\|^2 \geq |w_{\#1,t}^{(1)}(1)|^2 = \frac{|w_{\#1,t}^{(1),\|}|^2}{\|w_*\|^2}$. Furthermore, by Lemma 3 and that $|w_{\#1,t}^{(1),\|}| = c_{1,t}\|w_*\|^2$ and the definition of $c_{2,t}$, $\|w_{\#K,t}^{(k),\perp}\|_2 \lesssim \frac{|w_{\#1,t}^{(1),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k),\|}|}\frac{1}{\psi^t} = \frac{|w_{\#1,t}^{(1),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k)}[1]|\|w_*\|}\frac{1}{\psi^t} \leq c_{2,t}\|w_*\|$. So we have that

$$\begin{aligned}
&\|W_t^{\#K}\|_F^2 = \sum_{k=1}^K \|w_{\#K,t}^{(k)}\|^2 = \sum_{k=1}^K \big(\|w_{\#K,t}^{(k),\perp}\|^2 + \frac{1}{\|w_*\|^2}|w_{\#K,t}^{(k),\|}|^2\big) \\
&\lesssim \sum_{k=1}^K \big(\big(\frac{|w_{\#1,t}^{(1),\|}|\|w_{\#K,0}^{(k),\perp}\|}{|w_{\#K,0}^{(k)}[1]|\|w_*\|}\frac{1}{\psi^t}\big)^2 + \frac{1}{\|w_*\|^2}|w_{\#K,t}^{(k),\|}|^2\big) \leq K\|w_*\|^2 c_{2,t}^2 + \sum_{k=1}^K \frac{|w_{\#K,t}^{(k),\|}|^2}{\|w_*\|^2} \\
&\lesssim K\|w_*\|^2(c_{2,t}^2 + c_{1,t}^2),
\end{aligned} \tag{42}$$

where the last inequality uses (11). Therefore, by combining (40,41,42),

$$\text{dist}^2(W_t^{\#1}, w_*) - \text{dist}^2(W_t^{\#K}, w_*) \gtrsim \|w_*\|^2\big(2c_{1,t}(\sqrt{(1-2t\theta)}\sqrt{K}-1) - K(c_{1,t}^2 + c_{2,t}^2) + c_{1,t}^2\big). \tag{43}$$

□

E PROOF OF THEOREM 3

Theorem 3 *Suppose that $\|w_*\| > 1.1\sqrt{C}\gamma$ and $\gamma \geq 10\|w_0\|$. Assume that the step size satisfies $\eta \leq c/\|w_*\|^2$ for some sufficiently small constant $c > 0$. Then gradient descent for problem (12) has $T_{\gamma,C} \leq \frac{\log(\frac{\|w_*\|/\sqrt{C}-\gamma}{|w_0[1]|})}{\log(1+\eta\Delta_C)}$, where $\Delta_C := 6\gamma C(\sqrt{C}\|w_*\| - C\gamma) > 0$. Furthermore, for $0 \leq t \leq T_{\gamma,C}$, we have that $|w_t^{\|}| \geq (1+\eta\Delta_C)^t|w_0^{\|}|$ and $\|w_t^{\perp}\| \leq \gamma$.*

Proof. Recall the objective is

$$\min_{w\in\mathbb{R}^d} f(w) := \frac{1}{4n}\sum_{i=1}^n \big(C(x_i^\top w)^2 - y_i\big)^2. \tag{44}$$

With population gradient, the dynamics of gradient descent is

$$\begin{aligned}
w_{t+1}^{(1),\|} &= w_t^{(1),\|}\big(1 + \eta(3C\|w_*\|^2 - 3C^2\|w_t^{(1)}\|^2)\big) \\
w_{t+1}^{(1),\perp} &= w_t^{(1),\perp}\big(1 + \eta(C\|w_*\|^2 - 3C^2\|w_t^{(1)}\|^2)\big).
\end{aligned} \tag{45}$$

Since it is clear about the number of the student neuron (which is 1), in the following, we will suppress the subscript $\#1$ and the superscript (1) for the brevity of notations. Note that $w_t^{\|} = w_t[1]\|w_*\|$ so we have that

$$w_{t+1}[1] = w_t[1]\big(1 + \eta(3C\|w_*\|^2 - 3C^2\|w_t\|^2)\big). \tag{46}$$

For a number $\gamma > 0$, define $T_{\gamma,C} := \min\{t : ||w_t[1]| - \frac{\|w_*\|}{\sqrt{C}}| \leq \gamma \text{ and } \|w_t^{\perp}\| \leq \gamma\}$. We can decompose the square of the distance term as follows,

$$\text{dist}^2(W_t^{\#1}, \frac{\|w_*\|}{\sqrt{C}}) := ||w_t[1]| - \frac{\|w_*\|}{\sqrt{C}}|^2 + \|w_t^{\perp}\|^2.$$

In the latter part of this proof, we will show that $\|w_t^\perp\| \leq \gamma$ for all $t \leq T_{\gamma,C}$. Let us upper-bound the norm of $\|w_t\|^2$ for $t \leq T_\gamma$ as follows.

$$\begin{aligned}\|w_t\|^2 = w_t[1]^2 + \|w_t^\perp\|^2 &\leq (\frac{\|w_*\|}{\sqrt{C}} - \gamma)^2 + \gamma^2 \\ &= \frac{\|w_*\|^2}{C} - 2\gamma\frac{\|w_*\|}{\sqrt{C}} + 2\gamma^2, \end{aligned} \tag{47}$$

where the inequality is because $w_t[1]^2 \leq (\frac{\|w_*\|}{\sqrt{C}} - \gamma)^2$ for $t \leq T_{\gamma,C}$ by the definition and that $\|w_t^\perp\| \leq \gamma$ for all $t \leq T_{\gamma,C}$ proved in the latter part. Hence, we have that

$$\begin{aligned}|w_{t+1}[1]| &= |w_t[1]|\big(1 + \eta(3C\|w_*\|^2 - 3C^2\|w_t\|^2)\big) \\ &\geq |w_t[1]|\big(1 + \eta(3C\|w_*\|^2 - 3C^2(\frac{\|w_*\|^2}{C} - 2\gamma\frac{\|w_*\|}{\sqrt{C}} + 2\gamma^2))\big) \\ &:= |w_t[1]|\big(1 + \eta\Delta_C\big) = |w_0[1]|\big(1 + \eta\Delta_C\big)^t, \end{aligned} \tag{48}$$

with

$$\Delta_C := 6\gamma C(\sqrt{C}\|w_*\| - C\gamma).$$

Note that $\Delta_C > 0$ when $\|w_*\| > \sqrt{C}\gamma$.

Based on (48), it takes at most number of iterations

$$T_{\gamma,C} \leq \frac{\log(\frac{\|w_*\|/\sqrt{C} - \gamma}{|w_0[1]|})}{\log(1 + \eta\Delta_C)}$$

for $w_t[1]$ to rises above $\frac{\|w_*\|}{\sqrt{C}} - \gamma$ if $w_0[1] > 0$. Similarly, if $w_0[1] < 0$, it takes T_γ iterations for $w_0[1]$ satisfies $w_0[1] \leq -\frac{\|w_*\|}{\sqrt{C}} + \gamma$.

Now we switch to show that $0 \leq t \leq T_{\gamma,C}$, $\gamma \geq \|w_t^\perp\|$. We are going show that the perpendicular component $\|w_t^\perp\|$ starts decaying before it could have increased above γ. From the dynamics of $\|w_{t+1}^\perp\| = \|w_t^\perp\|\big(1+\eta(C\|w_*\|^2 - 3C^2\|w_t\|^2)\big)$, we see that once $\|w_t\|^2 \geq \frac{1}{3C}\|w_*\|^2$, the size of the perpendicular component $\|w_t^\perp\|$ starts decaying. On the other hand, it is increasing when $\|w_t\|^2 \leq \frac{1}{3C}\|w_*\|^2$, which leads to the following before the perpendicular component starts decaying,

$$\begin{aligned}\|w_t\|^2 \geq w_t^2[1] = \frac{1}{\|w_*\|^2}|w_t^\parallel|^2 &= \frac{1}{\|w_*\|^2}|w_{t-1}^\parallel|^2\big(1 + \eta(3C\|w_*\|^2 - 3C^2\|w_{t-1}\|^2)\big)^2 \\ &\geq \frac{1}{\|w_*\|^2}|w_{t-1}^\parallel|^2\big(1 + 2\eta C\|w_*\|^2\big)^2 \\ &\geq \frac{1}{\|w_*\|^2}|w_0^\parallel|^2\big(1 + 2\eta C\|w_*\|^2\big)^{2t} = |w_0[1]|^2\big(1 + 2\eta C\|w_*\|^2\big)^{2t} \end{aligned} \tag{49}$$

which means that that the size of $\|w_t\|^2$ grows at the rate at least $\big(1 + 2\eta C\|w_*\|^2\big)^2$ before the perpendicular component $\|w_t^\perp\|$ starts decaying.

The inequality (27) also implies that the number of iterations such that $\|w_t\|^2 \leq \frac{1}{3C}\|w_*\|^2$ is at most

$$t_C^* \leq \frac{1}{2}\frac{\log(\frac{\|w_*\|^2}{3C|w_0[1]|^2})}{\log\big((1 + 2\eta C\|w_*\|^2)^2\big)}. \tag{50}$$

After t_C^*, we have that $\|w_t^\perp\|$ is decaying. So we only have to show that for $0 \leq t \leq t_C^*$ the perpendicular component never rise above γ. It suffices to show that an upper bound of $\|w_{t_C}^\perp\|$ for

$0 \le t \le t_C^*$ is not greater than γ,

$$\begin{aligned}\|w_t^\perp\| &= \|w_{t-1}^\perp\|\big(1+\eta(C\|w_*\|^2 - 3C^2\|w_{t-1}\|^2)\big)\\ &\le \|w_{t-1}^\perp\|\big(1+\eta(C\|w_*\|^2 - 3C^2 w_{t-1}^2[1])\big)\\ &\le \|w_0^\perp\|\cdot\Pi_{s=0}^{t_C^*-1}\big(1+\eta(C\|w_*\|^2 - 3C^2 w_s^2[1])\big)\\ &\overset{(a)}{\le} \|w_0^\perp\|\cdot\Pi_{s=0}^{t_C^*-1}\big(1+\eta(C\|w_*\|^2 - 3C^2 w_0^2[1](1+\eta\Delta_C)^{2s})\big)\\ &\overset{?}{\le} \gamma, \end{aligned} \tag{51}$$

where (a) we use (26). By taking logarithm on the both sides of $\overset{?}{\le}$, it suffices to show that

$$\log\|w_0^\perp\| + \sum_{s=0}^{t_C^*-1}\log\big(1+\eta(C\|w_*\|^2 - 3C^2 w_0^2[1](1+\eta\Delta_C)^{2s})\big) \overset{?}{\le} \log\gamma. \tag{52}$$

Using the fact that $\log x \ge 1-\frac{1}{x}$ and that $\log(1+x)\le x$ for $x>-1$, it suffices to show that

$$\sum_{s=0}^{t_C^*-1}\eta(C\|w_*\|^2 - 3C^2 w_0^2[1](1+\eta\Delta_C)^{2s}) \overset{?}{\le} 1-\frac{\|w_0^\perp\|}{\gamma}.$$

which can be guaranteed if

$$\eta t_C^* C\|w_*\|^2 \le 3C^2\eta w_0^2[1]\frac{(1+\eta\Delta_C)^{2t_C^*}-1}{(1+\eta\Delta_C)^2-1} + \frac{9}{10}, \tag{53}$$

where we use that $\gamma \ge 10\|w_0^\perp\|$. So it suffices to have the step size satisfy $\eta = c/\|w_*\|^2$ for some sufficiently small constant $c>0$.

□

E.1 Optimization landscape

In this subsection, we analyze the optimization landscape of the following,

$$\min_{W\in\mathbb{R}^{d\times K}} f(W) := \tfrac{1}{4n}\sum_{i=1}^n\big(C(x_i^\top w^{(1)})^2 + C(x_i^\top w^{(2)})^2 + ... + C(x_i^\top w^{(K)})^2 - y_i\big)^2. \tag{54}$$

We define

$$\text{dist}(W, \tfrac{w_*}{\sqrt{C}}) := \min_{q\in\mathbb{R}^K:\|q\|_2\le 1}\|W - \tfrac{w_*}{\sqrt{C}}q^\top\|. \tag{55}$$

This is due to the observation that the set of global optimal solutions of (2) that achieves zero testing error is $\frac{w_*}{\sqrt{C}}q^\top \in \mathbb{R}^{d\times K}$ for any $q\in\mathbb{R}^K$ such that $\|q\|_2 = 1$. We denote $\nabla F(W) := \mathbb{E}_x[\nabla f(W)]$ and $\nabla^2 F(W) := \mathbb{E}_x[\nabla^2 f(W)]$, where $f(\cdot)$ corresponds to (54).

Lemma 7. *(locally strong convexity) Assume that* $\text{dist}(W_{t_0}, \frac{w_*}{\sqrt{C}}) := \|W_{t_0} - \frac{w_*}{\sqrt{C}}q_{t_0}^\top\|_F \le \nu\|w_*\|_2$, *where* $W_{t_0}\in\mathbb{R}^{d\times K}$, $\nu>0$, $q_{t_0} := \arg\min_{q\in\mathbb{R}^K:\|q\|_2\le 1}\|W_{t_0} - \frac{w_*}{\sqrt{C}}q^\top\|_F$, *and* $w_*\in\mathbb{R}^d$ *being the teacher neuron. Then*

$\textit{vec}(V)^\top\nabla^2 F(W_{t_0})\textit{vec}(V) \ge 2C\textit{tr}(q_{t_0}w_*^\top V q_{t_0}w_*^\top V) + (2C - 14C^{3/2}\nu - 2C^2\nu^2)\|w_*\|^2\|V\|_F^2$, *for any* $V\in\mathbb{R}^{d\times K}$, *where tr*$(\cdot)$ *denotes the matrix trace.*

Proof. For notation brevity, we will suppress script t_0 in the following (i.e. $W\leftarrow W_{t_0}$, $q\leftarrow q_{t_0}$). We have that

$$\begin{aligned}&\text{vec}(V)^\top\nabla^2 f(W)\text{vec}(V)\\ &= \frac{1}{m}\sum_{i=1}^m \text{vec}(V)^\top\big[\big((C\|x_i^\top W\|_2^2 - y_i)CI_K + 2C^2 W^\top x_i x_i^\top W\big)\otimes x_i x_i^\top\big]\text{vec}(V)\\ &= \frac{1}{m}\sum_{i=1}^m C\big(C\|x_i^\top W\|_2^2 - y_i\big)\text{vec}(V)^\top\text{vec}(x_i x_i^\top V) + \frac{C^2}{m}\sum_{i=1}^m\text{vec}(V)^\top\text{vec}(2x_i x_i^\top V W^\top x_i x_i^\top W).\\ &= \frac{1}{m}\sum_{i=1}^m\big[C\big(C\|x_i^\top W\|_2^2 - \|x_i^\top w_* q^\top\|_2^2\big)\|x_i^\top V\|_2^2 + 2C^2\big(x_i^\top W V^\top x_i\big)^2\big].\end{aligned}$$

By taking the expectation above over $x_i \sim N(0, I_d)$ and using Lemma 6, we have that

$$\mathbb{E}[\text{vec}(V)^\top \nabla^2 f(W)\text{vec}(V)] = C^2\|W\|_F^2\|V\|_F^2 + 2C^2\|V^\top W\|_F^2 - C\big(\|w_* q^\top\|_F^2\|V\|_F^2 + 2\|V^\top w_* q^\top\|_F^2\big) + 2C^2\big(\text{tr}(W^\top V)^2 + \text{tr}(W^\top V W^\top V) + \|WV^\top\|_F^2\big). \tag{56}$$

Now let us set $W = \frac{w_*}{\sqrt{C}}q^\top + \theta H$ with a H satisfying $\|H\|_F = 1$ and $\theta := \nu\|w_*\|$ for some number $\nu > 0$.

$$\begin{aligned}
\|W\|_F^2 &= \|\frac{w_*}{\sqrt{C}}q^\top\|_F^2 + \theta^2\|H\|_F^2 + 2\theta\langle\frac{w_*}{\sqrt{C}}q^\top, H\rangle \\
&\geq \|\frac{w_*}{\sqrt{C}}\|^2 - 2\theta\|\frac{w_*}{\sqrt{C}}q^\top\|_F\|H\|_F \\
&\geq \|\frac{w_*}{\sqrt{C}}\|^2 - 2\theta\|\frac{w_*}{\sqrt{C}}\|\|H\|_F \\
\|V^\top W\|_F^2 &= \|V^\top\frac{w_*}{\sqrt{C}}q^\top\|_F^2 + \theta^2\|V^\top H\|_F^2 + 2\theta\text{tr}(V^\top\frac{w_*}{\sqrt{C}}q^\top H^\top V) \\
&\geq \|V^\top\frac{w_*}{\sqrt{C}}q^\top\|_F^2 - 2\theta\|\frac{w_*}{\sqrt{C}}q^\top\|\|H\|_F\|V\|_F^2 \\
\|VW^\top\|_F^2 &= \|Vq(\frac{w_*}{\sqrt{C}})^\top\|_F^2 + \theta^2\|VH^\top\|_F^2 + 2\theta\text{tr}(V(\frac{w_*}{\sqrt{C}})^\top qHV^\top) \\
&\geq \|Vq(\frac{w_*}{\sqrt{C}})^\top\|_F^2 - 2\theta\|(\frac{w_*}{\sqrt{C}})^\top q\|\|H\|_F\|V\|_F^2 \\
&= \|Vq(\frac{w_*}{\sqrt{C}})^\top\|_F^2 - 2\theta\|(\frac{w_*}{\sqrt{C}})\|\|H\|_F\|V\|_F^2 \\
\text{tr}(W^\top VW^\top V) &= \frac{1}{C}\text{tr}(qw_*^\top Vqw_*^\top V) + 2\theta\text{tr}(H^\top Vq(\frac{w_*}{\sqrt{C}})^\top V) + \theta^2\text{tr}(H^\top VH^\top V) \\
&= \frac{1}{C}\text{tr}(qw_*^\top Vqw_*^\top V) - 2\theta\|(\frac{w_*}{\sqrt{C}})q^\top\|\|H\|_F\|V\|_F^2 - \theta^2\|H\|_F^2\|V\|_F^2.
\end{aligned} \tag{57}$$

Combining (56) and (57), together with the bilinear property of the expectation so that $\text{vec}(V)^\top\nabla^2 F(w)\text{vec}(V) = \mathbb{E}[\text{vec}(V)^\top\nabla^2 f(W)\text{vec}(V)]$, we have that

$$\begin{aligned}
\text{vec}(V)^\top\nabla^2 F(W)\text{vec}(V) \geq& 2C\text{tr}(qw_*^\top Vqw_*^\top V) + 2C\|Vqw_*^\top\|_F^2 \\
&- 14C^{3/2}\theta\|w_*\|\|H\|_F\|V\|_F^2 - 2\theta^2C^2\|H\|_F^2\|V\|_F^2 \\
\overset{(a)}{=}& 2C\text{tr}(qw_*^\top Vqw_*^\top V) + 2C\|w_*\|^2\|V\|_F^2 \\
&- 14C^{3/2}\theta\|w_*\|\|H\|_F\|V\|_F^2 - 2\theta^2C^2\|H\|_F^2\|V\|_F^2 \\
=& 2C\text{tr}(qw_*^\top Vqw_*^\top V) + (2C - 14C^{3/2}\nu - 2C^2\nu^2)\|w_*\|^2\|V\|_F^2,
\end{aligned} \tag{58}$$

where (a) is due to that $\|AB\|_F^2 \geq \sigma_r^2(A)\|B\|_F^2$ with r being the rank of A, which in our case $A := w_*q^\top$ is a rank one matrix and $B := V$; consequently $\sigma_1^2(A) = \|w_*q^\top\|_F^2 = \|w_*\|^2$. □

Lemma 8. *(smoothness) Assume that* $\text{dist}(W_{t_0}, \frac{w_*}{\sqrt{C}}) := \|W - \frac{w_*}{\sqrt{C}}q_{t_0}^\top\|_F \leq \nu\|w_*\|_2$ *where* $W_{t_0} \in \mathbb{R}^{d\times K}$, $\nu > 0$, $q_{t_0} := \arg\min_{q\in\mathbb{R}^K:\|q\|_2\leq 1}\|W - \frac{w_*}{\sqrt{C}}q^\top\|_F$, *and* $w_* \in \mathbb{R}^d$ *being the teacher neuron. Then,* $\|\nabla^2 F(W)\|_2 \leq (15C + 16C^2\nu^2)\|w_*\|^2$.

Proof. For notation brevity, we will suppress script t_0 in the following (i.e. $W \leftarrow W_{t_0}$, $q \leftarrow q_{t_0}$).

$$\begin{aligned}
\|\nabla^2 F(W)\|_2 &= \|\mathbb{E}\big[[C\big(C\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2\big)I_K + 2C^2 W^\top x x^\top W] \otimes x x^\top\big]\|_2 \\
&\le \|\mathbb{E}\big[[C\big|C\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2\big|I_K + 2C^2\|x^\top W\|_2^2 I_K] \otimes x x^\top\big]\|_2 \\
&\overset{(a)}{\le} \|\mathbb{E}\big[C\big|C\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2\big|x x^\top\big]\|_2 + 2C^2\|\mathbb{E}\big[\|x^\top W\|_2^2 x x^\top\big]\|_2 \\
&\overset{(b)}{=} C\|\mathbb{E}\big[(C\|x^\top W\|_2^2 - \|x^\top w_* q^\top\|_2^2)x x^\top\big]\|_2 + 2C^2\|\|W\|_F^2 I_d + 2WW^\top\|_2 \\
&\overset{(c)}{=} C\|(C\|W\|_F^2 - \|w_* q^\top\|_F^2)I_d + 2(CWW^\top - w_* w_*^\top)\|_2 + 2C^2\|\|W\|_F^2 I_d + 2WW^\top\|_2 \\
&\le 7C^2\|W\|_F^2 + C\|w_* q^\top\|_F^2 + 2C\|CWW^\top - w_* w_*^\top\|_2 \\
&\le (15C + 16C^2\nu^2)\|w_*\|^2,
\end{aligned} \tag{59}$$

where (a) is due to $\|I \otimes A\|_2 \le \|I\|_2\|A\|_2 = \|A\|_2$, (b,c) is due to Lemma 6, and the last inequality is by setting $W = \frac{w_*}{\sqrt{C}} q^\top + \theta H$ with a H satisfying $\|H\|_F = 1$ and $\theta := \nu\|w_*\|$.

□

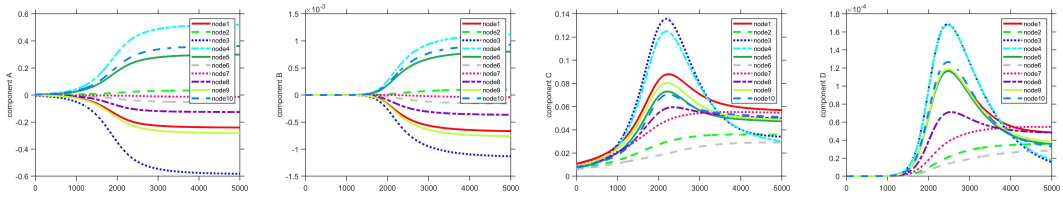
Lemma 7 and Lemma 8 together implies that when the the iterate is in the neighborhood of a global optimal solution, then gradient descent has a linear convergence rate. The proof essentially follows the same lines as the proof of Lemma 1. Hence, we omit its proof.

Lemma 9. *(locally linear convergence) Suppose that at time t_0,* $\text{dist}(W_{t_0}, w_*) := \|W_{t_0} - \frac{w_*}{\sqrt{C}} q_{t_0}^\top\| \le \nu\|w_*\|$ *where* $W_{t_0} \in \mathbb{R}^{d\times K}$, $\nu > 0$ *satisfies* $(2C - 14C^{3/2}\nu - 2C^2\nu^2) > 0$, *and* $q_{t_0} := \arg\min_{q\in\mathbb{R}^K : \|q\|_2 \le 1} \|W_{t_0} - \frac{w_*}{\sqrt{C}} q^\top\|$. *Then, gradient descent with the step size* $\eta \le \frac{2C - 14C^{3/2}\nu - 2C^2\nu^2}{(15C + 16C^2\nu^2)^2\|w_*\|^4}$ *generates the iterates* $\{W_t\}_{t\ge t_0}$ *satisfying*

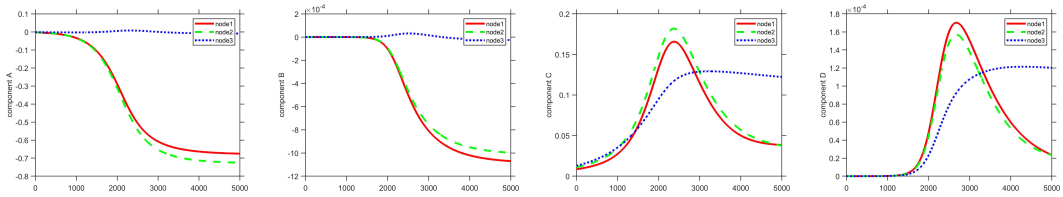
$$\text{dist}^2(W_{t+1}, w_*) \le (1 - \eta(2C - 14C^{3/2}\nu - 2C^2\nu^2))\text{dist}^2(W_t, w_*).$$

Remark: We will need ν in Lemma 9 to satisfy $(2C - 14C^{3/2}\nu - 2C^2\nu^2) > 0$ in order to show that gradient descent enters the linear convergence regime. For a fixed C, the condition translates to $\nu \le 0.14/\sqrt{C}$. Now we also see that the condition $\|w_*\| > 1.1\sqrt{C}\gamma$ in Theorem 3 is trivially satisfied when we set $2\gamma^2 = \nu^2\|w_*\|^2$, since $1 > 1.1\sqrt{C}\cdot 0.14/\sqrt{2}/\sqrt{C}$. Note that the reason why we set $2\gamma^2 = \nu^2\|w_*\|^2$ is because $\text{dist}^2(w_t, \frac{w_*}{\sqrt{C}}) = ||w_t[1]| - \|\frac{w_*}{\sqrt{C}}\||^2 + \|w_t^\perp\|_2^2 \le 2\gamma^2 = \nu^2\|w_*\|^2$ implies the iterate w_t is at the benign region that allows linear convergence when $t = T_{\gamma,C}$.

F MORE EMPIRICAL RESULTS

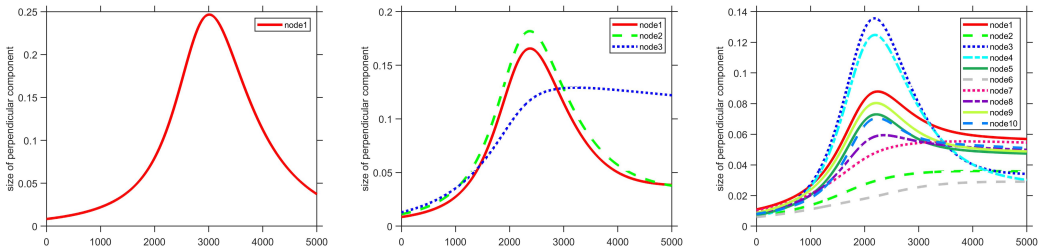


(a) Component A vs. t. (b) Component B vs. t. (c) Component C vs. t. (d) Component D vs. t.



(e) Component A vs. t. (f) Component B vs. t. (g) Component C vs. t. (h) Component D vs. t.

Figure 5: Subfigure (a) shows $w_t^{(k),\|}\big(1+\eta(3\|w_*\|^2-3\|w_t^{(k)}\|^2)\big)$ (i.e. component A of $w_{\#K,t}^{(k),\|}$) versus iteration t for each neuron k. Subfigure (b) plots $2\eta\sum_{j\neq k}^{K}((w_t^{(j)})^\top w_t^{(k)})w_t^{(j),\|}+\eta w_t^{(k),\|}\sum_{j\neq k}^{K}\|w_t^{(j)}\|^2$ (i.e. component B of $w_{\#K,t}^{(k),\|}$) versus iteration t for each neuron k. Subfigure (c) plots the norm of component C of $w_{\#K,t}^{(k),\perp}$, while subfigure (d) plots the norm of component D of $w_{\#K,t}^{(k),\perp}$ for each neuron k. Our empirical findings show that the components due to interaction of the other neurons (i.e. component B and D) are small (notice that the scale of the vertical axis of (a) and (b), (c) and (d) are different) compared to their counterparts (i.e. component A and C respectively), which suggests that $\theta, \vartheta \approxeq 10^{-4}$ on (10) empirically. Top row: number of neurons $K=10$. Bottom row: number of neurons $K=3$.



(a) $K=1$. (b) $K=3$. (c) $K=10$.

Figure 6: The perpendicular component $\|w_{\#K,t}^{(k),\perp}\|$ of each k over iterations t. We see that the perpendicular component remains small.

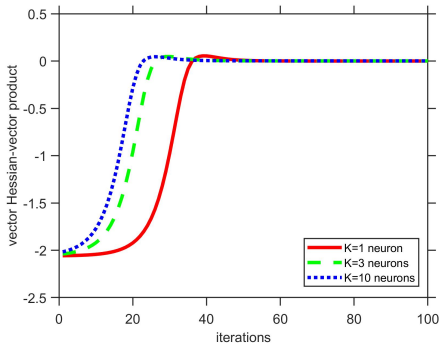
F.1 DIFFERENT STEP SIZES

Following the simulation as Figure 1, we tried different values of step sizes $\eta = \{1.0, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001\}$ for each model with different number of neurons $K = \{1, 3, 10\}$. We report the quantity (9) over iteration t,

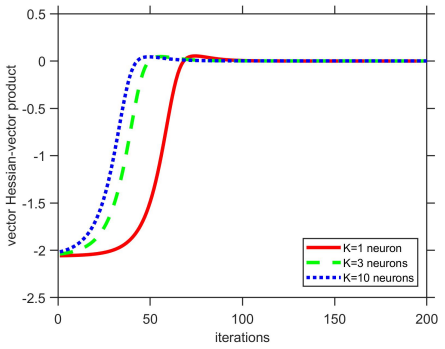
$$\text{vec}(w_* q_t^\top - W_t^{\#K})^\top \nabla^2 f(W_t^{\#K}) \text{vec}(w_* q_t^\top - W_t^{\#K}),$$

where $\nabla^2 f(W_t^{\#K}) \in \mathbb{R}^{dK \times dK}$ is the Hessian and $w_* q_t^\top$ is the closet global optimal solution to $W_t^{\#K}$ and the notation $\text{vec}(\cdot)$ represents the vectorization operation of its matrix argument. Recall that the quantity can be viewed as a measure of the strong convexity as mentioned in the main text. Specifically, if the quantity is larger than 0, then it suggests that the current optimization landscape is strongly convex with respect to $w_* q_t^\top$.

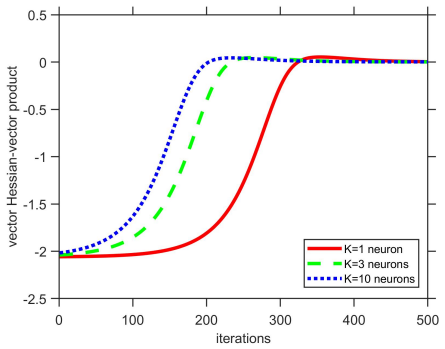
We found out that gradient descent with step size $\eta = \{1.0, 0.5, 0.4, 0.3\}$ either diverges or cannot converge towards zero testing error for all $K = \{1, 3, 10\}$, which means that $\eta = 0.2$ is basically the best step size of gradient descent for each model K. So the result suggests that even under optimal tuning of the step size η for each model K, gradient descent for an over-parametrized model converges faster than the case for a smaller model. We show some results of using different η on Figure 7, Figure 8, and Figure 9. Based on the empirical results, we conclude that the acceleration due to over-parametrization cannot be simply explained by that gradient descent uses a larger *effective* step size, as the impacts due to parameters η and K seem to be complementary in the experiment.



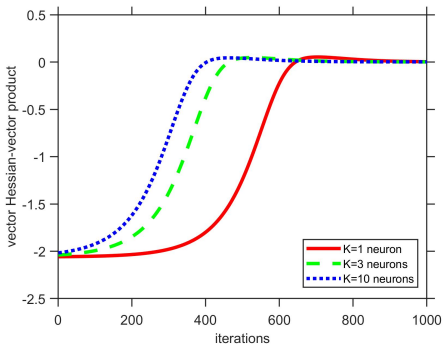
(a) ($\eta = 0.2$) quantity (9) over t.



(b) ($\eta = 0.05$) quantity (9) over t.

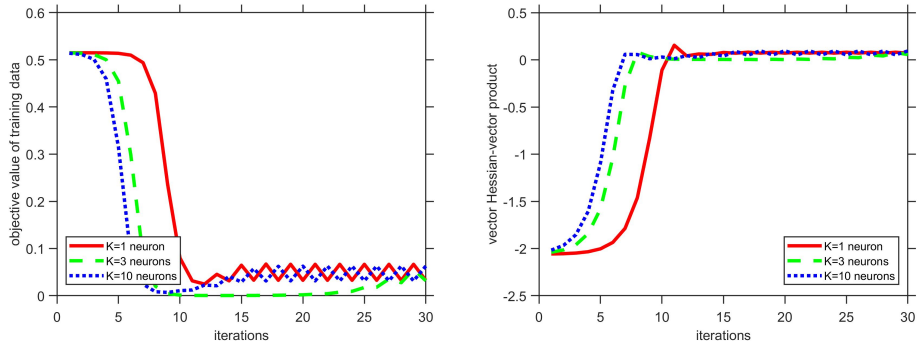


(c) ($\eta = 0.01$) quantity (9) over t.



(d) ($\eta = 0.005$) quantity (9) over t.

Figure 7: Gradient descent with different values of the step size. Note that the scales of the horizontal axes are different. Both the step size η and the degree of over-parametrization affect the time that gradient descent enters the linear convergence regime. A larger step size η and a larger number of neurons K help gradient descent to make progress faster.



(a) ($\eta = 0.5$) training error over t. (b) ($\eta = 0.5$) quantity (9) over t.

Figure 8: Gradient descent with $\eta = 0.5$. We see that gradient descent cannot converge to zero training (and testing) error. That is, the step size is too large to converge to a global optimal solution. Interestingly, we still observe that gradient descent requires fewer iterations to get closer to a global optimal point for a larger model, though it does not converge to a global optimal point using the large step size.

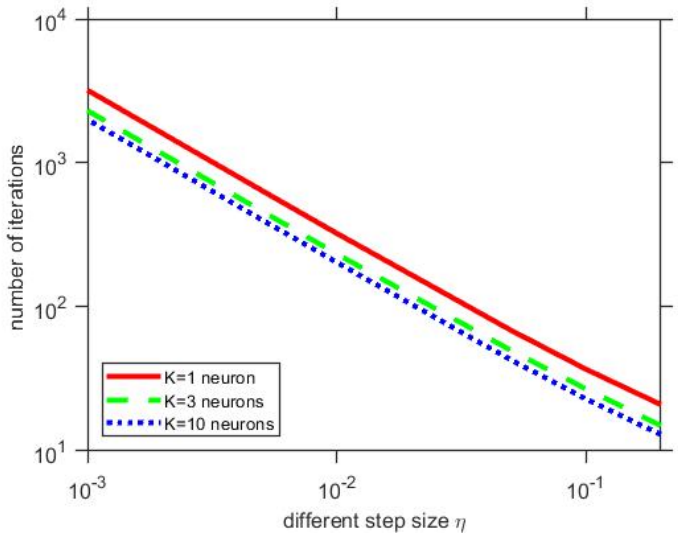


Figure 9: We plot the number of iterations required for the metric $\text{vec}(w_* q_t^\top - W_t^{\#K})^\top \nabla^2 f(W_t^{\#K}) \text{vec}(w_* q_t^\top - W_t^{\#K})$ to be positive (the y-axis) under different values of the step size η (the x-axis).