

Benefits of Jointly Training Autoencoders: An Improved Neural Tangent Kernel Analysis

Thanh V. Nguyen, Raymond K. W. Wong, Chinmay Hegde *

March 3, 2020

Abstract

A remarkable recent discovery in machine learning has been that deep neural networks can achieve impressive performance (in terms of both lower training error and higher generalization capacity) in the regime where they are massively over-parameterized. Consequently, over the past year, the community has devoted growing interest in analyzing optimization and generalization properties of over-parameterized networks, and several breakthrough works have led to important theoretical progress. However, the majority of existing work only applies to supervised learning scenarios and hence are limited to settings such as classification and regression.

In contrast, the role of over-parameterization in the *unsupervised* setting has gained far less attention. In this paper, we study the inductive bias of gradient descent for two-layer over-parameterized autoencoders with ReLU activation. We first provide theoretical evidence for the memorization phenomena observed in recent work using the property that infinitely wide neural networks under gradient descent evolve as linear models. We also analyze the gradient dynamics of the autoencoders in the finite-width setting. We make very few assumptions about the given training dataset (other than mild non-degeneracy conditions). Starting from a randomly initialized autoencoder network, we rigorously prove the linear convergence of gradient descent in two learning regimes, namely:

1. the *weakly-trained* regime where only the encoder is trained, and
2. the *jointly-trained* regime where both the encoder and the decoder are trained.

Our results indicate the considerable benefits of joint training over weak training for finding global optima, achieving a dramatic decrease in the required level of over-parameterization.

We also analyze the case of weight-tied autoencoders (which is a commonly used architectural choice in practical settings) and prove that in the over-parameterized setting, training such networks from randomly initialized points leads to certain unexpected degeneracies.

1 Introduction

Deep neural networks have achieved great success in a variety of applications such as image and speech recognition, natural language processing, and gaming AI. Remarkably, neural networks that achieve the state-of-the-art performance in each of these tasks are all massively over-parameterized, with far more weight parameters than the sample size of training data or the input dimension. Such networks can gain impressive performance in terms of both (near) zero training error and high generalization capacity, which seemingly contradicts the conventional wisdom of bias-variance tradeoffs. Surprising enough is the fact that (stochastic) gradient descent or its variants can effectively find global and generalizable solutions. Explaining this phenomenon has arguably become one of the fundamental tasks for demystifying deep learning.

*TN is with the ECE Department at Iowa State University. RW is with the Statistics Department at Texas A&M University. CH is with the Tandon School of Engineering at New York University. This work was supported in part by the National Science Foundation under grants DMS-1612985/1806063, CCF-1934904, CCF-1815101, CAREER CCF-2005804, and a faculty fellowship from the Black and Veatch Foundation.

As a consequence, there has been growing interest in understanding the power of the gradient descent for over-parameterized networks. Over the past year, a specific line of research [Li and Liang, 2018, Allen-Zhu et al., 2018, Zou et al., 2018, Du et al., 2018, Oymak and Soltanolkotabi, 2019, Arora et al., 2019a, Zou and Gu, 2019] has led to exciting theoretical progress. In particular, the seminal work of Du et al. [2018] shows that gradient descent on two-layer neural networks with ReLU activation provably converges to some global minimum at a geometric rate, provided a sufficiently large number of neurons that is of polynomial order in the sample size. The key idea that leads to this result is the following: once the network is sufficiently wide, gradient descent does not change the individual weights much, but results in a non-negligible change in the network output that exponentially reduces the training loss with iteration count. This line of thinking has been subsequently refined and linked to the stability of a special kernel, called the *neural tangent kernel* (NTK) [Jacot et al., 2018]. Arora et al. [2019a] showed that the minimum eigenvalue of the limiting kernel governs both the algorithmic convergence and the generalization performance.

Despite these exciting results, the majority of existing work has focused on *supervised* settings and hence are limited to tasks such as classification and regression. In contrast, the role of over-parameterization in the *unsupervised* setting (for tasks such as reconstruction, denoising, and visualization) has gained much less attention. An early related example in unsupervised learning can be traced back to learning over-complete dictionaries with sparse codes [Olshausen and Field, 1997]. Another example is the problem of learning mixtures of k well-separated spherical Gaussians, where Dasgupta and Schulman [2007] showed that initializing with $O(k \log k)$ centers enables expectation-maximization to correctly recover the k components.

Interesting (but limited) progress has been made towards understanding over-parameterization for autoencoders, a popular class of unsupervised models based on neural networks. Zhang et al. [2019] provided an extensive study of training highly over-parameterized autoencoders using a *single* sample. They empirically showed that when learned by gradient descent, autoencoders with different architectures can exhibit two inductive biases: memorization (i.e., learning the constant function) and generalization (i.e., learning the identity mapping) depending on the non-linearity and the network depth. Radhakrishnan et al. [2019] showed that over-parameterized autoencoder learning is empirically biased towards functions that concentrate around the training samples and hence exhibits memorization. Buhai et al. [2019] empirically showed that over-parameterization benefits learning in recovering generative models with single-layer latent variables (including the sparse coding model).

However, there has been a lack of theoretical evidence that supports these observations. Zhang et al. [2019] were able to prove a result for a simple one-layer linear case while Radhakrishnan et al. [2019] also proved the concentration of outputs near the training examples for a single-layer network under a data-restrictive setting. Moreover, none of the above papers have rigorously studied the training *dynamics* of autoencoder models. The loss surface of autoencoder training was first characterized in [Rangamani et al., 2017]. Subsequently, Nguyen et al. [2019] proved that under-parameterized (and suitably initialized) autoencoders performed (approximate) proper parameter learning in the regime of asymptotically many samples, building upon techniques in provable dictionary learning; cf. [Arora et al., 2015, Nguyen et al., 2018].

Our contributions. In this paper, we provide the first rigorous analysis of inductive bias of gradient descent and gradient dynamics of over-parameterized, shallow (two-layer) autoencoders. To examine the inductive bias, we use an infinite-width approximation to derive the output reconstruction in terms its input. For the gradient dynamics, we study different training schemes and establish upper bounds on the level of over-parameterization under which (standard) gradient descent, starting from randomly initialized weights, can linearly converge to global optima provided the training dataset obeys some mild assumptions. Our specific contributions are as follows:

1. First, we build upon the results by Lee et al. [2019] to characterize the evolution of autoencoder output via linearization and infinite-width approximation. Then, we establish the inductive bias of infinite-width autoencoders trained with gradient descent and provide insights into the memorization phenomena. While our analysis is asymptotic with respect to the network width, empirical results in [Lee et al., 2019, Zhang et al., 2019] strongly suggest that similar phenomena are exhibited at finite widths as well.

2. Next, we extend the results by Du et al. [2018] to the setting of over-parameterized two-layer autoencoders. This involves developing a version of the NTK for multiple outputs, which can be done in a straightforward manner by lifting the kernel matrix of a single output into a higher-dimensional space via Kronecker products.
3. Next, we study the gradient dynamics of the *weakly-trained*¹ case where the training is done only over the weights in the encoder layer. We obtain a bound on the number of hidden neurons (i.e., level of over-parameterization) required to achieve linear convergence of gradient descent, starting from random initialization, to global optimality.
4. Next, we study the gradient dynamics of the *jointly-trained* case where both the encoder and decoder are trained with gradient descent. We obtain a bound analogous to the weakly-trained case for the level of over-parameterization required for global convergence. Interestingly, our bound for over-parameterization in the jointly trained case is significantly better compared with the weakly-trained case.
5. Finally, we study a special family of autoencoders for which the encoder and decoder are *weight-tied*, i.e., the two layers share the same weights (this is a common architectural choice in practical applications). For the weight-tied case, we show that even without any training, $O(d/\epsilon)$ hidden units are able to achieve ϵ -test error where d is the input dimension. Indeed, as the number of hidden unit increases, the autoencoder approximately recovers an identity map. Since the identity map is not particularly useful in representation learning, we speculate that training of weight-tied autoencoders under over-parameterization may lead to unexpected degeneracies.

Techniques. Our analysis extends the techniques of Lee et al. [2019] and Du et al. [2018] for analyzing the global convergence of gradient descent in overparameterized neural networks using the neural tangent kernel. The special case of autoencoder networks is somewhat more complicated since we now have to deal with multiple outputs, but the use of Kronecker products enables us to derive concise NTK’s for our setting.

The work of Du et al. [2018] and subsequent papers study the weakly-trained case for the supervised setting where the second layer is fixed. We derive analogous bounds for the autoencoder setting. Moreover, we derive a new result for the jointly-trained case and obtain a significantly improved bound on the requisite level of over-parameterization. Our result is based on three key insights:

- (i) the linearization enables us to derive the autoencoder’s reconstruction for a given input as a linear combination of the training samples weighted by kernel scores;
- (ii) thanks to the linear decoder, the corresponding kernel is *smooth*, and the improved smoothness allows gradient descent to move greater amount from the initial point; and
- (iii) with this improved smoothness, we can derive a sharper characterization of the descent trajectory length in Frobenius norm instead of column-wise Euclidean norm.

2 Overview of main results

Notation. We use uppercase letters to denote matrices, and lowercase for vectors or scalars. An expectation is the notation C which represents a generic scalar constant, whose value can change from line to line. A vector is interpreted as a column vector by default. We denote by $x_i \in \mathbb{R}^d$ the i^{th} -column (or sample) of the data matrix X , and $W = [w_1, \dots, w_m] \in \mathbb{R}^{d \times m}$ denotes a weight matrix. Whenever necessary, we distinguish between the weight vector w_r at different algorithmic steps using an explicit $w_r(t)$ indexed by the step t . For a matrix $A = [a_1, \dots, a_m] \in \mathbb{R}^{d \times m}$, $\text{vec}(A) = [a_{11}, \dots, a_{d1}, \dots, a_{1m}, \dots, a_{dm}]^\top$ vectorizes the matrix A by stacking its columns. The symbol \otimes denotes the Kronecker product.

¹This distinction of weak- vs. joint-training has been introduced in earlier work such as Arora et al. [2019b].

We use $\mathcal{N}(\cdot)$ and $\text{Unif}(\cdot)$ to denote the Gaussian and uniform distributions respectively. We simply write \mathbb{E}_w instead of $\mathbb{E}_{w \sim \mathcal{N}(0, I)}$ for brevity. Throughout the paper, we refer to an arbitrary $\delta \in (0, 1)$ as the failure probability of some event under consideration.

2.1 Two-layer autoencoders

Our goal is to understand the inductive bias and the learning dynamics of learning two-layer autoencoders with gradient descent. We focus on the two-layer autoencoder architecture with the rectified linear unit (ReLU), defined by $\phi(z) = \max(z, 0)$ for any $z \in \mathbb{R}$. In below, when ϕ is applied to a vector or a matrix, the ReLU function is applied element-wisely. Given an input sample $x \in \mathbb{R}^d$, the autoencoder returns a reconstruction $u \in \mathbb{R}^d$ of x , given by

$$u = \frac{1}{\sqrt{md}} A \phi(W^\top x) = \frac{1}{\sqrt{md}} \sum_{r=1}^m a_r \phi(w_r^\top x),$$

where $W = [w_1, \dots, w_m]$ and $A = [a_1, \dots, a_m]$ are weight matrices of the first (encoder) and second (decoder) layers respectively. We do not consider bias terms in this work. However, in principle, the bias vector for the hidden layer can be regarded as the last column of W with the last dimension of x always being 1.

Remark 2.1 (Choice of scaling factor). Notice that we have scaled the output with $1/\sqrt{md}$, where $1/\sqrt{m}$ is the factor for the first layer and $1/\sqrt{d}$ for the second layer. Such scaling has been utilized in mathematical analyses of supervised networks [Jacot et al., 2018] as well as of autoencoders [Li and Nguyen, 2018]. Since the ReLU is homogeneous to scaling, such factors can technically be absorbed into the corresponding weight matrices W and A , but we find that keeping such factors explicit is crucial to understand the asymptotic behavior of neural network training as the network widths (i.e., m in this case) go to infinity.

Let us now set up the problem. Suppose that we are given n training samples $X = [x_1, x_2, \dots, x_n]$. We assume that each weight is randomly and independently initialized. Then, we train the autoencoder via gradient descent over the usual squared-error reconstruction loss:

$$L(W, A) = \frac{1}{2} \sum_{i=1}^n \|x_i - \frac{1}{\sqrt{md}} A \phi(W^\top x_i)\|^2 = \frac{1}{2} \sum_{i=1}^n \|x_i - u_i\|^2. \quad (2.1)$$

Throughout the paper, unless otherwise specified, we make the following assumptions:

Assumption 1. All training samples are normalized, i.e., $\|x_i\| = 1$ for $i = 1, \dots, n$.

We gather the training samples into the data matrix $X = [x_1, x_2, \dots, x_n]$ and define $\lambda_n \triangleq \|X^\top X\|$. Assumption 1 implies that $\|X\|_F = \sqrt{n}$ and hence $1 \leq \lambda_n \leq n$. We regard λ_n as a parameter that depends on the data geometry. For certain families of matrices (e.g., those with independent Gaussian entries), $\lambda_n \sim O(\max(n/d, 1))$, which can be $o(n)$ depending on how large n is in terms of d . We note that throughout our analysis, X is regarded as fixed, and we will focus on the randomness in the weights.

Assumption 2. Consider a random vector $w \sim \mathcal{N}(0, I)$ and define $\tilde{x}_i = \mathbb{1}[w^\top x_i \geq 0] x_i$ for each $i \in [n]$. Let $\tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_n]$. Assume $\min(\lambda_{\min}(\mathbb{E}_w[\tilde{X}^\top \tilde{X}]), \lambda_{\min}(\mathbb{E}_w[\phi(X^\top w) \phi(w^\top X)])) = \lambda_0 > 0$.

The matrix $\mathbb{E}_w[\tilde{X}^\top \tilde{X}]$ is the so-called Gram matrix from the kernel induced by the ReLU transformation and has been extensively studied in [Xie et al., 2016, Tsuchida et al., 2017, Du et al., 2018, Arora et al., 2019a]. Although this condition is difficult to interpret, one sufficient condition established in [Oymak and Soltanolkotabi, 2019] (Lemma H.1 and Lemma H.2) is that as long as the squared minimum singular value $\sigma_{\min}^2(X \star X) > 0$ where \star denotes the Khatri-Rao product, then Assumption 2 holds. In this sense, our assumption is similar to that of Oymak and Soltanolkotabi [2019] and slightly weaker than of Du et al. [2018], which only require $\lambda_{\min}(\mathbb{E}_w[\tilde{X}^\top \tilde{X}]) > 0$.

The above assumptions about the data are relatively mild, which are in sharp contrast with assuming a specific generative model for the data (e.g., dictionary models, mixture of Gaussians [Nguyen et al., 2019, Buhai et al., 2019]) that have so far been employed to analyze autoencoder gradient dynamics.

2.2 Learning dynamics

Depending on which weight variables are being optimized, we consider three training regimes:

- **Weakly-trained case:** This corresponds to the regime where the loss function (2.1) is optimized over the weights W while keeping A fixed. A different form of weak training is to fix the encoder and optimize (2.1) over A . Indeed, this practice is perhaps a folklore: it corresponds to standard kernel regression where the global convergence depends on the Hessian associated with random ReLU features. We do not pursue this case any further since kernel methods are well understood, but note in passing that the Hessian will eventually show up in our analysis.
- **Jointly-trained case:** This corresponds to the regime that (2.1) is optimized over both W and A . This case matches practical neural network training, and performs better than the weakly trained case. We will show that the contrast between weakly-trained and jointly-trained cases arises due to the nature of the different NTK's and our analysis may pave the way to better understanding of autoencoder training.
- **Weight-tied case:** Weight-tying is another common practice in training autoencoders. Here, one sets the encoder and decoder weights to be the same, i.e., $A = W$, and optimizes (2.1) over the common variables W . We study this problem from the perspective of over-parameterization and show that this case leads to somewhat unexpected degeneracies.

We adopt the framework introduced in [Du et al., 2018]. Our proofs proceed generally as follows:

- (i) We will consider the continuous flow of the autoencoder outputs $U(t) = [u_1(t), u_2(t), \dots, u_n(t)] \in \mathbb{R}^{d \times n}$ corresponding to the samples in X at time t . This continuous flow can be morally viewed as the execution of gradient descent with infinitesimal learning rate. This enables us to write:

$$\frac{d\text{vec}(U(t))}{dt} = K(t)\text{vec}(X - U(t)),$$

where $K(t)$ is a kernel matrix.

- (ii) From this characterization, we can infer that the spectrum of $K(t)$ governs the dynamics of the outputs. To derive explicit convergence bounds, we will first prove that $K(0)$ has positive minimum eigenvalue with high probability. This is achieved via using concentration arguments over the random initialization. Then, we will upper-bound the movement of each individual weight vector from the initial guess and hence bound the deviation of $K(t)$ from $K(0)$ in terms of spectral norm.
- (iii) By discretizing the continuous-time analysis, we will obtain analogous bounds for gradient descent with a properly chosen step size and show that gradient descent linearly converges to a global solution.

Our convergence results are informally stated in the following theorems:

Theorem 2.1 (Informal version of Theorems 5.1 and 5.2). *Consider an autoencoder that computes output $u = \frac{1}{\sqrt{md}}A\phi(W^\top x)$ where the weight vectors are initialized with independent vectors $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{-1, 1\}^d)$ for all $r \in [m]$. For any $\delta \in (0, 1)$ and $m \geq C \frac{n^5 d^4 \lambda_n}{\lambda_0^4 \delta^3}$ for some large enough constant C , the gradient descent over W linearly converges to a global minimizer with probability at least $1 - \delta$ over the randomness in the initialization.*

Theorem 2.2 (Informal version of Theorems 6.1 and 6.2). *Consider an autoencoder that computes output $u = \frac{1}{\sqrt{md}}A\phi(W^\top x)$ where the weight vectors are initialized with independent vectors $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{-1, 1\}^d)$ for all $r \in [m]$. For any $\delta \in (0, 1)$ and $m \geq C \frac{nd\lambda_n^3}{\lambda_0^4 \delta^2}$ for some large enough constant C , the gradient descent jointly over W and A linearly converges to a global minimizer with probability at least $1 - \delta$ over the randomness in the initialization.*

Comparisons with existing work. We summarize the quantitative implications of our results in Table 1. In this table, we compare with Du et al. [2018], Oymak and Soltanolkotabi [2019], Zou and Gu [2019] that achieve the best known bounds to our knowledge.

We emphasize that the factor d in our bounds arises due to the fact that our network produces high-dimensional outputs (dimension d in the case of autoencoders) while the previous works have focused on scalar outputs. Note also that the input dimension d is implicitly hidden in λ_0 and λ_n .

For weakly-trained networks with a single output, we (slightly) improve the order of over-parameterization: $m = \Omega\left(\frac{n^5 \lambda_n}{\lambda_0^4 \delta^3}\right)$ over the previous bound $\Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$ in Du et al. [2018, Theorem 3.2] by explicitly exposing the role of the spectral norm λ_n of the data.

For the jointly-trained regime, we obtain a significantly improved bound over Du et al. [2018, Theorem 3.3]. Our result is consistent with Oymak and Soltanolkotabi [2019, Theorem 6.3], but we have both layers jointly trained; the proof technique in Oymak and Soltanolkotabi [2019, Theorem 6.3] is different from ours (bounding Jacobian perturbations), and does not seem to be easily extended to the jointly trained case.

Let us better understand the intuition behind the bounds in Table 1 in terms of the dimension d and the sample size n . We emphasize that in the fairly typical regime of machine learning where $n \geq d$ and $\lambda_n \sim n/d$, the level of over-parameterization for the single output is moderate (of order n^4/d^3). Since autoencoders have an output dimension d , the factor- d in the bounds is natural in the jointly-trained case by characterizing the trajectory length by Frobenius norm. This is consistent with the result in Zou and Gu [2019]. Our bound is different from that in Zou and Gu [2019] in that we make assumption on the minimum eigenvalue λ_0 while they assume a lower bound on the sample separation Δ . A direct universal comparison between the two bounds is difficult; however, Oymak and Soltanolkotabi [2019] shows an upper bound $\lambda_0 \geq \Delta/100n^2$. Finally, we note that initializing A with i.i.d. Rademacher entries keeps our analysis in line with previous work, and an extension to Gaussian random initialization of A should be straightforward.

| Regime | Reference | Single output | Multiple output |
|----------------|---------------------------------|---|--|
| Weakly-trained | Du et al. [2018] | $C \frac{n^6}{\lambda_0^4 \delta^3}$ | \times |
| | This work | $C \frac{n^5 \lambda_n}{\lambda_0^4 \delta^3}$ | $C \frac{n^5 d^4 \lambda_n}{\lambda_0^4 \delta^3}$ |
| | Oymak and Soltanolkotabi [2019] | $C \frac{n \lambda_n^3}{\lambda_0^4}$ | \times |
| Joint-trained | Du et al. [2018] | $C \frac{n^6 \log(m/\delta)}{\lambda_0^4 \delta^3}$ | \times |
| | Zou and Gu [2019] | $C \frac{n^8}{\Delta^4}$ | $C \frac{n^8 d}{\Delta^4}$ |
| | This work | $C \frac{n \lambda_n^3}{\lambda_0^4 \delta^2}$ | $C \frac{n d \lambda_n^3}{\lambda_0^4 \delta^2}$ |

Table 1: Comparison of our over-parameterization bounds with the known results in [Du et al., 2018, Theorem 3.2 and Theorem 3.3], [Oymak and Soltanolkotabi, 2019, Theorem 6.3] and [Zou and Gu, 2019, Table 1]. Here, d is the input dimension, n is the training size, λ_0 is the smallest eigenvalue of the Gram matrix, λ_n is the maximum eigenvalue of the covariance matrix and C is some sufficiently large constant. Δ is the smallest distance between any pair of distinct training points.

2.3 Inductive bias

The following theorem establishes a result on the inductive bias of the infinitely wide autoencoders trained with gradient descent.

Theorem 2.3. *Let $K^\infty = \mathbb{E}_{W(0), A(0)}[K(0)]$. Assume $\lambda_{\min}(K^\infty) > 0$ and let $\eta_{\text{critical}} = 2(\lambda_{\max}(K^\infty) + \lambda_{\min}(K^\infty))^{-1}$. Under gradient descent with learning rate $\eta < \eta_{\text{critical}}$, for every normalized $x \in \mathbb{R}^d$ as the*

width $m \rightarrow \infty$, the autoencoder output $f_t(x)$ at step t converges to $\mu_t(x) + \gamma_t(x)$, with:

$$\begin{aligned}\mu_t(x) &\rightarrow \sum_{i=1}^n \Lambda_i x_i, \\ \gamma_t(x) &\rightarrow f_0(x) - \sum_{i=1}^n \Lambda_i f_0(x_i)\end{aligned}$$

where each $\Lambda_i \in \mathbb{R}^{d \times d}$ depends on the kernel score between the input x and each training sample x_i and K^∞ . $f_0(x)$ is the autoencoder reconstruction of x at initialization.

We prove this result in Section 3.2. Essentially, Theorem 2.3 generalizes the simple result in Zhang et al. [2019, Theorem 1] to non-linear autoencoders and multiple-sample training despite its asymptotic nature. The closer the new test input x is to the span of training data X , the more its reconstruction concentrates around these seen points. This coincides with the observation about “memorization” by Radhakrishnan et al. [2019].

3 The Neural Tangent Kernel and Linearized Autoencoders

3.1 NTK for general autoencoders

Let us first derive the neural tangent kernels for general autoencoders (possibly deep and with more than 2 layers) with multiple outputs in a compact form. Given n i.i.d samples $X = [x_1, x_2, \dots, x_n]$ and the autoencoder $f(\theta, x)$, we consider minimizing the squared-error reconstruction loss:

$$L(\theta) = \frac{1}{2} \sum_{i=1}^n \|x_i - f(\theta, x_i)\|^2 = \frac{1}{2} \sum_{i=1}^n \|x_i - u_i\|^2,$$

where θ is a vector that stacks all the network parameters (e.g. W and A) and $u_i = f(\theta, x_i) \in \mathbb{R}^d$ denotes the corresponding output for every $i = 1, 2, \dots, n$. The evolution of gradient descent on $L(\theta)$ with an infinitesimally small learning rate is represented by the following ordinary differential equation (ODE):

$$\frac{d\theta(t)}{dt} = -\nabla_\theta L(\theta(t)). \quad (3.1)$$

The time-dependent NTK for autoencoders can be characterized as follows:

Lemma 3.1. *Denote by $U(t) = [u_1(t), u_2(t), \dots, u_n(t)] \in \mathbb{R}^{d \times n}$ the corresponding outputs of all the samples in X , i.e., $u_i(t) = f(\theta(t), x_i)$. The dynamics of $U(t)$ is given by the ODE:*

$$\frac{d\text{vec}(U(t))}{dt} = K(t)\text{vec}(X - U(t)),$$

where $K(t)$ is an $nd \times nd$ positive semi-definite kernel matrix whose (i, j) -th block of size $d \times d$ is:

$$\left(\frac{\partial}{\partial \theta} f(\theta, x_i) \right) \cdot \left(\frac{\partial}{\partial \theta} f(\theta, x_j) \right)^\top.$$

Proof. Note that in the supervised learning setting with a single output, the (i, j) -th block is a single scalar equal to the inner product of two gradients. We prove this using simple calculus. The gradient of the loss over the parameters θ is

$$\nabla_\theta L(\theta) = - \sum_{i=1}^n \frac{\partial u_i^\top}{\partial \theta} (x_i - u_i),$$

where $\partial u_i / \partial \theta$ denotes the Jacobian matrix of the output vector u_i with respect to θ . Combining with (3.1), the continuous-time dynamics of the prediction for each sample $i \in [n]$ is specified as

$$\begin{aligned} \frac{du_i}{dt} &= \frac{\partial u_i}{\partial \theta} (-\nabla_{\theta} L(\theta)) \\ &= \sum_{j=1}^n \frac{\partial u_i}{\partial \theta} \frac{\partial u_j^{\top}}{\partial \theta} (x_j - u_j). \end{aligned}$$

Vectorizing $\frac{dU(t)}{dt}$, we get

$$\frac{d\text{vec}(U(t))}{dt} = K(t) \text{vec}(X - U(t)),$$

where $K(t)$ (or K) is an $nd \times nd$ matrix whose (i, j) -block is of size $d \times d$:

$$K_{i,j} = \frac{\partial u_i}{\partial \theta} \frac{\partial u_j^{\top}}{\partial \theta} = \left(\frac{\partial}{\partial \theta} f(\theta, x_i) \right) \cdot \left(\frac{\partial}{\partial \theta} f(\theta, x_j) \right)^{\top}.$$

One can easily verify that $K(t)$ is positive semi-definite. ■

If the parameters $\theta(0)$ are assumed to be stochastic, then the (deterministic) neural tangent kernel (NTK) is defined as:

$$(K^{\infty})_{i,j} = \mathbb{E}_{\theta(0)} \left[\left(\frac{\partial}{\partial \theta} f(\theta, x_i) \Big|_{\theta=\theta(0)} \right) \cdot \left(\frac{\partial}{\partial \theta} f(\theta, x_j) \Big|_{\theta=\theta(0)} \right)^{\top} \right]. \quad (3.2)$$

Note that K^{∞} is time-independent. If the network is randomly initialized and its width is allowed to grow infinitely large, $K(t)$ converges to K^{∞} , and remains constant during training. Our goal is to show that if the width is sufficiently large (not necessarily infinite), then $K(t) \approx K(0) \approx K^{\infty}$, and the gradient dynamics are governed by the spectrum of K^{∞} .

3.2 Linearized Autoencoders

While the NTK allows us to analyze the gradient dynamics of autoencoders, it does not provide a straightforward characterization of the reconstruction given any new input. This makes it difficult to reason about the inductive bias of the over-parameterization and gradient descent for autoencoders, which were empirically studied in [Zhang et al., 2019, Radhakrishnan et al., 2019]. Here, we theoretically justify these results by using linearization and infinite approximation based on the result of Lee et al. [2019].

For the autoencoder $f(\theta, x)$, we denote by $\theta(t)$ the parameter vector at time t and by $\theta(0)$ its initial value. Let us simplify the notation by denoting $f_t(x) = f(\theta(t), x)$ and $f_t(X) = [f_t(x_1), \dots, f_t(x_n)]$. Recall the training objective:

$$L(\theta) = \frac{1}{2} \sum_{i=1}^n \|x_i - f(\theta, x_i)\|^2,$$

and the gradient flow characterization of the training dynamics:

$$\frac{d\theta(t)}{dt} = -\nabla_{\theta} L(\theta(t)). \quad (3.3)$$

Consider the following linearized autoencoder via the first order Taylor expansion of $f_t(x)$ around $\theta(0)$:

$$f_t^{\text{lin}}(x) \triangleq f_0(x) + \frac{\partial f_0(x)}{\partial \theta} \cdot \omega(t).$$

Here, $\omega_t = \theta(t) - \theta(0)$ is the parameter movement from its initialization. The first term $f_0(x)$ or the initial reconstruction of x remains unchanged during training over θ whereas the second term captures the dynamics with respect to the parameters, governed by:

$$\dot{\omega}(t) = - \sum_{i=1}^n \left(\frac{\partial f_0(x_i)}{\partial \theta} \right)^\top (x_i - f_t^{\text{lin}}(x_i)), \quad (3.4)$$

$$\begin{aligned} f_t^{\text{lin}}(x) &= - \sum_{i=1}^n \frac{\partial f_0(x)}{\partial \theta} \left(\frac{\partial f_0(x_i)}{\partial \theta} \right)^\top (x_i - f_t^{\text{lin}}(x_i)). \\ &= -K_0(x, X) \text{vec}(X - f_t^{\text{lin}}(X)). \end{aligned} \quad (3.5)$$

where we denote

$$\begin{aligned} \nabla_\theta f_0(X)^\top &\triangleq \left[\left(\frac{\partial f_0(x_1)}{\partial \theta} \right)^\top, \dots, \left(\frac{\partial f_0(x_n)}{\partial \theta} \right)^\top \right], \\ K_0(x, X) &\triangleq \frac{\partial f_0(x)}{\partial \theta} \nabla_\theta f_0(X)^\top \in \mathbb{R}^{d \times nd}, \\ \mathcal{K}_0 &\triangleq \nabla_\theta f_0(X) \nabla_\theta f_0(X)^\top \in \mathbb{R}^{nd \times nd}. \end{aligned}$$

The last quantity is known as the neural tangent kernel matrix evaluated at $\theta(0)$, which is presented in the earlier section. Following from Lee et al. [2019], we have the closed form solutions for the ODEs in (3.4) and (3.5) as follows:

$$\omega(t) = -\nabla_\theta f_0(X)^\top \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) \text{vec}(X - f_0^{\text{lin}}(X)), \quad (3.6)$$

$$\text{vec}(f_t^{\text{lin}}(X)) = (I - e^{-\mathcal{K}_0 t}) \text{vec}(X) + e^{-\mathcal{K}_0 t} \text{vec}(f_0(X)). \quad (3.7)$$

Moreover, given any new input x , the linearized output is $f_t^{\text{lin}}(x) = \mu_t(x) + \gamma_t(x)$ where the signal and noise terms are given by

$$\mu_t(x) = K_0(x, X) \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) \text{vec}(X), \quad (3.8)$$

$$\gamma_t(x) = f_0(x) - K_0(x, X) \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) \text{vec}(f_0(X)). \quad (3.9)$$

These equations characterize the dynamics of reconstruction (up to scaling) for the linearized network. Now, we establish the connection between the infinitely wide autoencoder and its linearized version, and prove Theorem 2.3.

Proof of Theorem 2.3. We simply invoke Theorem 2.1 in [Lee et al., 2019] for the autoencoder case. Denote by $K^\infty = \mathbb{E}_{W(0), A(0)}[K(0)]$ the neural tangent kernel of the two-layer autoencoder. Assume $\lambda_{\min}(K^\infty) > 0$ and let $\eta_{\text{critical}} \triangleq 2(\lambda_{\max}(K^\infty) + \lambda_{\min}(K^\infty))^{-1}$. Lee et al. [2019] shows that under gradient descent with learning rate $\eta < \eta_{\text{critical}}$, for every $x \in \mathbb{R}^d$ such that $\|x\| \leq 1$, as the width $m \rightarrow \infty$, the autoencoder $f_t(x)$ converges in $f_t^{\text{lin}}(x)$ given by Equation (3.8) and Equation (3.9). ■

3.3 NTK for two-layer autoencoders

Let us now specialize to the case of two-layer autoencoders with ReLU activation. Since we consider the two training regimes, including the *weakly-trained* and *jontly-trained*, we first give the expression of a few *base* kernels whose appropriate compositions produce the final kernel for each individual case. The precise derivation of each regime is given in the next few sections.

Again, we consider the reconstruction loss:

$$L(W, A) = \frac{1}{2} \sum_{i=1}^n \|x_i - \frac{1}{\sqrt{md}} A \phi(W^\top x_i)\|^2 = \frac{1}{2} \sum_{i=1}^n \|x_i - u_i\|^2,$$

where the weights are independently initialized such that:

$$w_r(0) \sim \mathcal{N}(0, I), \quad a_r(0) \sim \text{Unif}\{-1, 1\}^d, \quad r = 1, \dots, m.$$

Here the minimization can be either over the encoder weights W , or the decoder weights A , or both W and A . Let us denote

$$\tilde{X}_r(t) = [\mathbb{1}[w_r(t)^\top x_1 \geq 0]x_1, \dots, \mathbb{1}[w_r(t)^\top x_n \geq 0]x_n].$$

If we fix A and optimize the loss $L(W, A)$ over W , we get

$$G(t) = \frac{1}{md} \sum_{r=1}^m \tilde{X}_r(t)^\top \tilde{X}_r(t) \otimes a_r a_r^\top.$$

If we fix W and optimize the loss $L(W, A)$ over A , we get

$$H(t) = \frac{1}{md} \sum_{r=1}^m \phi(X^\top w_r(t)) \phi(w_r(t)^\top X) \otimes I,$$

Writing these kernels in Kronecker product form allows us to clearly visualize the connection to the supervised learning case, and enables characterization of their spectrum. Intuitively, in the jointly-trained case, since both W and A depend on t , an invocation of the chain rule leads to the sum $G(t) + H(t)$ being the “effective” kernel that governs the dynamics.

In the infinite-width limit where $m \rightarrow \infty$, the NTKs in the corresponding training regimes reduce to compositions of the following fixed deterministic kernels:

$$\begin{aligned} G^\infty &= \mathbb{E}_{w(0), a(0)} [\tilde{X}(0)^\top \tilde{X}(0) \otimes a(0) a(0)^\top] = \mathbb{E}_{w(0)} [\tilde{X}(0)^\top \tilde{X}(0)] \otimes I_d, \\ H^\infty &= \mathbb{E}_{w(0)} \left[\phi(X^\top w(0)) \phi(w(0)^\top X) \right] \otimes I. \end{aligned}$$

Somewhat curiously, we will show that the crucial component of the *time-dependent* kernel in the jointly-trained regime, $H(t)$ (within $H(t) + G(t)$), is better-behaved than the corresponding kernel in the weakly-trained regime, $G(t)$, thanks to its better Lipschitz smoothness, even though the respective *limiting* kernels are the same. This improved smoothness allows us to derive a much better bound on kernel perturbations with respect to changing weights, and this results in a significant improvement in the level of over-parameterization (Theorem 2.2).

4 Inductive Biases of Over-parameterized Autoencoders

In principle, the training dynamics of over-parameterized autoencoders are similar to those of supervised networks. However, the generalization properties or inductive biases of the over-parameterization are different and underexplored. In this section, we rigorously analyze the observations in [Zhang et al., 2019, Radhakrishnan et al., 2019] using the results we have developed.

4.1 One-sample training

This training setting was exclusively studied in Zhang et al. [2019] with interesting insights on the memorization phenomenon and the role of the depth and width. They were able to give some theoretical evidence for their observation in a simple one-layer linear case. Using linearization, we generalize this result for non-linear networks. We particularly focus on the two-layer architecture, but the results can be extended to networks of any depth. Although our result is asymptotic, Lee et al. [2019] showed that networks with finite, large width exhibit the same inductive bias.

Suppose we have access to only one sample x or the training data $X = x$. For a test input x' , $f_t^{\text{lin}}(x') = \mu_t(x') + \gamma_t(x')$ where

$$\mu_t(x') = K_0(x', x) \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) x, \quad (4.1)$$

$$\gamma_t(x') = f_0(x') - K_0(x', x) \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) f_0(x). \quad (4.2)$$

As the learning rate for gradient descent is sufficiently small, the autoencoder output $f_t(x') \rightarrow f_t^{\text{lin}}(x')$ as $m \rightarrow \infty$. In the infinite-width limit, the neural tangent kernel at $t = 0$ converges to:

$$\begin{aligned} \mathcal{K}_0 &= \frac{1}{md} \sum_{r=1}^m \mathbb{1}[w_r(0)^\top x \geq 0] a_r(0) a_r(0)^\top + \phi(w_r(0)^\top x)^2 I \\ &\rightarrow \frac{1}{d} \mathbb{E}[\mathbb{1}[w_r(0)^\top x \geq 0] a_r(0) a_r(0)^\top + (w_r(0)^\top x)^2 I] = \frac{I}{d}, \end{aligned}$$

since $w_r(0), a_r(0)$ are independent and $w_r(0)^\top x \sim \mathcal{N}(0, 1)$. Therefore, the reconstruction is governed by the similarity between x' and x via the kernel function. Specifically,

$$\begin{aligned} \mu_t(x') &\rightarrow d \cdot K_0(x', x) (1 - e^{-t/d}) x, \\ \gamma_t(x') &\rightarrow f_0(x') - d \cdot K_0(x', x) (1 - e^{-t/d}) f_0(x). \end{aligned}$$

Moreover, as $m \rightarrow \infty$,

$$\begin{aligned} K_0(x', x) &= \frac{\partial f_0(x')}{\partial \theta} \left(\frac{\partial f_0(x)}{\partial \theta} \right)^\top \\ &= \frac{1}{md} \sum_{r=1}^m \mathbb{1}[w_r(0)^\top x' \geq 0, w_r(0)^\top x \geq 0] x'^\top x a_r(0) a_r(0)^\top + \phi(w_r(0)^\top x') \phi(w_r(0)^\top x) I \\ &\rightarrow \frac{1}{d} \mathbb{E}_w[\mathbb{1}[w^\top x' \geq 0, w^\top x \geq 0] x'^\top x] + \mathbb{E}_w[\phi(w^\top x') \phi(w^\top x) I] \\ &= \langle x', x \rangle \frac{\pi - \arccos(\langle x', x \rangle)}{\pi d} \cdot I + \frac{1}{2\pi d} \sqrt{1 - \langle x', x \rangle^2} I. \end{aligned}$$

When x' is close to x , $K_0(x', x) \sim I/d$, the signal term $\mu_t(x') \approx d \cdot K_0(x', x) x \approx x$ dominates the zero-mean noise term $\gamma_t(x') \approx f_0(x') - f_0(x)$, then the reconstruction is close to x that explains the memorization. When x' is far from x , $\mu_t(x') \sim 0$ while $\gamma_t(x')$ is a random, so the reconstruction is governed by a random noise. See [Zhang et al., 2019] for more details on the empirical evidence.

4.2 Multiple-sample training

For the training with many samples, Radhakrishnan et al. [2019] showed that overparameterized autoencoders exhibit memorization by learning functions that concentrate near the training examples. They proved that single-layer autoencoders project data onto the span of the training examples. We provide another intuition based on the reconstruction of the linearized networks. For an arbitrary input x' ,

$$\begin{aligned} \mu_t(x') &= K_0(x', X) \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) \text{vec}(X), \\ \gamma_t(x') &= f_0(x') - K_0(x', X) \mathcal{K}_0^{-1} (I - e^{-\mathcal{K}_0 t}) \text{vec}(f_0(X)). \end{aligned}$$

The signal part of the reconstruction is a linear combination of training samples weighted by the kernel $K_0(x', x_i)$ and the eigenvalues of the kernel matrix \mathcal{K}_0 . Therefore, as $m \rightarrow \infty$ and t is sufficiently large,

$$\begin{aligned} \mu_t(x') &\rightarrow \sum_{i=1}^n d \cdot K_0(x', x_i) x_i, \\ \gamma_t(x) &\rightarrow f_0(x) - \sum_{i=1}^n d \cdot K_0(x', x_i) x_i f_0(x_i). \end{aligned}$$

The closer the new test input x' is to the span of training data X , the more its reconstruction concentrates around these seen points. This coincides with the observation about “memorization” by Radhakrishnan et al. [2019].

5 Weakly-trained Autoencoders

We now analyze various training regimes; these will follow from different compositions of the above NTK’s. In each of the analyses, we will first set up the corresponding NTK, study the gradient dynamics with infinitesimal step size (gradient flow), and then appropriately discretize the flow to get our final results.

5.1 Gradient flow

Consider the weakly-trained regime with the objective function:

$$L(W) = \frac{1}{2} \sum_{i=1}^n \|x_i - \frac{1}{\sqrt{md}} A \phi(W^\top x_i)\|^2, \quad (5.1)$$

where the corresponding minimization is *only* performed over W . Suppose that the weight matrices W and A are randomly initialized such that

$$w_{ij}(0) \sim \mathcal{N}(0, 1), \quad a_{ij}(0) \sim \text{Unif}(\{\pm 1\})$$

are drawn independently for each all (i, j) . After the initialization, we keep A fixed throughout and apply gradient descent learning over W with step size η :

$$W(k+1) = W(k) - \eta \nabla_W L(W(k)), \quad k = 0, 1, 2, \dots$$

Let us derive the neural tangent kernel for this training regime. We first calculate the gradient of $L(W)$ with respect to W . Since $A \phi(W^\top x) = \sum_{r=1}^m a_r \phi(w_r^\top x)$ for any $x \in \mathbb{R}^d$, it is convenient to compute the gradient with respect to each column w_r . The gradient $\nabla_{w_r} L(W)$ of the loss in (5.1) over w_r is given by:

$$\nabla_{w_r} L = - \sum_{i=1}^n J_r(u_i)^\top (x_i - u_i) = - \frac{1}{\sqrt{md}} \sum_{i=1}^n \mathbb{1}[w_r^\top x_i \geq 0] a_r x_i^\top (x_i - u_i), \quad (5.2)$$

where $J_r(u_i)^2$ denotes the Jacobian matrix of the output vector u_i with respect to w_r :

$$J_r(u_i) = \frac{1}{\sqrt{md}} a_r x_i^\top \phi'(w_r^\top x) = \frac{1}{\sqrt{md}} \mathbb{1}[w_r^\top x_i \geq 0] a_r x_i^\top. \quad (5.3)$$

Let us consider the gradient flow for the weight vector $w_r(t)$ via the following ODE:

$$\frac{dw_r(t)}{dt} = -\nabla_{w_r} L(W(t)). \quad (5.4)$$

Using (5.2) and (5.4), the continuous-time dynamics of the prediction for each sample $i \in [n]$ is:

$$\frac{du_i}{dt} = \sum_{j=1}^n \left(\sum_{r=1}^m J_r(u_i) J_r^\top(u_j) \right) (x_j - u_j).$$

Vectorizing $\frac{dU(t)}{dt}$, we get the equation that characterizes the dynamics of $U(t)$:

$$\frac{d\text{vec}(U(t))}{dt} = \frac{1}{d} K(t) \text{vec}(X - U(t)), \quad (5.5)$$

²Note that $\phi(z)$ is differentiable everywhere except at $z = 0$, at which the derivative will be considered as 0.

where $K(t)$ is the $nd \times nd$ matrix whose (i, j) -block is of size $d \times d$ and defined as

$$K(t)_{i,j} = d \sum_{r=1}^m J_r(u_i) J_r^\top(u_j) = \frac{1}{m} \sum_{r=1}^m \mathbb{1}[w_r(t)^\top x_i \geq 0, w_r(t)^\top x_j \geq 0] x_i^\top x_j a_r a_r^\top.$$

If we denote

$$\tilde{X}_r(t) = [\mathbb{1}[w_r(t)^\top x_1 \geq 0]x_1, \dots, \mathbb{1}[w_r(t)^\top x_n \geq 0]x_n].$$

then we can write $K(t)$ in Kronecker form:

$$K(t) = \frac{1}{m} \sum_{r=1}^m \tilde{X}_r(t)^\top \tilde{X}_r(t) \otimes a_r a_r^\top.$$

Since $W(0)$ and $A(0)$ are randomly initialized, in the limit as $m \rightarrow \infty$, $K(0)$ converges to the NTK:

$$\begin{aligned} K^\infty &= \mathbb{E}_{W(0), A(0)}[K(0)] \\ &= \mathbb{E}_{w(0), a(0)}[\tilde{X}(0)^\top \tilde{X}(0) \otimes a(0)a(0)^\top] \\ &= \mathbb{E}_w[\tilde{X}^\top \tilde{X}] \otimes I, \end{aligned}$$

where the last step follows from the independence of $w(0)$ and $a(0)$.

By Assumption 2, $\lambda_{\min}(K^\infty) = \lambda_{\min}(\mathbb{E}_w[\tilde{X}^\top \tilde{X}]) = \lambda_0 > 0$. In other words, the NTK kernel is strictly positive definite. We want to bound the minimum eigenvalue of $K(0)$ at the initialization $W(0)$ and prove $K(t) \approx K(0) \approx K^\infty$ when m is large enough.

Now, we state the main theorem for the convergence of the gradient flow:

Theorem 5.1 (Linear convergence of gradient flow, weakly-trained regime). *Suppose Assumptions 1 and 2 hold. Suppose at initialization that the weights are independently drawn such that $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{\pm 1\}^d)$ for all $r \in [m]$. If $m \geq C \frac{n^5 d^4 \lambda_n}{\lambda_0^4 \delta^2}$ for a constant $C > 0$, then with probability at least $1 - \delta$*

$$\|X - U(t)\|_F^2 \leq \exp\left(-\frac{\lambda_0 t}{d}\right) \|X - U(0)\|_F^2.$$

To prove this theorem, we use the auxiliary results from Lemmas 5.1, 5.2 and 5.3.

Lemma 5.1. *For any $\delta \in (0, 1)$, if $m \geq C \frac{\lambda_n^2 d \log(nd/\delta)}{\lambda_0^2}$ for some large enough constant C , then with probability at least $1 - \delta$, one obtains $\|K(0) - K^\infty\| \leq \lambda_0/4$ and $\lambda_{\min}(K(0)) \geq 3\lambda_0/4$.*

The proof of this Lemma is given in Appendix B.

Remark 5.1. Compared with the results in [Du et al., 2018, Song and Yang, 2019], our bound exposes the dependence on the data X through the spectral norm of X and the dimension d . When λ_n is much smaller than n , our bound improves over these aforementioned results. For example, if the training samples are drawn from certain distributions (e.g., Gaussians, or from sparsely used dictionary models), the bound can be as low as $m \sim \tilde{O}(d)$.

The next step in our analysis is to upper bound the spectral norm of the kernel perturbation, $\|K(t) - K(0)\|$, with high probability.

Lemma 5.2. *Suppose $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{\pm 1\}^d)$ are drawn independently for all $r \in [m]$. For any $\delta \in (0, 1)$ and some $R > 0$, with probability at least $1 - \delta$:*

$$\sup_{\{\tilde{w}=(\tilde{w}_1, \dots, \tilde{w}_m): \|\tilde{w}_r - w_r\| \leq R, \forall r \in [m]\}} \|K(\tilde{w}) - K(w)\| < \frac{2n^2 d R}{\delta}, \quad (5.6)$$

where $K(w) = \frac{1}{m} \sum_{r=1}^m \tilde{X}(w_r)^\top \tilde{X}(w_r) \otimes a_r a_r^\top$.

Remark 5.2. One may ask why not to directly bound $K(t) - K(0)$ for each time t but need the supremum over the ball near each w_r . Basically, since $w(t)$ depends on $W(0)$ and $A(0)$, directly working on $K(t) - K(0)$ is difficult. The uniform bound (5.6) allows us to overcome this dependence when applied to $K(t) - K(0)$.

Note that in this lemma we use $K(w)$ to indicate that the kernel K is being evaluated at the weight vectors w_r and ignore the time index t . In this Lemma, we use $\tilde{X}(w_r)$ to denote \tilde{X}_r evaluated at w_r .

Proof. For simplicity of notation, we use $\sup_{\tilde{w}}$ to represent the supremum in (5.6), and $\sup_{\tilde{w}_r}$ to represent $\sup_{\{\tilde{w}_r: \|\tilde{w}_r - w_r\| \leq R\}}$. To prove this lemma, we work on the Frobenius norm instead of the spectral norm. Let us first write

$$z_{ijr} = \mathbb{1}[\tilde{w}_r^\top x_i \geq 0, \tilde{w}_r^\top x_j \geq 0] - \mathbb{1}[w_r^\top x_i \geq 0, w_r^\top x_j \geq 0].$$

Next,

$$\begin{aligned} \|K(\tilde{w}) - K(w)\|_F^2 &\leq \|K(\tilde{w}) - K(w)\|_F^2 = \frac{1}{m^2} \sum_{i,j=1}^n \|x_i^\top x_j \sum_{r=1}^m z_{ijr} a_r a_r^\top\|_F^2 \\ &\leq \frac{1}{m^2} \left(\sum_{i,j=1}^n \left\| \sum_{r=1}^m z_{ijr} a_r a_r^\top \right\|_F \right)^2. \end{aligned}$$

The last step follows from the fact that $|x_i^\top x_j| \leq 1$ due to Cauchy-Schwartz. Therefore,

$$\begin{aligned} \sup_{\tilde{w}} \|K(\tilde{w}) - K(w)\| &\leq \frac{1}{m} \sup_{\tilde{w}} \sum_{i,j=1}^n \left\| \sum_{r=1}^m z_{ijr} a_r a_r^\top \right\|_F \\ &\leq \frac{1}{m} \sup_{\tilde{w}} \sum_{i,j=1}^n \sum_{r=1}^m |z_{ijr}| \|a_r a_r^\top\|_F \\ &\leq \frac{d}{m} \sum_{i,j=1}^n \sum_{r=1}^m \sup_{\tilde{w}_r} |z_{ijr}|, \end{aligned}$$

since $\|a_r a_r^\top\|_F = \|a_r\|^2 = d$. Now we take expectation over the random vector w_r 's on both sides:

$$\mathbb{E}_w [\sup_{\tilde{w}_r} \|K(\tilde{w}) - K(w)\|] \leq \frac{d}{m} \sum_{i,j=1}^n \sum_{r=1}^m \mathbb{E}_w [\sup_{\tilde{w}_r} |z_{ijr}|].$$

Next, we bound $\mathbb{E}_w [\sup_{\tilde{w}_r} |z_{ijr}|]$. By definition of z_{ijr} ,

$$\begin{aligned} |z_{ijr}| &= |\mathbb{1}[w_r^\top x_i \geq 0, w_r^\top x_j \geq 0] - \mathbb{1}[\tilde{w}_r^\top x_i \geq 0, \tilde{w}_r^\top x_j \geq 0]| \\ &\leq |\mathbb{1}[w_r^\top x_i \geq 0] - \mathbb{1}[\tilde{w}_r^\top x_i \geq 0]| + |\mathbb{1}[w_r^\top x_j \geq 0] - \mathbb{1}[\tilde{w}_r^\top x_j \geq 0]| \\ &\leq \mathbb{1}[|w_r^\top x_i| \leq R] + \mathbb{1}[|w_r^\top x_j| \leq R]. \end{aligned} \tag{5.7}$$

The last step follows from the results in [Du et al., 2018, Lemma 3.2]. So we get

$$\begin{aligned} \mathbb{E}_w [\sup_{\tilde{w}_r} |z_{ijr}|] &\leq \mathbb{E}_w [\mathbb{1}[|w_r^\top x_i| \leq R] + \mathbb{1}[|w_r^\top x_j| \leq R]] \\ &= 2\mathbb{P}_{z \sim \mathcal{N}(0,1)}[|z| \leq R] \leq \frac{4R}{\sqrt{2\pi}} < 2R. \end{aligned}$$

Therefore,

$$\mathbb{E}_w [\sup_{\tilde{w}} \|K(\tilde{w}) - K(w)\|] < 2n^2 dR.$$

Finally, by Markov's inequality, with probability at least $1 - \delta$:

$$\sup_{\tilde{w}} \|K(\tilde{w}) - K(w)\| < \frac{2n^2 d R}{\delta}.$$

■

Corollary 5.1. *Suppose $\|w_r(t) - w_r(0)\| \leq R \triangleq \frac{\lambda_0 \delta}{8n^2 d}$ for all $r \in [m]$ and $t \geq 0$ with probability at least $1 - \delta$. We have*

$$\lambda_{\min}(K(t)) > \frac{\lambda_0}{2}$$

with probability at least $1 - 3\delta$ if $m \geq C \frac{\lambda_n^2 d \log(nd/\delta)}{\lambda_0^2}$.

Proof. This is the direct consequence of Lemma 5.1 and Lemma 5.2. Since $\|w_r(t) - w_r(0)\| \leq R = \frac{\lambda_0 \delta}{8n^2 d}$ with probability at least $1 - \delta$ for all $t \geq 0$, then

$$\|K(t) - K(0)\| < 2n^2 d R \delta = \frac{\lambda_0}{4}$$

with probability at least $1 - 2\delta$. Using Weyl's inequality, we can bound:

$$\lambda_{\min}(K(t)) \geq \lambda_{\min}(K(0)) - \|K(t) - K(0)\| > \lambda_0/2$$

with probability at least $1 - 3\delta$ if $m \geq C \frac{\lambda_n^2 d \log(nd/\delta)}{\lambda_0^2}$ as stated in Lemma 5.1.

■

In what follows, we show that $\|w_r(t) - w_r(0)\| \leq R$ with high probability if m is sufficiently large.

Lemma 5.3. *Fix $t > 0$. Suppose $\lambda_{\min}(K(s)) \geq \lambda_0/2$ for all $0 \leq s < t$. Then,*

$$\|X - U(s)\|_F^2 \leq \exp\left(-\frac{\lambda_0 s}{d}\right) \|X - U(0)\|_F^2.$$

Also, for each $r = 1, 2, \dots, m$:

$$\|w_r(t) - w_r(0)\| \leq \frac{d\sqrt{\lambda_n} \|X - U(0)\|_F}{\sqrt{m} \lambda_0} \triangleq R'.$$

Proof. For all $s \in [0, t)$, we have

$$\begin{aligned} \frac{d}{ds} \|\text{vec}(X - U(s))\|_2^2 &= -2 \text{vec}(X - U(s))^\top \frac{1}{d} K(s) \text{vec}(X - U(s)) \\ &\leq -\frac{2}{d} \lambda_{\min}(K(s)) \|\text{vec}(X - U(s))\|^2 \\ &\leq -\frac{\lambda_0}{d} \|X - U(s)\|_F^2 \end{aligned}$$

by the assumption $\lambda_{\min}(K(s)) \geq \lambda_0/2$. Therefore, the loss at time s is upper-bounded by

$$\begin{aligned} \|X - U(s)\|_F^2 &= \|\text{vec}(X - U(s))\|^2 \\ &\leq \exp\left(-\frac{\lambda_0 s}{d}\right) \|\text{vec}(X - U(0))\|^2 \\ &\leq \exp\left(-\frac{\lambda_0 s}{d}\right) \|X - U(0)\|_F^2, \end{aligned} \tag{5.8}$$

which decays exponentially with time s at rate λ_0/d .

To upper bound the movement of the weights $\|w_r(t) - w_r(0)\|$, we use the above result while expanding the derivative of $w_r(s)$ over time $0 \leq s < t$:

$$\begin{aligned}
\left\| \frac{d}{ds} w_r(s) \right\| &= \left\| -\nabla_{w_r} L(W(s)) \right\| \\
&= \left\| \frac{1}{\sqrt{md}} \sum_{i=1}^n \mathbb{1}[w_r^\top x_i \geq 0] x_i a_r^\top (x_i - u_i(s)) \right\| \\
&= \left\| \frac{1}{\sqrt{md}} \tilde{X}_r (X - U(s))^\top a_r \right\| \\
&\leq \frac{\|X\| \|a_r\|}{\sqrt{md}} \|X - U(s)\|_F \\
&\leq \sqrt{\frac{\lambda_n}{m}} \exp(-\lambda_0 s/d) \|X - U(0)\|_F,
\end{aligned}$$

where the last step follows from $\|a_r\|^2 = d$, $\|X\|^2 = \lambda_n$ and Eq. (5.8). From the differential equation, $w_r(s)$ is continuous for all $s \in [0, t]$, and so is $\|w_r(s) - w_r(0)\|$. Consequently, we can take the limit for $t' \rightarrow t$:

$$\begin{aligned}
\|w_r(t) - w_r(0)\|_2 &= \lim_{t' \rightarrow t} \|w_r(t') - w_r(0)\|_2 \leq \lim_{t' \rightarrow t} \int_0^{t'} \left\| \frac{d}{ds} w_r(s) \right\| ds \\
&\leq \lim_{t' \rightarrow t} \int_0^{t'} \frac{\sqrt{\lambda_n} \exp(-\lambda_0 s/d) \|X - U(0)\|_F}{\sqrt{m}} ds \\
&\leq \frac{d\sqrt{\lambda_n} \|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R',
\end{aligned}$$

since $\exp(-\lambda_0 s/d)$ is continuous at $s = t$. Therefore, we finish the proof. ■

Lemma 5.4. *If $R' < R$, then $\lambda_{\min}(K(t)) \geq \frac{1}{2}\lambda_0$ for all $t \geq 0$. Moreover, $\|w_r(t) - w_r(0)\| \leq R'$ and $\|X - U(t)\|_F^2 \leq \exp(-\frac{\lambda_0 t}{d}) \|X - U(0)\|_F^2$ for all $r \in [m]$.*

Proof. We will prove this by contradiction. Assume the conclusion does not hold, meaning there exists t_0 such that:

$$t_0 = \inf \{t > 0 : \lambda_{\min}(K(t)) \leq \lambda_0/2\}.$$

We will argue that $t_0 > 0$ using the continuity. Since $w_r(t)$ is continuous in t , $K(t)$ and $\lambda_{\min}(K(t))$ are also continuous. Therefore, there exists $t' > 0$ such that for any $0 < \epsilon < \lambda_0/4$ we have

$$\lambda_{\min}(K(t')) > \lambda_{\min}(K(0)) - \epsilon > \lambda_0/2.$$

Since $t_0 > 0$, then for any $0 \leq s < t_0$, $\lambda_{\min}(K(s)) \geq \lambda_0/2$. By Lemma 5.3, we have for all $r \in [m]$:

$$\|w_r(t_0) - w_r(0)\| \leq R' < R.$$

Corollary 5.1 implies that $\lambda_0(K(t_0)) > \lambda_0/2$, which is a contradiction.

Therefore, we have proved the first part. For the second part, we have for all $t \geq 0$, $\lambda_{\min}(K(t)) \geq \frac{1}{2}\lambda_0$ and it follows from Lemma 5.3 that: $\|w_r(t) - w_r(0)\| \leq R'$ for all $r \in [m]$ and $\|X - U(t)\|_F^2 \leq \exp(-\frac{\lambda_0 t}{d}) \|X - U(0)\|_F^2$. ■

Now, we bound $\|X - U(0)\|_F$ to upper bound R' .

Claim 5.1. For any $\delta \in (0, 1)$, then $\|X - U(0)\|_F^2 \leq \frac{2n}{\delta}$ with probability at least $1 - \delta$.

Proof. We prove this using Markov's inequality. We use the independence between $A(0)$ and $W(0)$ to derive expressions for the expectation. In this proof, the expectations are evaluated over $W(0)$ and $A(0)$.

$$\begin{aligned}
\mathbb{E}[\|X - U(0)\|_F^2] &= \|X\|_F^2 + \frac{1}{md} \mathbb{E}[\|A(0)\phi(W(0)^T X)\|_F^2] \\
&= n + \frac{1}{md} \mathbb{E}[\text{trace}(\phi(X^T W(0))A(0)\phi(W(0)^T X)] \\
&= n + \frac{1}{md} \text{trace}(\mathbb{E}[\phi(X^T W(0))A(0)^T A(0)\phi(W(0)^T X)]) \\
&= n + \frac{1}{m} \text{trace}(\mathbb{E}[\phi(X^T W(0))\phi(W(0)^T X)]) \\
&= n + \sum_{i=1}^n \mathbb{E}_w[\phi(w^\top x_i)^2] \\
&= n + n \mathbb{E}_{z \in \mathcal{N}(0,1)}[z^2 \mathbb{1}[z \geq 0]] = \frac{3n}{2},
\end{aligned}$$

where in the fourth step we use $\mathbb{E}[A(0)^T A(0)] = dI$, and in the last step we use the independence of the columns of $W(0)$. Using Markov, we get:

$$\|X - U(0)\|_F^2 \leq \frac{2n}{\delta}$$

with probability at least $1 - \delta$. ■

Proof of Theorem 5.1. If the following condition holds

$$R' = \frac{d\sqrt{\lambda_n}\|X - U(0)\|_F}{\sqrt{m}\lambda_0} \leq R = \frac{\delta\lambda_0}{8n^2d},$$

then Lemma 5.3 follows. Using the condition with the bound $\|X - U(0)\|_F \leq \sqrt{2n/\delta}$ in Claim 5.1, we obtain $m = \Omega\left(\frac{n^5 d^4 \lambda_n}{\lambda_0^4 \delta^3}\right)$. This bound dominates the order of m required for the concentration of $K(0)$ in the Corollary 5.1, and therefore Theorem 5.1 follows. ■

5.2 Gradient descent

The above result for gradient flow can be viewed as a convergence rate for gradient descent in the weakly-trained regime with infinitesimally small step size. We now derive a convergence rate for gradient descent with finite step sizes.

Theorem 5.2. Suppose Assumptions 1 and 2 hold. The initial weights are independently drawn such that $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{\pm 1\}^d)$ for all $r \in [m]$. If $m \geq C \frac{n^5 d^4 \lambda_n}{\lambda_0^4 \delta^3}$ for some large enough constant C , then with probability at least $1 - \delta$ the gradient descent on W with step size $\eta = \Theta(\frac{\lambda_0}{nd\lambda_n})$,

$$\|X - U(k)\|_F^2 \leq \left(1 - \frac{\eta\lambda_0}{2d}\right)^k \|X - U(0)\|_F^2 \quad (5.9)$$

for $k = 0, 1, \dots$

We will prove Theorem 5.2 by induction. The base case when $k = 0$ is trivially true. Assume Eq. (5.9) holds for $k' = 0, 1, \dots, k$, then we show it holds for $k' = k + 1$. To this end, we first prove $\|w_r(k+1) - w_r(0)\|$ is small enough; then we use that property to bound $\|X - U(k+1)\|_F^2$.

Lemma 5.5. *If (5.9) holds for $k' = 0, 1, \dots, k$, then we have for all $r \in [m]$,*

$$\|w_r(k+1) - w_r(0)\| \leq \frac{4d\sqrt{\lambda_n}\|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R'.$$

Proof. We use the expression of the gradient in (5.2), which is:

$$\begin{aligned} \nabla_{w_r} L(W(k)) &= - \sum_{i=1}^n \frac{1}{\sqrt{md}} \mathbb{1}[w_r(k)^T x_i \geq 0] x_i a_r^\top (x_i - u_i(k)) \\ &= - \frac{1}{\sqrt{md}} \tilde{X}_r(k) (X - U(k))^\top a_r. \end{aligned}$$

Then, the difference of the weight vector w_r is:

$$\begin{aligned} \|w_r(k+1) - w_r(0)\| &= \eta \left\| \sum_{k'=0}^k \nabla_{w_r} L(w_r(k')) \right\| \\ &= \eta \left\| \sum_{k'=0}^k \frac{1}{\sqrt{md}} \tilde{X}_r(k') (X - U(k'))^\top a_r \right\| \\ &\leq \eta \frac{\|X\|}{\sqrt{md}} \sum_{k'=0}^k \|\text{vec}(X - U(k'))\|_F \|a_r\| \\ &\leq \eta \frac{\sqrt{\lambda_n}}{\sqrt{m}} \sum_{k'=0}^k \left(1 - \frac{\eta\lambda_0}{2}\right)^{k'/2} \|X - U(0)\|_F \\ &\leq \eta \frac{\sqrt{\lambda_n}}{\sqrt{m}} \|X - U(0)\|_F \sum_{k'=0}^{\infty} \left(1 - \frac{\eta\lambda_0}{2d}\right)^{k'/2} \\ &= \eta \frac{\sqrt{\lambda_n}}{\sqrt{m}} \|X - U(0)\|_F \frac{1}{\eta\lambda_0/(4d)} \\ &= \frac{4d\sqrt{\lambda_n}\|X - U(0)\|_F}{\sqrt{m}\lambda_0}, \end{aligned}$$

where the third step and the fourth step follow from the facts that $\|\tilde{X}_r(k')\| \leq \|X\| = \sqrt{\lambda_n}$ and $\|a_r\| = \sqrt{d}$. The last step follows because $\sum_{i=0}^{\infty} (1 - \eta\lambda_0/2)^{i/2} \leq \frac{4d}{\eta\lambda_0}$. ■

Now, let us derive the form of $X - U(k+1)$. First, we compute the difference of the prediction between two consecutive steps, similar to deriving $\frac{du_i(t)}{dt}$. For each $i \in [n]$, we have

$$\begin{aligned} u_i(k+1) - u_i(k) &= \frac{1}{\sqrt{md}} \sum_{r=1}^m a_r (\phi(w_r(k+1)^\top x_i) - \phi(w_r(k)^\top x_i)) \\ &= \frac{1}{\sqrt{md}} \sum_{r=1}^m a_r \left(\phi \left(\left(w_r(k) - \eta \nabla_{w_r} L(W(k)) \right)^\top x_i \right) - \phi(w_r(k)^\top x_i) \right). \end{aligned} \quad (5.10)$$

We split the right hand side into two parts: $v_{1,i}$ represents the terms that the activation pattern does not change and $v_{2,i}$ represents the remaining term that pattern may change. Formally speaking, for each $i \in [n]$, we define

$$S_i = \{r \in [m] : \mathbb{1}[w_r(k+1)^\top x_i \geq 0] = \mathbb{1}[w_r(k)^\top x_i \geq 0]\}, \text{ and } S_i^\perp = [m] \setminus S_i.$$

Then, we can formally define $v_{1,i}$ and $v_{2,i}$ as follows:

$$\begin{aligned} v_{1,i} &\triangleq \frac{1}{\sqrt{md}} \sum_{r \in S_i} a_r \left(\phi \left(\left(w_r(k) - \eta \nabla_{w_r} L(W(k)) \right)^\top x_i \right) - \phi(w_r(k)^\top x_i) \right), \\ v_{2,i} &\triangleq \frac{1}{\sqrt{md}} \sum_{r \in S_i^\perp} a_r \left(\phi \left(\left(w_r(k) - \eta \nabla_{w_r} L(W(k)) \right)^\top x_i \right) - \phi(w_r(k)^\top x_i) \right). \end{aligned}$$

We write $v_1 = (v_{1,1}^\top, v_{1,2}^\top, \dots, v_{1,n}^\top)^\top$ and do the same for v_2 , so

$$\text{vec}(U(k+1) - U(k)) = v_1 + v_2.$$

In order to analyze $v_1 \in \mathbb{R}^n$, we define K and $K^\perp \in \mathbb{R}^{nd \times nd}$ as follows:

$$\begin{aligned} K(k)_{i,j} &= \frac{1}{m} \sum_{r=1}^m x_i^\top x_j \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] a_r a_r^\top, \\ K(k)_{i,j}^\perp &= \frac{1}{m} \sum_{r \in S_i^\perp} x_i^\top x_j \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] a_r a_r^\top. \end{aligned}$$

Next, we write $\phi(z) = z \mathbb{1}[z \geq 0]$ to make use the definition of S_i and expand the form of $\nabla_{w_r} L(W(k))$:

$$\begin{aligned} v_{1,i} &= \frac{1}{\sqrt{md}} \sum_{r \in S_i} a_r \left(-\eta \nabla_{w_r} L(W(k)) \right)^\top x_i \mathbb{1}[w_r(k)^\top x_i \geq 0] \\ &= \frac{\eta}{md} \sum_{j=1}^n x_i^\top x_j \sum_{r \in S_i} \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] a_r a_r^\top (x_j - u_j) \\ &= \frac{\eta}{d} \sum_{j=1}^n (K_{i,j}(k) - K_{i,j}^\perp(k)) (x_j - u_j), \end{aligned}$$

Then, we can write v_1 as:

$$v_1 = \frac{\eta}{d} (K(k) - K^\perp(k)) \text{vec}(X - U(k)), \quad (5.11)$$

and expand $\|X - U(k+1)\|_F^2$:

$$\begin{aligned} \|X - U(k+1)\|_F^2 &= \|\text{vec}(X - U(k+1))\|^2 \\ &= \|\text{vec}(X - U(k)) - \text{vec}(U(k+1) - U(k))\|_F^2 \\ &= \|X - U(k)\|_F^2 - 2\text{vec}(X - U(k))^\top \text{vec}(U(k+1) - U(k)) \\ &\quad + \|U(k+1) - U(k)\|_F^2. \end{aligned}$$

We can further expand the second term above using (6.19) as below:

$$\begin{aligned} &\text{vec}(X - U(k))^\top \text{vec}(U(k+1) - U(k)) \\ &= \text{vec}(X - U(k))^\top (v_1 + v_2) \\ &= \text{vec}(X - U(k))^\top v_1 + \text{vec}(X - U(k))^\top v_2 \\ &= \frac{\eta}{d} \text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k)) - \frac{\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp (X - U(k)) \\ &\quad + \text{vec}(X - U(k))^\top v_2. \end{aligned}$$

We define and bound the following quantities and bound them in Claims B.1, B.2, B.3 and B.4.

$$\begin{aligned} C_1 &= -\frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k)), \\ C_2 &= \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp (X - U(k)), \\ C_3 &= -2 \text{vec}(X - U(k))^\top v_2, \\ C_4 &= \|U(k+1) - U(k)\|_F^2. \end{aligned}$$

Proof of Theorem 5.2. We are now ready to prove the induction hypothesis. What we need to is to prove

$$\|X - U(k')\|_F^2 \leq \left(1 - \frac{\eta\lambda_0}{2d}\right)^{k'} \|X - U(0)\|_F^2$$

holds for $k' = k + 1$ with probability at least $1 - \delta$. In fact,

$$\begin{aligned} \|X - U(k+1)\|_F^2 &= \|X - U(k)\|_F^2 + C_1 + C_2 + C_3 + C_4 \\ &\leq \|X - U(k)\|_F^2 \left(1 - \frac{\eta\lambda_0}{d} + 8\eta nR + 8\eta nR + \eta^2 n\lambda_n\right), \end{aligned}$$

with probability at least $1 - \delta$ where the last step follows from Claim B.1, B.2, B.3, and B.4.

Choice of η and R . We need to choose η and R such that

$$\left(1 - \frac{\eta\lambda_0}{d} + 8\eta nR + 8\eta nR + \eta^2 n\lambda_n\right) \leq 1 - \frac{\eta\lambda_0}{2d}. \quad (5.12)$$

If we set $\eta = \frac{\lambda_0}{4nd\lambda_n}$ and $R = \frac{\lambda_0}{64nd}$, we have

$$8\eta nR + 8\eta nR = 16\eta nR \leq \frac{\eta\lambda_0}{4d}, \quad \text{and} \quad \eta^2 n\lambda_n \leq \frac{\eta\lambda_0}{4d}.$$

Finally,

$$\|X - U(k+1)\|_F^2 \leq \left(1 - \frac{\eta\lambda_0}{2d}\right) \|X - U(k)\|_F^2.$$

holds with probability at least $1 - \delta$ if $2n \exp(-mR) \leq \delta/3$.

Lower bound on the level of over-parameterization m . We require for any $\delta \in (0, 1)$ that

$$R' = \frac{4d\sqrt{\lambda_n}\|X - U(0)\|_F}{\sqrt{m}\lambda_0} < R = \min \left\{ \frac{\lambda_0}{64nd}, \frac{\lambda_0\delta}{2n^2d} \right\},$$

where the first bound on R comes from the gradient descent whereas the second is required in Lemma 5.2.

By Claim 5.1 that $\|X - U(0)\|_F \leq \sqrt{\frac{2n}{\delta}}$ with probability at least $1 - \delta$, then we require

$$m \geq C \frac{n^5 \lambda_n d^4}{\lambda_0^4 \delta^3},$$

for a sufficiently large constant $C > 0$ so that the descent holds with probability $1 - \delta$. ■

We give proofs for Claims B.1, B.2, B.3, and B.4 in Appendix B.

6 Jointly-trained Autoencoders

In the previous section, we analyzed the gradient dynamics of a two-layer autoencoder under the weakly-trained regime. We now analyze the jointly-trained regime where the loss is optimized over both sets of layer weights. For consistency of our presentation, we reuse some key notations in this section; for example, $K(t), U(t)$ have the same interpretation as before but possess a different closed form.

6.1 Gradient flow

The loss function we consider for this *jointly-trained* regime is the same:

$$L(W, A) = \frac{1}{2} \sum_{i=1}^n \|x_i - \frac{1}{\sqrt{md}} A \phi(W^\top x_i)\|^2. \quad (6.1)$$

The difference is that the optimization is now taken over *both* weights W and A . To make the comparison easier, the matrices W and A are randomly initialized in the same way such that

$$w_{ij}(0) \sim \mathcal{N}(0, 1), \quad a_{ij}(0) \sim \text{Unif}(\{-1, 1\})$$

are drawn independently for each pair (i, j) . W and A are then updated using gradient descent with step size η :

$$W(k+1) = W(k) - \eta \nabla_W L(W(k), A(k)), \quad k = 0, 1, \dots \quad (6.2)$$

$$A(k+1) = A(k) - \eta \nabla_A L(W(k), A(k)), \quad k = 0, 1, \dots \quad (6.3)$$

Similar to the previous case, we derive the gradients of $L(W, A)$ with respect the column w_r of W and a_r of A . The gradient $\nabla_{w_r} L(W, A)$ is the same in (5.2) in Section 5.1 whereas $\nabla_{a_r} L(W, A)$ is standard:

$$\nabla_{w_r} L(W, A) = -\frac{1}{\sqrt{md}} \sum_{i=1}^n \mathbb{1}[w_r^\top x_i \geq 0] x_i a_r^\top (x_i - u_i), \quad (6.4)$$

$$\nabla_{a_r} L(W, A) = -\frac{1}{\sqrt{md}} \sum_{i=1}^n \phi(w_r^\top x_i) (x_i - u_i). \quad (6.5)$$

Consider two ODEs, one for each weight vector over the continuous time t :

$$\frac{dw_r(t)}{dt} = -\nabla_{w_r} L(W(t), A(t)), \quad (6.6)$$

$$\frac{da_r(t)}{dt} = -\nabla_{a_r} L(W(t), A(t)). \quad (6.7)$$

Using (6.4), (6.5), (6.6) and (6.7), the continuous-time dynamics of the predicted output, $u_i(t)$, for sample x_i is given by:

$$\begin{aligned} \frac{du_i(t)}{dt} &= \frac{d}{dt} \left(\frac{1}{\sqrt{md}} \sum_{r=1}^m a_r \phi(w_r^\top x_i) \right) \\ &= \frac{1}{\sqrt{md}} \sum_{r=1}^m \left(J_{w_r}(a_r \phi(w_r^\top x_i)) \frac{dw_r}{dt} + J_{a_r}(a_r \phi(w_r^\top x_i)) \frac{da_r}{dt} \right) \\ &= \frac{1}{\sqrt{md}} \sum_{r=1}^m (1[w_r^\top x_i \geq 0] a_r^\top x_i (-\nabla_{w_r} L(W, A)) + \phi(w_r^\top x_i) (-\nabla_{a_r} L(W, A))) \\ &= -\frac{1}{\sqrt{md}} \sum_{j=1}^n \sum_{r=1}^m (1[w_r^\top x_i \geq 0, w_r^\top x_j \geq 0] x_i^\top x_j a_r a_r^\top + \phi(w_r^\top x_i) \phi(w_r^\top x_j) I) (x_j - u_j). \end{aligned}$$

In these expressions, we skip the dependence of the weight vectors on time t and simply write them as w_r and a_r . Vectorizing $\frac{dU(t)}{dt}$, we get to the key equation that characterizes the dynamics of $U(t)$:

$$\frac{d\text{vec}(U(t))}{dt} = \frac{1}{d} \left(G(t) + H(t) \right) \text{vec}(X - U(t)). \quad (6.8)$$

In the above equation, $G(t)$ is a size- $nd \times nd$ matrix of the form:

$$G(t) = \frac{1}{m} \sum_{r=1}^m \tilde{X}_r(t)^\top \tilde{X}_r(t) \otimes a_r(t) a_r(t)^\top, \quad (6.9)$$

where $\tilde{X}_r(t) = [\mathbb{1}[w_r(t)^\top x_1 \geq 0]x_1, \dots, \mathbb{1}[w_r(t)^\top x_n \geq 0]x_n]$, while $H(t)$ is a size- $nd \times nd$ matrix:

$$H(t) = \frac{1}{m} \sum_{r=1}^m \phi(X^\top w_r(t)) \phi(w_r(t)^\top X) \otimes I. \quad (6.10)$$

Let us emphasize again that $G(t)$ is precisely the kernel that governs the dynamics for the weakly-trained case. On the other hand, $H(t)$ is a Kronecker form of the Hessian of the loss function derived with respect to A , using the features produced at the output of the ReLU activations.

As shown in Section 3, assuming randomness and independence of $W(0)$ and $A(0)$, we can prove that as $m \rightarrow \infty$, $H(0)$ and $G(0)$ converge to the corresponding NTKs whose minimum eigenvalues are assumed to be positive. More specifically, we have

$$\begin{aligned} G^\infty &= \mathbb{E}_{W(0), A(0)}[G(0)] \\ &= \mathbb{E}_{w(0), a(0)}[\tilde{X}(0)^\top \tilde{X}(0) \otimes a(0) a(0)^\top] \\ &= \mathbb{E}_w[\tilde{X}^\top \tilde{X}] \otimes I. \end{aligned} \quad (6.11)$$

and

$$\begin{aligned} H^\infty &= \mathbb{E}_{W(0), A(0)}[H(0)] \\ &= \mathbb{E}_{w(0), a(0)}[\phi(X^\top w(0)) \phi(w(0)^\top X)] \otimes I. \end{aligned} \quad (6.12)$$

Denote the time-dependent kernel $K(t) = G(t) + H(t)$. Since both $G(t)$ and $H(t)$ are positive semi-definite, we only focus on $H(t)$ for reasons that will become clear shortly.

Since $G(t)$ is also positive definite with high probability (Section 5.1), the flow convergence can be also boosted by the positive definiteness of G^∞ . By Assumption 2,

$$\lambda_{\min}(K^\infty) \geq \lambda_{\min}(H^\infty) \geq \lambda_0 > 0.$$

Since $G(0)$ is positive semi-definite, in order to bound the minimum eigenvalue of $K(0)$, all we need is to bound that of $H(0)$. Importantly, we observe that the smoothness of the kernel $H(t)$ is much better as a function of the deviation of the weights from the initialization. This allows the weights to change with a larger amount than merely using $G(t)$, and enables us to significantly reduce the number of parameters required for the gradient to reach a global optimum.

Our main result for gradient flow of the jointly-trained autoencoder is given by:

Theorem 6.1 (Linear convergence of gradient flow, jointly-trained regime). *Suppose Assumptions 1 and 2 hold. The initial weights are independently drawn such that $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{-1, 1\}^d)$ for all $r \in [m]$. If $m \geq C \frac{n\lambda_0^3 d}{\lambda_0^4 \delta^2}$ for some large enough constant C , then with probability at least $1 - \delta$,*

$$\|X - U(t)\|_F^2 \leq \exp\left(-\frac{\lambda_0 t}{d}\right) \|X - U(0)\|_F^2.$$

Remark 6.1. We initialize the second-layer weights A with independent Rademacher entries. This is for convenience of analysis because such A has constant-norm columns. However, similar results should easily follow for initialization with more practical schemes (for example, i.i.d. Gaussians).

We will first state and prove a few auxiliary results in Lemmas 6.1, 6.2, and 6.6 and then use them to prove Theorem 6.1.

Lemma 6.1. *For any $\delta \in (0, \frac{\lambda_0}{12d\lambda_n^2})$, if $m \geq C \frac{nd\lambda_n \log^2(nd/\delta)}{\lambda_0^2}$ for some large enough constant C , then with probability at least $1 - 1/(2nd)^{2 \log nd} - m\delta$, one obtains $\|H(0) - H^\infty\| \leq \lambda_0/4$ and $\lambda_{\min}(H(0)) \geq 3\lambda_0/4$.*

The proof of this Lemma is deferred to Appendix C.

Lemma 6.2. *Suppose $\|W(t) - W(0)\|_F \leq R_w$. Then,*

$$\|H(t) - H(0)\| \leq \frac{\lambda_n}{m} (2\|W(0)\| + R_w) R_w.$$

Particularly, if $\frac{\lambda_n}{m} (2\|W(0)\| + R_w) R_w \leq \frac{\lambda_0}{4}$, then $\|H(t) - H(0)\| \leq \frac{\lambda_0}{4}$. Therefore, $\lambda_{\min}(K(t)) > \frac{\lambda_0}{2}$ if $\lambda_{\min}(H(0)) \geq \frac{3\lambda_0}{4}$.

Remark 6.2. Let us compare with Lemma 5.2. Note that compared with that bound, $O(n^2 d R_w)$ on the kernel perturbation, here the spectral norm bound on $H(t) - H(0)$ is significantly better in two ways:

- (i) the bound scales with $1/\sqrt{m}$, which later determines the over-parameterization and
- (ii) the movement is now characterized by the total $\|W(t) - W(0)\|_F$. This is possible due to the smoothness of the ReLU activation, which is the reason why we focus on $H(t)$ instead of $G(t)$.

Proof. We apply the triangle inequality and use the Lipschitz property of the rectified linear unit to bound the difference. Recall that

$$H(t) = \frac{1}{m} \sum_{r=1}^m \phi(X^\top w_r(t)) \phi(w_r(t)^\top X) = \frac{1}{m} \phi(X^\top W(t)) \phi(W(t)^\top X).$$

Then, we can upper bound the perturbation as follows:

$$\begin{aligned} \|H(t) - H(0)\| &\leq \frac{1}{m} \|\phi(X^\top W(t)) \phi(W(t)^\top X) - \phi(X^\top W(0)) \phi(W(0)^\top X)\| \\ &\leq \frac{1}{m} \|\phi(X^\top W(t))\| \|\phi(W(t)^\top X) - \phi(W(0)^\top X)\| \\ &\quad + \frac{1}{m} \|\phi(X^\top W(t)) - \phi(X^\top W(0))\| \|\phi(W(0)^\top X)\| \\ &\leq \frac{1}{m} \|X\|^2 (\|W(t)\| + \|W(0)\|) \|W(t) - W(0)\|_F \\ &\leq \frac{\lambda_n}{m} (2\|W(0)\| + \|W(t) - W(0)\|) \|W(t) - W(0)\|_F \\ &\leq \frac{\lambda_n}{m} (2\|W(0)\| + R_w) \lambda_n R_w. \end{aligned}$$

In the third step, we use the fact that the ReLU function is 1-Lipschitz and $\|\phi(X^\top W(t))\| \leq \|X\| \|W(t)\|$. The last step follows by $\|W(t) - W(0)\|_F \leq R_w$.

Using the condition and Weyl's inequality, one can easily show that $\lambda_{\min}(K(t)) \geq \lambda_{\min}(H(t)) > \lambda_0/2$. ■

We have proved that as long as the weight matrix $W(t)$ do not change much over t , the minimum eigenvalue of $K(t)$ stays positive. Next, we show that this implies the exponential decay of the loss with iteration, and give a condition under which the weights do not change much.

Lemma 6.3. Fix $t > 0$. Suppose $\lambda_{\min}(K(s)) \geq \frac{\lambda_0}{2}$ for all $0 \leq s < t$. Then

$$\|X - U(s)\|_F^2 \leq \exp\left(-\frac{\lambda_0 s}{d}\right) \|X - U(0)\|_F^2.$$

Proof. We have $\lambda_{\min}(K(s)) \geq \frac{\lambda_0}{2}$, then

$$\begin{aligned} \frac{d}{ds} \left(\|\text{vec}(X - U(s))\|_2^2 \right) &= -2\text{vec}(X - U(s))^\top \frac{K(s)}{d} \text{vec}(X - U(s)) \\ &\leq -\frac{2}{d} \lambda_{\min}(K(s)) \|\text{vec}(X - U(s))\|_2^2 \\ &\leq -\frac{\lambda_0}{d} \|X - U(s)\|_F^2, \end{aligned} \tag{6.13}$$

since $\lambda_{\min}(H(s)) \geq \frac{\lambda_0}{2}$. Therefore,

$$\begin{aligned} \|X - U(s)\|_F^2 &= \|\text{vec}(X - U(s))\|_2^2 \leq \exp\left(-\frac{\lambda_0 s}{d}\right) \|\text{vec}(X - U(0))\|_2^2 \\ &\leq \exp\left(-\frac{\lambda_0 s}{d}\right) \|X - U(0)\|_F^2. \end{aligned}$$

■

Lemma 6.4. Fix $t > 0$. Suppose $\lambda_{\min}(K(s)) \geq \frac{\lambda_0}{2}$ and $\|A(s) - A(0)\|_F \leq R_a$ for all $0 \leq s < t$. For all $r \in [m]$, we have

$$\|W(t) - W(0)\|_F \leq \frac{2\sqrt{d\lambda_n}(\|A(0)\| + R_a)\|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R'_w.$$

Proof. For $s \in [0, t)$, we have

$$\begin{aligned} \frac{d}{ds} w_r(s) &= -\nabla_{w_r} L(W(s), A(s)) \\ &= \frac{1}{\sqrt{md}} \sum_{i=1}^n \mathbb{1}[w_r^\top x_i \geq 0] x_i a_r(s)^\top (x_i - u_i(s)) \\ &= \frac{1}{\sqrt{md}} \tilde{X}_r (X - U(s))_r^\top a_r(s). \end{aligned}$$

Then, one can bound the entire weight matrix as follows:

$$\begin{aligned} \left\| \frac{d}{ds} W(s) \right\|_F &\leq \frac{\|X\|}{\sqrt{md}} \|(X - U(s))^\top A(s)\|_F \\ &\leq \frac{\sqrt{\lambda_n}}{\sqrt{md}} \|X - U(s)\|_F \|A(s)\| \\ &\leq \frac{\sqrt{\lambda_n}(\|A(0)\| + R_a)}{\sqrt{md}} \exp\left(-\frac{\lambda_0 s}{2d}\right) \|X - U(0)\|_F. \end{aligned}$$

In the second step, we use the fact $\|CD\|_F \leq \|C\|_F \|D\|$ for any matrices C, D , and $\|X\|^2 = \lambda_n$. The last step follows from $\|A(s)\| \leq \|A(0)\| + R_a$ and Lemma 6.3. Using the same continuity, we have

$$\begin{aligned}
\|W(t) - W(0)\|_F &\leq \lim_{t' \rightarrow t} \int_0^{t'} \left\| \frac{d}{ds} W(s) \right\|_F \\
&\leq \lim_{t' \rightarrow t} \int_0^{t'} \frac{\sqrt{\lambda_n}(\|A(0)\| + R_a) \exp\left(-\frac{\lambda_0 s}{2d}\right) \|X - U(0)\|_F}{\sqrt{md}} ds \\
&\leq \frac{2\sqrt{\lambda_n}(\|A(0)\| + R_a) \|X - U(0)\|_F}{\sqrt{md}} \lim_{t' \rightarrow t} \int_0^{t'} \exp\left(-\frac{\lambda_0 s}{2d}\right) ds \\
&\leq \frac{2\sqrt{d\lambda_n}(\|A(0)\| + R_a) \|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R'_w.
\end{aligned}$$

■

Lemma 6.5. Fix $t > 0$. Suppose $\lambda_{\min}(K(s)) \geq \frac{\lambda_0}{2}$ and $\|W(s) - W(0)\|_F \leq R_w$ for all $0 \leq s < t$, then for $r = 1, 2, \dots, m$

$$\|A(t) - A(0)\|_F \leq \frac{2\sqrt{d\lambda_n}(\|W(0)\| + R_w) \|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R'_a.$$

Proof. For $s \in [0, t]$, we use the gradient derived in (6.5) and (6.7) to obtain:

$$\begin{aligned}
\frac{d}{ds} a_r(s) &= -\nabla_{a_r} L(W(s), A(s)) \\
&= \frac{1}{\sqrt{md}} \sum_{i=1}^n \phi(w_r^\top x_i) (x_i - u_i(s)) \\
&= \frac{1}{\sqrt{md}} (X - U(s)) \phi(X^\top w_r(s))
\end{aligned}$$

Then, one can write

$$\begin{aligned}
\left\| \frac{d}{ds} A(s) \right\|_F &= \left\| \frac{1}{\sqrt{md}} (X - U(s)) \phi(X^\top W(s)) \right\|_F \\
&\leq \frac{\sqrt{\lambda_n}}{\sqrt{md}} \|X - U(s)\|_F \|W(s)\| \\
&\leq \frac{\sqrt{\lambda_n}(\|W(0)\| + R_w)}{\sqrt{md}} \exp\left(-\frac{\lambda_0 s}{2d}\right) \|X - U(0)\|_F,
\end{aligned}$$

where we use $\|X\| \leq \sqrt{\lambda_n}$, $\|W(s)\| \leq \|W(0)\| + R_w$. The last step follows from Lemma 6.3. Now, we integrate out s :

$$\begin{aligned}
\|A(t) - A(0)\|_F &\leq \int_0^t \left\| \frac{d}{ds} A(s) \right\|_F \leq \int_0^t \frac{\sqrt{\lambda_n}(\|W(0)\| + R_w) \exp\left(-\frac{\lambda_0 s}{2d}\right) \|X - U(0)\|_F}{\sqrt{m}} ds \\
&\leq \frac{2\sqrt{d\lambda_n}(\|W(0)\| + R_w) \|X - U(0)\|_F}{\sqrt{m}\lambda_0} = R'_a,
\end{aligned}$$

which is what we need.

■

Lemma 6.6. If $R'_w < R_w$ and $R'_a < R_a$, then for all $t \geq 0$, we have

- (i) $\lambda_{\min}(K(t)) \geq \frac{\lambda_0}{2}$; and for all $r \in [m]$, $\|W(t) - W(0)\|_F \leq R'_w$, $\|A(t) - A(0)\|_F \leq R'_a$;
- (ii) If (i) holds, then $\|X - U(t)\|_F^2 \leq \exp\left(-\frac{\lambda_0}{d}\right) \|X - U(0)\|_F^2$.

Proof. Suppose on the contrary that

$$\mathcal{T} = \left\{ t \geq 0 : \lambda_{\min}(K(t)) \leq \frac{\lambda_0}{2} \text{ or } \|W(t) - W(0)\|_F > R'_w \text{ or } \|A(t) - A(0)\|_F > R'_a \right\}. \quad (6.14)$$

is not an empty set. Therefore, $t_0 \triangleq \inf \mathcal{T}$ exists. Using the same continuity argument as in Lemma 5.3, one can verify that $t_0 > 0$. First, if $\lambda_{\min}(K(t_0)) \leq \frac{\lambda_0}{2}$, then by Lemma 6.2, $\|W(t_0) - W(0)\|_F > R_w > R'_w$, which is a contradiction because it violates the minimality of t_0 .

The other two cases are similar, so we will prove one of them. If it holds true that

$$\|W(t) - W(0)\|_F > R'_w.$$

The definitions of t_0 and \mathcal{T} implies that for any $s \in [0, t_0)$, $\lambda_{\min}(K(s)) \geq \frac{\lambda_0}{2}$ and $\|A(s) - A(0)\|_F \leq R'_a$. Then, Lemma 6.4 leads to:

$$\|W(t_0) - W(0)\|_F \leq R'_w,$$

which is a contradiction. Therefore, we have finish the proof. ■

Proof of Theorem 6.1. With the results, we can prove the Theorem. From Lemma 6.6, if $R'_w \leq R_w$ and $R'_a \leq R_a$, then

$$\|X - U(t)\|_F^2 \leq \exp\left(-\frac{\lambda_0}{d}\right) \|X - U(0)\|_F^2.$$

We only need the conditions $R'_w = R'_a \leq R_w = R_a$ to satisfy for this to work. The conditions are

$$\begin{aligned} \frac{\lambda_n}{m} (2\|W(0)\| + R_w) R_w &\leq \frac{\lambda_0}{4}; \\ \text{and } R_w < R'_w &= \frac{2\sqrt{d\lambda_n}(\|A(0)\| + R_a) \|X - U(0)\|_F}{\sqrt{m}\lambda_0}. \end{aligned}$$

Note that $\|X - U(0)\|_F^2 \leq 3n/2\delta$ with probability at least $1 - \delta$. Also, using a standard bound on sub-Gaussian matrices, we have $\|W(0)\| \leq 2\sqrt{m} + \sqrt{d}$ and $\|A(0)\| \leq 2\sqrt{m} + \sqrt{d}$ with probability at least $1 - 2\exp(-m)$. Then if we have the order of $m \geq \Omega\left(\frac{nd\lambda_n^3}{\lambda_0^4\delta^2}\right)$. Therefore, we finished the proof for the gradient flow Theorem. ■

6.2 Gradient descent

As above, we will now appropriately discretize the gradient flow to obtain a convergence result for gradient descent with finite step size for the jointly-trained regime.

Theorem 6.2. *Suppose Assumptions 1 and 2 hold. At initialization, suppose the weights are independently drawn from $w_r \sim \mathcal{N}(0, I)$ and $a_r \sim \text{Unif}(\{-1, 1\}^d)$ for all $r \in [m]$. If $m \geq C \frac{n\lambda_n^3 d}{\lambda_0^4 \delta^2}$ for some large enough constant C , then with probability at least $1 - \delta$ the gradient descent on W with step size $\eta = \Theta(\frac{\lambda_0}{n\lambda_n})$,*

$$\|X - U(k)\|_F^2 \leq \left(1 - \frac{\eta\lambda_0}{2d}\right)^k \|X - U(0)\|_F^2. \quad (6.15)$$

We will prove 6.2 by induction. The base case when $k = 0$ is trivially true. Assume holds for $k' = 0, 1, \dots, k$ and we want to show (6.15) for $k' = k + 1$. First, we prove that $\|W(k+1) - W(0)\|_F$ and $\|A(k+1) - A(0)\|_F$ are small enough, and we then use that to bound $\|X - U(k+1)\|_F^2$.

In this section, we define and assume that

$$R_w < \frac{4\sqrt{d\lambda_n}(\|A(0)\| + R_a)\|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R'_w \quad \text{and} \quad R_a < \frac{4\sqrt{d\lambda_n}(\|W(0)\| + R_w)\|X - U(0)\|_F}{\sqrt{m}\lambda_0} \triangleq R'_a.$$

First, we show the following auxiliary lemma.

Lemma 6.7. *If the condition (6.15) holds for $k' = 0, 1, \dots, k$, then we have*

$$\|W(k+1) - W(0)\|_F \leq R'_w, \quad \text{and} \quad \|A(k+1) - A(0)\|_F \leq R'_a$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Proof. We prove this by induction. Clearly, both hold when $k' = 0$. Assuming that both hold for $k' \leq k$. We will prove both hold for $k' = k + 1$.

We use the expression of the gradients over w_r and a_r in (6.4) and (6.5):

$$\begin{aligned} \nabla_{w_r} L(W(k), A(k)) &= - \sum_{i=1}^n \frac{1}{\sqrt{md}} \mathbb{1}[w_r(k)^T x_i \geq 0] x_i a_r(k)^\top (x_i - u_i(k)) \\ &= - \frac{1}{\sqrt{md}} \tilde{X}_r(k) (X - U(k)) a_r(k), \\ \nabla_{a_r} L(W(k), A(k)) &= - \sum_{i=1}^n \frac{1}{\sqrt{md}} \phi(w_r(k)^\top x_i) (x_i - u_i(k)) \\ &= - \frac{1}{\sqrt{md}} (X - U(k)) \phi(X^\top w_r(k)). \end{aligned}$$

Then, the difference of the weight matrix W is:

$$\begin{aligned} \|W(k+1) - W(0)\|_F &\leq \eta \frac{\|X\|}{\sqrt{md}} \sum_{k'=0}^k \|\text{vec}(X - U(k'))\|_F (\|A(0)\| + R_a) \\ &\leq \eta \frac{\sqrt{\lambda_n}(\|A(0)\| + R_a)}{\sqrt{m}} \sum_{k'=0}^k \left(1 - \frac{\eta\lambda_0}{2d}\right)^{k'/2} \|X - U(0)\|_F \\ &\leq \eta \frac{\sqrt{\lambda_n}(\|A(0)\| + R_a)}{\sqrt{m}} \|X - U(0)\|_F \sum_{k'=0}^{\infty} \left(1 - \frac{\eta\lambda_0}{2d}\right)^{k'/2} \\ &= \eta \frac{\sqrt{\lambda_n}(\|A(0)\| + R_a)}{\sqrt{m}} \|X - U(0)\|_F \frac{4d}{\eta\lambda_0} \\ &= \frac{4\sqrt{d\lambda_n}(\|A(0)\| + R_a)\|X - U(0)\|_F}{\sqrt{m}\lambda_0} = R'_w, \end{aligned}$$

where the third step and fourth step follow from $\|\tilde{X}_r(k')\| \leq \|X\| = \sqrt{\lambda_n}$ and the induction hypothesis $\|A(k')\| \leq \|A(0)\| + R_a$. The last step follows from

$$\sum_{i=0}^{\infty} (1 - \eta\lambda_0/2)^{i/2} \leq \frac{4d}{\eta\lambda_0}.$$

Similarly, we bound the difference of the weight matrix A between time $k+1$ and 0:

$$\begin{aligned}
\|A(k+1) - A(0)\|_F &= \eta \left\| \sum_{k'=0}^k \frac{1}{\sqrt{md}} (X - U(k')) \phi(X^\top W(k')) \right\|_F \\
&\leq \eta \frac{\|X\|}{\sqrt{md}} \sum_{k'=0}^k \|X - U(k')\|_F \|W(k')\| \\
&\leq \eta \frac{\sqrt{\lambda_n}}{\sqrt{md}} \sum_{k'=0}^k \left(1 - \frac{\eta\lambda_0}{2d}\right)^{k'/2} \|X - U(0)\|_F (\|W_0\| + R_w) \\
&= \eta \frac{\sqrt{\lambda_n} (\|W_0\| + R_w) \|X - U(0)\|_F}{\sqrt{m}} \frac{4}{\eta\lambda_0} \\
&= \frac{4\sqrt{\lambda_n} (\|W_0\| + R_w) \|X - U(0)\|_F}{\sqrt{m}\lambda_0} = R'_a
\end{aligned}$$

where the third step and fourth step follow from the facts that $\|\phi(X^\top W(k'))\| \leq \|X\| \|W(k')\|$ and $\|X\| = \sqrt{\lambda_n}$, and $\|W(k')\| \leq \|W(0)\| + R_w$.

We have therefore shown that $\|W(k') - w(0)\|_F \leq R'_w$ and $\|A(k') - A(0)\|_F \leq R'_a$ for $k' = k+1$. \blacksquare

Now, we expand $\|X - U(k+1)\|_F^2$ in terms of the step k . Recall the update rule in (6.2) and (6.3) that

$$W(k+1) = W(k) - \eta \nabla_W L(W(k), A(k)), \quad k = 0, 1, \dots \quad (6.16)$$

$$A(k+1) = A(k) - \eta \nabla_A L(W(k), A(k)), \quad k = 0, 1, \dots \quad (6.17)$$

where the gradients is given above. Next, we compute the difference of the prediction between two consecutive steps, a discrete version of $\frac{du_i(t)}{dt}$. For each $i \in [n]$, we have

$$\begin{aligned}
u_i(k+1) - u_i(k) &= \frac{1}{\sqrt{md}} \sum_{r=1}^m (a_r(k+1) \phi(w_r(k+1)^\top x_i) - a_r(k) \phi(w_r(k)^\top x_i)) \\
&= \frac{1}{\sqrt{md}} \sum_{r=1}^m \left((a_r(k) - \eta \nabla_{a_r} L) \phi \left((w_r(k) - \eta \nabla_{w_r} L)^\top x_i \right) - a_r(k) \phi(w_r(k)^\top x_i) \right). \quad (6.18)
\end{aligned}$$

For a particular r , if the activation pattern does not change, we can write the inside term as:

$$\begin{aligned}
&(a_r(k) - \eta \nabla_{a_r} L) \phi \left((w_r(k) - \eta \nabla_{w_r} L)^\top x_i \right) - a_r(k) \phi(w_r(k)^\top x_i) \\
&= \left(-\eta a_r(k) (\nabla_{w_r} L)^\top - \eta (\nabla_{a_r} L) w_r(k)^\top + \eta^2 (\nabla_{a_r} L) (\nabla_{w_r} L)^\top \right) x_i \mathbb{1}[w_r(k)^\top x_i \geq 0],
\end{aligned}$$

where the first part corresponds to kernel $G(t)$ and the second part corresponds to the $H(t)$ shown up in the gradient flow analysis. With this intuition, we split the right hand side into two parts. $v_{1,i}$ represents the terms that the pattern does not change and $v_{2,i}$ represents the remaining term that pattern may changes.

For each $i \in [n]$, we define $S_i = \{r \in [m] : \mathbb{1}[w_r(k+1)^\top x_i \geq 0] = \mathbb{1}[w_r(k)^\top x_i \geq 0]\}$, and $S_i^\perp = [m] \setminus S_i$. Then, we write $v_{1,i}$ and $v_{2,i}$ as follows:

$$\begin{aligned}
v_{1,i} &\triangleq \frac{1}{\sqrt{md}} \sum_{r \in S_i} (a_r(k+1) \phi(w_r(k+1)^\top x_i) - a_r(k) \phi(w_r(k)^\top x_i)), \\
v_{2,i} &\triangleq \frac{1}{\sqrt{md}} \sum_{r \in S_i^\perp} (a_r(k+1) \phi(w_r(k+1)^\top x_i) - a_r(k) \phi(w_r(k)^\top x_i)).
\end{aligned}$$

We further write $v_1 = (v_{1,1}^\top, v_{1,2}^\top, \dots, v_{1,n}^\top)^\top$ and do the same for v_2 . Hence, we write

$$\text{vec}(U(k+1) - U(k)) = v_1 + v_2.$$

In order to analyze $v_1 \in \mathbb{R}^n$, we provide definition of $G, G^\perp \in \mathbb{R}^{nd \times nd}$ and $H, H^\perp \in \mathbb{R}^{nd \times nd}$,

$$\begin{aligned} G(k)_{i,j} &= \frac{1}{m} \sum_{r=1}^m x_i^\top x_j \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] a_r(t) a_r(t)^\top, \\ G(k)_{i,j}^\perp &= \frac{1}{m} \sum_{r \in S_i^\perp} x_i^\top x_j \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] a_r(t) a_r(t)^\top, \\ H(k)_{i,j} &= \frac{1}{m} \sum_{r=1}^m \phi(w_r(t)^\top x_i) \phi(w_r(t)^\top x_j) I, \\ H(k)_{i,j}^\perp &= \frac{1}{m} \sum_{r \in S_i^\perp} \phi(w_r(t)^\top x_i) \phi(w_r(t)^\top x_j) I. \end{aligned}$$

Using the fact that $\phi(z) = z \mathbb{1}[z \geq 0]$ and the definition of S_i , we expand the forms of the gradients in 6.4 and 6.5 and get:

$$\begin{aligned} v_{1,i} &= -\frac{1}{\sqrt{md}} \sum_{r \in S_i} \eta a_r(k) (\nabla_{w_r} L)^\top x_i \mathbb{1}[w_r(k)^\top x_i \geq 0] \\ &\quad - \frac{1}{\sqrt{md}} \sum_{r \in S_i} \eta (\nabla_{a_r} L) w_r(k)^\top x_i \mathbb{1}[w_r(k)^\top x_i \geq 0] \\ &\quad + \frac{1}{\sqrt{md}} \sum_{r \in S_i} \eta^2 (\nabla_{a_r} L) (\nabla_{w_r} L)^\top x_i \mathbb{1}[w_r(k)^\top x_i \geq 0] \\ &= \frac{\eta}{d} \sum_{j=1}^n (G(k)_{i,j} - G(k)_{i,j}^\perp + H(k)_{i,j} - H(k)_{i,j}^\perp) (x_j - u_j) + v_{3,i}, \end{aligned}$$

where v_3 will be treated as a perturbation:

$$v_{3,i} = \frac{\eta^2}{\sqrt{md}} \sum_{r \in S_i} (\nabla_{a_r} L) (\nabla_{w_r} L)^\top x_i \mathbb{1}[w_r(k)^\top x_i \geq 0].$$

Then, we can write v_1 as:

$$v_1 = \frac{\eta}{d} (K(k) - K^\perp(k)) \text{vec}(X - U(k)) + v_3, \quad (6.19)$$

in which $K(k) = G(k) + H(k)$ — the discrete NTK kernel and $K^\perp(k) = H^\perp(k) + G^\perp(k)$. Lastly, we come to the main prediction dynamics in discrete time for $\text{vec}(U(k+1) - U(k))$ as:

$$\text{vec}(U(k+1) - U(k)) = \frac{\eta}{d} (K(k) - K^\perp(k)) \text{vec}(X - U(k)) + v_2 + v_3.$$

Using this equation, we can rewrite $\|X - U(k+1)\|_F^2$ in terms of $X - U(k)$ as follows:

$$\begin{aligned}
\|X - U(k+1)\|_F^2 &= \|\text{vec}(X - U(k+1))\|^2 \\
&= \|\text{vec}(X - U(k)) - \text{vec}(U(k+1) - U(k))\|_F^2 \\
&= \|X - U(k)\|_F^2 - 2\text{vec}(X - U(k))^\top \text{vec}(U(k+1) - U(k)) \\
&\quad + \|U(k+1) - U(k)\|_F^2 \\
&= \|X - U(k)\|_F^2 - \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k)) \\
&\quad + \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp \text{vec}(X - U(k)) \\
&\quad - \frac{2\eta}{d} \text{vec}(X - U(k))^\top (v_2 + v_3) \\
&\quad + \|U(k+1) - U(k)\|_F^2.
\end{aligned}$$

We define and upper bound each of the following terms

$$\begin{aligned}
C_1 &= -\frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k)), \\
C_2 &= \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp \text{vec}(X - U(k)), \\
C_3 &= -\frac{2\eta}{d} \text{vec}(X - U(k))^\top v_2, \\
C_4 &= -\frac{2\eta}{d} \text{vec}(X - U(k))^\top v_3, \\
C_5 &= \|U(k+1) - U(k)\|_F^2.
\end{aligned}$$

Notice that C_1 can be upper bounded in terms of $\lambda_{\min}(K(k)) \geq \lambda_{\min}(H(k))$, which is ensured as long as the movement in the weight is sufficiently small (shown in Lemma 6.7.) C_2 can be upper bounded also using the kernel with bound on its spectral norm.

Proof of Theorem 6.2. We will prove Theorem 6.2 by induction. The base case when $k = 0$ is trivially true. Assume that the claim holds for $k' = 0, 1, \dots, k$ and we want to show that (6.15) also holds for $k' = k + 1$. For $k' = k + 1$, we have

$$\|X - U(k+1)\|_F^2 = \|X - U(k)\|_F^2 + C_1 + C_2 + C_3 + C_4 + C_5$$

Now, we invoke the bound for each of these terms from Claims C.4, C.5, C.6, C.7 and C.8 in Appendix C and Lemma 6.7. Then, we want to choose η and R_w such that

$$\left(1 - \frac{\eta\lambda_0}{d} + \frac{8\eta\lambda_n}{d} + \frac{16\eta^2\sqrt{n\lambda_n}}{d} + \frac{8\eta^2\sqrt{n\lambda_n^2}}{d} \|X - U(0)\|_F + \frac{64\eta^2\lambda_n^2}{md}\right) \leq \left(1 - \frac{\eta\lambda_0}{2d}\right). \quad (6.20)$$

If we set $\eta = \frac{\lambda_0}{64n\lambda_n}$ and use $\|X - U(0)\|_F \leq C\sqrt{n}$, we have the two dominating terms are

$$\frac{8\eta\lambda_n}{d} \leq \frac{\eta\lambda_0}{8d}, \quad \text{and} \quad \frac{8\eta^2\sqrt{n\lambda_n^2}}{d} \|X - U(0)\|_F \leq \frac{\eta\lambda_0}{8d}.$$

This implies that

$$\|X - U(k+1)\|_F^2 \leq \left(1 - \frac{\eta\lambda_0}{2d}\right) \|X - U(k)\|_F^2.$$

Lower bound on m . We require for any $\delta \in (0, 1)$ that

$$\frac{\lambda_n}{m}(4\|W(0)\| + R_w)R_w \leq \frac{\lambda_0}{4},$$

$$R_w \leq \frac{4\sqrt{d\lambda_n}(\|A(0)\| + R_a)\|X - U(0)\|_F}{\sqrt{m}\lambda_0}$$

and

$$2\exp(-m) \leq \delta$$

where the first bound on R_w comes from the result on gradient descent and the condition in Lemma 6.2, whereas the second bound is required by the above Claims. By Claim 5.1 that $\|X - U(0)\|_F \leq \sqrt{\frac{2n}{\delta}}$ for arbitrary $\delta \in (0, 1)$, then we require

$$m \geq C \frac{nd\lambda_n^3}{\lambda_0^4\delta^2}$$

for a sufficiently large constant $C > 0$ so that the claim holds with probability $1 - \delta$. ■

7 Weight-tied Autoencoders

We conclude with the case of training two-layer autoencoders whose weights are shared (i.e., $A = W$). This is a common architectural choice in practice, and indeed previous theoretical analysis for autoencoders [Rangamani et al., 2017, Nguyen et al., 2019, Li and Phan-Minh, 2019] have focused on this setting. We will show that somewhat surprisingly, allowing the network to be over-parameterized in this setting leads to certain degeneracies. First, we prove:

Lemma 7.1. *Let x be any fixed sample. The weight W is randomly initialized such that $w_r \sim \mathcal{N}(0, \sigma^2 I)$ independently for $r = 1, 2, \dots, m$, then*

$$\mathbb{E}_{w_r \sim \mathcal{N}(0, \sigma^2 I), \forall r} [\|x - \frac{1}{m} W \phi(W^\top x)\|^2] = \left(\frac{\sigma^2}{2} - 1\right)^2 \|x\|^2 + \frac{(2d+3)\|\sigma^2 x\|^2}{4m}.$$

Particularly, when $\|x\| = 1$, $\sigma^2 = 2$, then

$$\mathbb{E}_{w_r \sim \mathcal{N}(0, \sigma^2 I), \forall r} [\|x - \frac{1}{m} W \phi(W^\top x)\|^2] = \frac{2d+3}{m}.$$

For an arbitrary small $\epsilon > 0$, the expected reconstruction loss is at most ϵ if $m \geq \Omega(d/\epsilon)$.

Remark 7.1. This Lemma has a few interesting implications. First, when $\sigma^2 = 2$, then

$$\mathbb{E}_{w_r \sim \mathcal{N}(0, 2I), \forall r} [\|x - u\|^2] = \frac{(2d+3)\|x\|^2}{m},$$

which does not exceed ϵ if $m \geq 3d/\epsilon$ for $\epsilon > 0$. Provided that the data samples are normalized, if m is sufficiently large, even with random initialization the reconstruction loss is very close to zero *without any need for training*. Therefore, mere over-parameterization already gets us to near-zero loss; the autoencoder mapping $\frac{1}{m} W \phi(W^\top x) \approx x$ for any unit-norm x . It suggests that training of weight-tied autoencoders under high levels of over-parameterization may be degenerated.

Proof. We will use \mathbb{E}_{w_r} as a shorthand for $\mathbb{E}_{w_r \sim \mathcal{N}(0, \sigma^2 I)}$. We expand the reconstruction loss:

$$\begin{aligned} \|x - \frac{1}{m} W \phi(W^\top x)\|^2 &= \|x - \frac{1}{m} W \phi(W^\top x)\|^2 = \|x - \frac{1}{m} \sum_{r=1}^m \phi(w_r^\top x) w_r\|^2 \\ &= \|x\|^2 - \frac{2}{m} \sum_{r=1}^m x^\top \phi(w_r^\top x) w_r + \frac{1}{m^2} \sum_{r,s \in [m]} \phi(w_r^\top x) w_r^\top \phi(w_s^\top x) w_s. \end{aligned} \quad (7.1)$$

Because ϕ is ReLU and the distribution of w_r is symmetric, we have:

$$\mathbb{E}_w[\phi(w^\top x)w] = \frac{1}{2} \mathbb{E}_w[ww^\top]x = \frac{\sigma^2 x}{2},$$

since $\mathbb{E}_w[ww^\top] = I$. Then, by the independence of the columns in W (more details or split up, one more step),

$$\begin{aligned} \mathbb{E}_{w_r}[\|x - u\|^2] &= \|x\|^2 - \frac{2}{m} \sum_{r=1}^m \frac{\|x\|^2}{2} + \frac{1}{m^2} \sum_{r,s \in [m], r \neq s} \frac{\|\sigma^2 x\|^2}{4} + m \mathbb{E}_w[\phi(w^\top x)^2 \|w\|^2] \\ &= (1 - \sigma^2) \|x\|^2 + \frac{m-1}{4m} \|\sigma^2 x\|^2 + \frac{1}{m} \mathbb{E}_w[\phi(w^\top x)^2 \|w\|^2]. \end{aligned} \quad (7.2)$$

Now, we compute the last term:

$$\mathbb{E}_w[\phi(w^\top x)^2 \|w\|^2] = \frac{d+2}{2} \|\sigma^2 x\|^2. \quad (7.3)$$

Due to the normalization $\|x\| = 1$, we can also write $\frac{1}{\sigma} w = ux + v$ such that $x^\top v = 0$, then $u = \langle w, x \rangle \sim \mathcal{N}(0, 1)$ and $v \sim \mathcal{N}(0, I - xx^\top)$ are conditionally independent given x . Note that since the conditional distribution of u is unchanged with respect to x , this implies that u is independent of x ; as a result, u and v are (unconditionally) independent.

Also, denote $\alpha_q = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[z^q \mathbb{1}(z \geq 0)]$ for the exact value. Using Stein's Lemma, we can compute the exact values: $\alpha = \mathbb{E}_w[u \mathbb{1}(u \geq 0)] = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[z \mathbb{1}(z \geq 0)] = \frac{1}{\sqrt{2\pi}}$, $\beta = \mathbb{E}_w[u^2 \mathbb{1}(u \geq 0)] = \frac{1}{2}$, $\gamma = \mathbb{E}_w[u^4 \mathbb{1}(u \geq 0)] = \frac{3}{2}$, which are all positive. Write $\phi(z) = \max(0, z) = \mathbb{1}(z \geq 0)z$, and

$$\begin{aligned} \mathbb{E}_{w \sim \mathcal{N}(0, I)}[\phi(w^\top x)^2 \|w\|^2] &= \mathbb{E}_w[\mathbb{1}(\langle w, x \rangle \geq 0) \langle w, x \rangle^2 \|w\|^2] \\ &= \mathbb{E}_w[\mathbb{1}(u \geq 0) u^2 (u^2 + \|v\|^2)] \quad (\|x\| = 1) \\ &= \mathbb{E}_w[\mathbb{1}(u \geq 0) u^4] + \mathbb{E}_w[\mathbb{1}(u \geq 0) u^2] \mathbb{E}_w[\|v\|^2] \quad (\text{cond. independence of } u \text{ and } v.) \\ &= \frac{d+2}{2}. \end{aligned} \quad (7.4)$$

Changing variables by scaling the variance:

$$\mathbb{E}_{w \sim \mathcal{N}(0, \sigma^2 I)}[\phi(w^\top x)^2 \|w\|^2] = \sigma^4 \mathbb{E}_{w \sim \mathcal{N}(0, I)}[\phi(w^\top x)^2 \|w\|^2] = (2d+4) \|\sigma^2 x\|^2.$$

Combining with (7.2):

$$\mathbb{E}_{w_r \sim \mathcal{N}(0, \sigma^2 I)}[\|x - u\|^2] = \left(\frac{\sigma^2}{2} - 1\right)^2 \|x\|^2 + \frac{(2d+3) \|\sigma^2 x\|^2}{4m}. \quad (7.5)$$

The second result directly follows from the Lemma with the specific values of $\|x\|, \sigma$ plugged in. ■

References

- Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.
- Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. *arXiv preprint arXiv:1811.03962*, 2018.
- Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes overparameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.
- Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes overparameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.
- Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.
- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019a.
- Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *arXiv preprint arXiv:1906.04688*, 2019.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.
- Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.
- Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019.
- Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Memorization in overparameterized autoencoders. 2019.
- Rares-Darius Buhai, Andrej Risteski, Yoni Halpern, and David Sontag. Benefits of overparameterization in single-layer latent variable generative models. *arXiv preprint arXiv:1907.00030*, 2019.
- Akshay Rangamani, Anirbit Mukherjee, Ashish Arora, Tejaswini Ganapathy, Amitabh Basu, Sang Chin, and Trac D Tran. Sparse coding and autoencoders. *arXiv preprint arXiv:1708.03735*, 2017.
- Thanh V. Nguyen, Raymond K. W. Wong, and Chinmay Hegde. On the dynamics of gradient descent for autoencoders. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2858–2867, 16–18 Apr 2019.
- Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on Learning Theory*, pages 113–149, 2015.
- Thanh V Nguyen, Raymond K W Wong, and Chinmay Hegde. A provable approach for double-sparse coding. In *Proc. Conf. American Assoc. Artificial Intelligence*, Feb. 2018.
- Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

- Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019b.
- Ping Li and Phan-Minh Nguyen. On random deep weight-tied autoencoders: Exact asymptotic analysis, phase transitions, and implications to training. 2018.
- Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. *arXiv preprint arXiv:1611.03131*, 2016.
- Russell Tsuchida, Farbod Roosta-Khorasani, and Marcus Gallagher. Invariance of weight distributions in rectified mlps. *arXiv preprint arXiv:1711.09090*, 2017.
- Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.
- Ping Li and Nguyen Phan-Minh. On random deep weight-tied autoencoders: exact asymptotic analysis, phase transition, and implications to training. In *ICLR*, 2019.
- Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149. JMLR. org, 2017.

A Useful Facts

Lemma A.1 (Stein’s Lemma). *For a random vector $w \in \mathbb{R}^d$ such that $w \sim \mathcal{N}(0, I)$ and function $h(w) : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is weakly differentiable with Jacobian $D_w h$, we have*

$$\mathbb{E}_{w \sim \mathcal{N}(0, I)} [wh(w)^\top] = \mathbb{E}_{w \sim \mathcal{N}(0, I)} [(D_w h)^\top].$$

Lemma A.2. *Denote $S_i = \{r \in [m] : \mathbb{1}[w_r(k+1)^\top x_i \geq 0] = \mathbb{1}[w_r(k)^\top x_i \geq 0]\}$, and $S_i^\perp = [m] \setminus S_i$. If $\|w_r(k) - w_r(0)\| \leq R$, then*

$$\sum_{r=1}^m \mathbb{1}[r \in S_i^\perp] \leq 4mR$$

with probability at least $1 - n \exp(-mR)$.

This result is borrowed from the proof of [Song and Yang, 2019, Claim 4.10].

B Weakly-trained Autoencoders

B.1 Concentration of $K(0)$ – Proof of Lemma 5.1

Proof. Recall that $K(0) = \frac{1}{m} \sum_{r=1}^m \tilde{X}_r(0)^\top \tilde{X}_r(0) \otimes a_r a_r^\top$. In this proof, we omit the argument $t = 0$ in $\tilde{X}_r(0)$, and simply write \tilde{X}_r for clarity.

Consider the random matrix $Z_r = \tilde{X}_r^\top \tilde{X}_r \otimes a_r a_r^\top$ and $\bar{Z}_r = \mathbb{E}_{w_r}[\tilde{X}^\top \tilde{X}] \otimes I$. Note that Z_r is positive semi-definite. One can easily show two facts:

$$\|Z_r\| = \|\tilde{X}_r^\top \tilde{X}_r \otimes a_r a_r^\top\| = \|\tilde{X}_r^\top \tilde{X}_r\| \|a_r a_r^\top\| = \|a_r\|^2 \|\tilde{X}_r^\top \tilde{X}_r\| \leq d\lambda_n,$$

in which we use $\|a_r\|^2 = d$; and

$$\|\tilde{X}_r^\top \tilde{X}_r\| = \sup_{\|b\|=1} \|\tilde{X}_r b\|^2 \leq \sup_{\|b\|=1} \left\| \sum_i b_i x_i \right\|^2 = \|X^\top X\| = \lambda_n. \quad (\text{B.1})$$

Similarly, $\|\bar{Z}_r\| \leq \lambda_n$, and hence $\|Z_r - \bar{Z}_r\| \leq (d+1)\lambda_n$. Moreover,

$$\begin{aligned} \mathbb{E}_{w_r, a_r}[(Z_r - \bar{Z}_r)^2] &= \mathbb{E}_{w_r, a_r}[(Z_r - \bar{Z}_r)(Z_r - \bar{Z}_r)^\top] \\ &= \mathbb{E}_{w_r, a_r}[Z_r Z_r^\top] - \bar{Z}_r^2 \\ &= \mathbb{E}_{w_r, a_r}[(\tilde{X}_r^\top \tilde{X}_r)^2 \otimes \|a_r\|^2 a_r a_r^\top] - (\mathbb{E}_w[\tilde{X}^\top \tilde{X}])^2 \otimes I \\ &\leq d \mathbb{E}_{w_r}[(\tilde{X}_r^\top \tilde{X}_r)^2] \otimes I. \end{aligned}$$

By the above argument, $\|\mathbb{E}_{w_r}[(\tilde{X}_r^\top \tilde{X}_r)^2]\| \leq \lambda_n^2$, so $\|\sum_r \mathbb{E}((Z_r - \bar{Z}_r)^2)\| \leq md\lambda_n^2$.

From matrix Bernstein's inequality [Tropp, 2012, Theorem 1.4 of],

$$\mathbb{P}[\|mK(0) - mK^\infty\| \geq \epsilon] \leq nd \exp\left(-\frac{\epsilon^2/2}{(d+1)\lambda_n \epsilon/3 + md\lambda_n^2}\right).$$

Since the second term in the denominator of the exponent dominates ($\lambda_0 \leq \lambda_n$), we get

$$m \geq C \frac{\lambda_n^2 d \log(nd/\delta)}{\lambda_0^2}$$

where we pick $\epsilon = m\lambda_0/4$. Therefore,

$$\|K(0) - K^\infty\| \leq \lambda_0/4$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$. By Weyl's inequality, we have with the same probability:

$$\lambda_{\min}(K(0)) \geq 3\lambda_0/4. \quad \blacksquare$$

B.2 Proof of supporting claims

To prove the bounds in Claims B.1, B.2, B.3, and B.4, we use the bound $\|w_r(k+1) - w_r(0)\| \leq R'$ for all $r \in [m]$ in Lemma 5.5. In what follows, we assume $R' < R$, which is the weight movement allowed to achieve Lemma 5.2. This assumption holds with high probability as long as m is large enough.

Claim B.1. *Let $C_1 = -\frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k))$. Then we have*

$$C_1 \leq -\frac{\eta\lambda_0}{d} \|X - U(k)\|_F^2.$$

with probability at least $1 - \delta$.

Proof. Using Lemma 5.5, we have $\|w_r(k) - w_r(0)\| \leq R' < R$ for all $r \in [m]$. By Lemma 5.2, we have

$$\|K(k) - K(0)\| < \frac{\lambda_0}{4}.$$

Therefore, $\lambda_{\min}(K(k)) \geq \lambda_0/2$ with probability at least $1 - \delta$. As a result,

$$\text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k)) \geq \frac{\lambda_0}{2} \|X - U(k)\|^2 = \frac{\lambda_0}{2} \|X - U(k)\|_F^2,$$

and $C_1 \leq -\frac{\eta\lambda_0}{d} \|X - U(k)\|_F^2$ with probability at least $1 - \delta$. \blacksquare

Claim B.2. Let $C_2 = \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp (X - U(k))$. We have

$$C_2 \leq 8\eta n R \|X - U(k)\|_F^2.$$

with probability at least $1 - n \exp(-mR)$.

Proof. All we need is to bound $K(k)^\perp$. A simple upper bound is

$$\begin{aligned} \|K(k)^\perp\|^2 &\leq \sum_{i,j=1}^n \|K(k)_{i,j}^\perp\|_F^2 \\ &\leq \sum_{i,j=1}^n \left\| \frac{1}{m} \sum_{r \in S_i^\perp} x_i^\top x_j \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] a_r a_r^\top \right\|_F^2 \\ &\leq d^2 \sum_{i,j=1}^n \left(\frac{1}{m} \sum_{r \in S_i^\perp} \mathbb{1}[w_r(k)^\top x_i \geq 0, w_r(k)^\top x_j \geq 0] \right)^2 \\ &\leq d^2 \sum_{i,j=1}^n \left(\frac{1}{m} \sum_{r=1}^m \mathbb{1}[r \in S_i^\perp] \right)^2 \\ &\leq 16n^2 d^2 R^2 \end{aligned}$$

with probability $1 - n \exp(-mR)$ where the last step follows from Lemma A.2. Then, with that same probability $\|K(k)^\perp\| \leq \|K(k)^\perp\|_F \leq 4ndR$, and

$$\begin{aligned} C_2 &= \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp (X - U(k)) \\ &\leq \frac{2\eta}{d} \|K(k)^\perp\| \|X - U(k)\|_F^2 \\ &\leq 8\eta n R \|X - U(k)\|_F^2. \end{aligned}$$

■

Claim B.3. Let $C_3 = -2\text{vec}(X - U(k))^\top v_2$, then with probability at least $1 - n \exp(-mR)$

$$C_3 \leq 8\eta n R \|X - U(k)\|_F^2.$$

Proof. We have $C_3 \leq 2\|X - U(k)\|_F \|v_2\|$. Using the Lipschitz property of ϕ , we have

$$\begin{aligned}
\|v_2\|^2 &= \sum_{i=1}^n \|v_{2,i}\|^2 \\
&\leq \sum_{i=1}^n \left\| \frac{1}{\sqrt{md}} \sum_{r \in S_i^\perp} a_r (\phi(w_r(k+1)^\top x_i) - \phi(w_r(k)^\top x_i)) \right\|^2 \\
&\leq \frac{\eta^2}{m} \sum_{i=1}^n \left(\sum_{r \in S_i^\perp} \left| (\nabla_{w_r} L(W(k)))^\top x_i \right| \right)^2 \\
&\leq \frac{\eta^2}{m} \sum_{i=1}^n \left(\sum_{r=1}^m \mathbb{1}[r \in S_i^\perp] \left| (\nabla_{w_r} L(W(k)))^\top x_i \right| \right)^2 \\
&\leq \frac{\eta^2}{m} \max_r \left\| \nabla_{w_r} L(W(k)) \right\|^2 \sum_{i=1}^n \left(\sum_{r=1}^m \mathbb{1}[r \in S_i^\perp] \right)^2 \\
&\leq \frac{\eta^2 \lambda_n}{m^2} \|X - U(k)\|_F^2 \sum_{i=1}^n \left(\sum_{r=1}^m \mathbb{1}[r \in S_i^\perp] \right)^2 \\
&\leq \frac{\eta^2 \lambda_n}{m^2} \|X - U(k)\|_F^2 \sum_{i=1}^n (4mR)^2 \\
&\leq 16n\lambda_n R^2 \eta^2 \|X - U(k)\|_F^2 \\
&\leq 16n^2 R^2 \eta^2 \|X - U(k)\|_F^2.
\end{aligned}$$

with probability $1 - n \exp(-mR)$. The sixth step we use

$$\begin{aligned}
\|\nabla_{w_r} L(W(k))\| &= \left\| \frac{1}{\sqrt{md}} \tilde{X}_r(k) (X - U(k))^\top a_r \right\| \\
&\leq \frac{\sqrt{\lambda_n}}{\sqrt{m}} \|X - U(k)\|_F,
\end{aligned}$$

and the last step follows from Lemma A.2 that $\sum_{r=1}^m \mathbb{1}[r \in S_i^\perp] \leq 4mR$ with probability at least $1 - n \exp(-mR)$. Substitute the bound into C_3 , then we finish the proof. \blacksquare

Claim B.4. Let $C_4 = \|U(k+1) - U(k)\|_F^2$. Then we have

$$C_4 \leq \eta^2 n \lambda_n \|X - U(k)\|_F^2.$$

Proof. Previously in Lemma 5.5, we proved that

$$\begin{aligned}
\|\nabla_{w_r} L(W(k))\| &= \left\| \frac{1}{\sqrt{md}} \tilde{X}_r(k) (X - U(k))^\top a_r \right\| \\
&\leq \frac{\sqrt{\lambda_n}}{\sqrt{m}} \|X - U(k)\|_F.
\end{aligned}$$

Expand the form of $U(k+1) - U(k)$ and use the Lipschitz of ReLU to get

$$\begin{aligned}
C4 &= \sum_{i=1}^n \|u_i(k+1) - u_i(k)\|^2 \\
&= \frac{1}{md} \sum_{i=1}^n \left\| \sum_{r=1}^m a_r (\phi(w_r(k+1)^\top x_i) - \phi(w_r(k)^\top x_i)) \right\|^2 \\
&\leq \eta^2 \sum_{i=1}^n \frac{1}{m} \left(\sum_{r=1}^m \left\| \nabla_{w_r} L(W(k)) \right\| \right)^2 \\
&\leq \eta^2 \sum_{i=1}^n \frac{1}{m} \left(\sum_{r=1}^m \frac{\sqrt{\lambda_n}}{\sqrt{m}} \|X - U(k)\|_F \right)^2 \\
&= \eta^2 n \lambda_n \|X - U(k)\|_F^2 \\
&\leq n^2 \eta^2 \|X - U(k)\|_F^2.
\end{aligned}$$

Therefore, we finish the proof. ■

C Jointly-trained Autoencoders

C.1 Concentration of $H(0)$

We re-state and prove the concentration of $H(0)$ in Lemma 6.1.

Lemma C.1. *For any $\delta \in (0, \frac{\lambda_0}{12d\lambda_n^2})$, if $m \geq C \frac{\max(n,d)\lambda_n \log^2(nd/\delta)}{\lambda_0^2}$ for some large enough constant C , then with probability at least $1 - 1/(2nd)^{2 \log nd} - m\delta$, one obtains $\|H(0) - H^\infty\| \leq \lambda_0/4$ and $\lambda_{\min}(H(0)) \geq 3\lambda_0/4$.*

Proof. Recall that

$$H(0) = \frac{1}{m} \sum_{r=1}^m \phi(X^\top w_r(0)) \phi(w_r(0)^\top X) \otimes I, \quad (\text{C.1})$$

and $H^\infty = \mathbb{E}_{W(0), A(0)}[H(0)]$. Our goal is to show the concentration of $\sum_{r=1}^m \phi(X^\top w_r(0)) \phi(w_r(0)^\top X)$. Let us use w_r to mean $w_r(0)$ and denote the r^{th} -random matrix as

$$Z_r = \phi(X^\top w_r) \phi(w_r^\top X).$$

We use Lemma B.7 of Zhong et al. [2017] and verify the required conditions by the next results.

Claim C.1 (Condition I for $H(0)$). *The following are true:*

- (i). $\|Z_r\| \leq \lambda_n \|w_r\|^2$.
- (ii). $\mathbb{P}_{w_r}[\|w_r\|^2 \leq 4d\sqrt{\log(2/\delta)}] \geq 1 - \delta$ for any $\delta \in (0, 1)$.

Proof. (i) We have

$$\|Z_r\| = \|\phi(X^\top w_r) \phi(w_r^\top X)\| = \|\phi(X^\top w_r)\|^2 \leq \|X^\top X\| \|w_r\|^2 \leq \lambda_n \|w_r\|^2,$$

which gives the first part (i). For the second part, we use the fact $\|w_r\|^2$ is a chi-squared random variable with d degrees of freedom, and sub-exponential with sub-exponential norm $2\sqrt{d}$, meaning that:

$$\mathbb{P}_{w_r}[\|\|w_r\|^2 - d\| \geq \epsilon] \leq 2 \exp\left(-\frac{\epsilon^2}{8d}\right)$$

For any $\delta \in (0, 1)$ and $\epsilon^2 = 9d \log(2/\delta) \geq 8d \log(2/\delta)$, we have

$$|\|w_r\|^2 - d| \leq 3\sqrt{d \log(2/\delta)}$$

with probability at least $1 - \delta$. Then, $\|w_r\|^2 \leq d + 3\sqrt{d \log(2/\delta)} \leq 4d\sqrt{\log(2/\delta)}$ with probability at least $1 - \delta$. ■

Claim C.2 (Condition II for H(0)). $\|\mathbb{E}[Z_r Z_r^\top]\| \leq 3n\lambda_n$.

Proof. We have

$$\begin{aligned} Z_r Z_r^\top &= \phi(X^\top w_r) \phi(w_r^\top X) \phi(X^\top w_r) \phi(w_r^\top X) \\ &\preceq \|\phi(X^\top w_r)\|^2 \|w_r\|^2 X^\top X \preceq \sum_{l=1}^n \phi(w^\top x_l)^2 \|w_r\|^2 X^\top X. \end{aligned}$$

We need is to compute $\mathbb{E}_{w_r}[\phi(w^\top x_l)^2 \|w_r\|^2]$, which is already done in (7.3), Section 7. To be precise, we have

$$\mathbb{E}[\phi(w_r^\top x_l)^2 \|w_r\|^2] = \frac{d+2}{2} \leq d$$

where $\tilde{x}_i = x_i 1[w_r^\top x_i \geq 0]$ for each $i \in n$. Then, we can write

$$\|\mathbb{E}_{w_r}[Z_r Z_r^\top]\| \leq nd\lambda_n. \quad \blacksquare$$

Claim C.3 (Condition IV for H(0)). $\sup_{\{b: \|b\|=1\}} (\mathbb{E}[(b^\top Z_r b)^2])^{1/2} \leq \sqrt{3}d\lambda_n$.

Proof. Recall $Z_r = \phi(X^\top w_r) \phi(w_r^\top X)$, and for any unit-norm vector $b \in \mathbb{R}^n$

$$(b^\top Z_r b)^2 = \|b^\top \phi(X^\top w_r)\|^4 \leq \|\phi(X^\top w_r)\|^4 \leq \lambda_n^2 \|w_r\|^4.$$

Moreover, $\|w_r\|^2$ is a chi-squared random variable with d degree of freedom, so

$$\mathbb{E}[\|w_r\|^4] = 3d^2.$$

Therefore, $\sup_{\{b: \|b\|=1\}} (\mathbb{E}[(b^\top Z_r b)^2])^{1/2} \leq \sqrt{3}d\lambda_n. \quad \blacksquare$

Proof of Lemma 6.1. With the conditions fulfilled in Claims C.1, C.2 and C.3, we can now apply [Zhong et al., 2017, Lemma B.7] to show the concentration of $K(0)$:

$$\left\| \frac{1}{m} \sum_{r=1}^m Z_r - \mathbb{E}[Z_r] \right\| \leq \epsilon \|\mathbb{E}[Z_r]\|$$

with probability $1 - 1/n^{2t} - n\delta$ for any $t \geq 1$, $\epsilon \in (0, 1)$ and $\delta < \epsilon \|\mathbb{E}[Z_r]\| / (2\sqrt{3}d\lambda_n)^2$.

For the target bound, we choose $\epsilon \|\mathbb{E}[Z_r]\| = \lambda_0/4$, $t = \log(2nd)$ and note that $\lambda_0 \leq \|\mathbb{E}[Z_r]\| = \|K^\infty\| \leq \lambda_n$. Therefore, with probability $1 - 1/(2nd)^{2 \log nd} - m\delta$ for any $\delta \in (0, \frac{\lambda_0}{12d^2\lambda_n^2})$, then

$$\|H(0) - H^\infty\| \leq \frac{\lambda_0}{4}$$

if m satisfies

$$m \geq 18 \log^2(2nd) \frac{nd\lambda_n + \lambda_n^2 + (4d\sqrt{\log(2/\delta)})\lambda_n\lambda_0/4}{\lambda_0^2},$$

which means $m = C \frac{nd\lambda_n \log^2(nd) \log(1/\delta)}{\lambda_0^2}$. ■

C.2 Proof of supporting claims

In the proof of the next claims, we assume that $\|W(0)\| \geq R'_w > R_w$ and $\|A(0)\| \geq R'_a > R_a$. Also, assume $d \ll m$. These conditions will hold with high probability when m is large enough.

Claim C.4. Let $C_1 = -\frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k))$. We have

$$C_1 \leq -\frac{\eta\lambda_0}{d} \|X - U(k)\|_F^2.$$

Proof. Since we have proved that $\|W(k) - W(0)\|_F \leq R'_w$, using Lemma 6.2 with the choice of $R_w < R'_w$, we have

$$\|H(k) - H(0)\| \leq \frac{\lambda_0}{4}.$$

Moreover, $G(k)$ is p.s.d, therefore $\lambda_{\min}(K(k)) \geq \lambda_{\min}(H(k)) \geq \lambda_0/2$, and as a result,

$$\text{vec}(X - U(k))^\top K(k) \text{vec}(X - U(k)) \geq \frac{\lambda_0}{2} \|X - U(k)\|^2 = \frac{\lambda_0}{2} \|X - U(k)\|_F^2.$$

and $C_1 \leq -\frac{\eta\lambda_0}{d} \|X - U(k)\|_F^2$. ■

Claim C.5. Let $C_2 = \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp \text{vec}(X - U(k))$. We have

$$C_2 \leq \frac{8\eta\lambda_n}{d} \|X - U(k)\|_F^2$$

with probability at least $1 - 2\exp(-m)$.

Proof. We need to bound the spectral norm of $K(k)^\perp$, defined as $K(k)^\perp = G(k)^\perp + H(k)^\perp$. We will bound their spectral norms. We have

$$\begin{aligned} \|G(k)^\perp\| &= \left\| \frac{1}{m} \sum_{r=1}^m \text{diag}(\mathbb{1}[r \in S_i^\perp]) \tilde{X}_r^\top \tilde{X}_r \otimes a_r(k) a_r(k)^\top \right\| \\ &\leq \frac{1}{m} \|X\|^2 \left\| \sum_{r=1}^m a_r(k) a_r(k)^\top \mathbb{1}[r \in S_i^\perp] \right\| \\ &\leq \frac{\lambda_n}{m} \|A(k)\|^2 \\ &\leq \frac{4\lambda_n \|A(0)\|^2}{m}, \end{aligned}$$

where we use the assumption $R_a \leq \|A(0)\|$. Similarly, using $R_w \leq \|W(0)\|$ we have

$$\begin{aligned} \|H(k)^\perp\| &= \left\| \frac{1}{m} \sum_{r=1}^m \text{diag}(\mathbb{1}[r \in S_i^\perp]) \phi(X^\top w_r(k)) \phi(w_r(k)^\top X) \otimes I \right\| \\ &\leq \frac{1}{m} \|X\|^2 \|W(k)\|^2 \\ &\leq \frac{4\lambda_n \|W(0)\|^2}{m}. \end{aligned}$$

Moreover, using a standard bound on sub-Gaussian matrices, we have $\|W(0)\| \leq 2\sqrt{m} + \sqrt{d}$ and $\|A(0)\| \leq 2\sqrt{m} + \sqrt{d}$ with probability at least $1 - 2\exp(-m)$. Then,

$$\begin{aligned} C_2 &= \frac{2\eta}{d} \text{vec}(X - U(k))^\top K(k)^\perp \text{vec}(X - U(k)) \\ &\leq \frac{8\eta\lambda_n}{d} \end{aligned}$$

with probability at least $1 - 2\exp(-m)$. ■

We re-state the results in the proof in Lemma 6.4 and Lemma 6.5 to bound the remaining terms C_3, C_4, C_5 :

$$\|\nabla_W L(W(k), A(k))\|_F \leq \frac{\sqrt{\lambda_n}}{\sqrt{md}} \|X - U(k)\|_F \|A(k)\|. \quad (\text{C.2})$$

$$\|\nabla_A L(W(k), A(k))\|_F \leq \frac{\sqrt{\lambda_n}}{\sqrt{md}} \|X - U(k)\|_F \|W(k)\|. \quad (\text{C.3})$$

Claim C.6. Let $C_3 = -\frac{2\eta}{d} \text{vec}(X - U(k))^\top v_2$. We have

$$C_3 \leq \frac{16\eta^2}{d} \sqrt{n\lambda_n} \|X - U(k)\|_F^2$$

with probability at least $1 - 3\exp(-m)$.

Proof. We have $\text{vec}(X - U(k))^\top v_2 \leq \|v_2\| \|X - U(k)\|_F$, so we need to bound $\|v_2\|$. Let $D_i = \text{diag}(\mathbb{1}[1 \in S_i^\perp], \dots, \mathbb{1}[m \in S_i^\perp])$, then:

$$\begin{aligned} \|v_2\|^2 &= \sum_{i=1}^n \|v_{2,i}\|^2 \\ &= \sum_{i=1}^n \left\| \frac{1}{\sqrt{md}} \sum_{r \in S_i^\perp} (a_r(k+1)\phi(w_r(k+1)^\top x_i) - a_r(k)\phi(w_r(k)^\top x_i)) \right\|^2 \\ &= \frac{1}{md} \sum_{i=1}^n \left\| A(k+1)D_i\phi(W(k+1)^\top x_i) - A(k)D_i\phi(W(k)^\top x_i) \right\|^2 \\ &\leq \frac{2\eta^2}{md} \sum_{i=1}^n \left\| (\nabla_A L)D_i\phi(W(k+1)^\top x_i) \right\|^2 + \left\| A(k)D_i(\nabla_W L)^\top x_i \right\|^2 \\ &\leq \frac{2n\eta^2}{md} \left(\|\nabla_A L\|_F^2 \|W(k+1)\|^2 + \|A(k)\|^2 \|\nabla_W L\|_F^2 \right) \\ &\leq \frac{2n\lambda_n\eta^2}{m^2d^2} \|X - U(k)\|_F^2 \left(\|W(k+1)\|^2 \|W(k)\|^2 + \|A(k)\|^4 \right) \\ &\leq \frac{64n\lambda_n\eta^2}{d^2} \|X - U(k)\|_F^2 \end{aligned}$$

with probability at least $1 - 3\exp(-m)$. Since $\|W(k+1)\| \leq 2\|W(0)\|$ and $\|A(k+1)\| \leq 2\|A(0)\|$. Therefore, with that probability

$$C_3 \leq \frac{16\eta^2}{d} \sqrt{n\lambda_n} \|X - U(k)\|_F^2. \quad \text{■}$$

Claim C.7. Let $C_4 = -\frac{2\eta}{d} \text{vec}(X - U(k))^\top v_3$. We have

$$C_4 \leq \frac{8\eta^2}{d} \sqrt{n\lambda_n^2} \|X - U(k)\|_F^2 \|X - U(0)\|_F$$

with probability at least $1 - 2\exp(-m)$.

Proof. We have $\text{vec}(X - U(k))^\top v_3 \leq \|v_3\| \|X - U(k)\|_F$. We want to bound $\|v_3\|$. Let $D'_i = \text{diag}(\mathbb{1}[w_1(k)^\top x_i \geq 0], \dots, \mathbb{1}[w_m(k)^\top x_i \geq 0])$

$$\begin{aligned}
\|v_3\|^2 &= \sum_{i=1}^n \|v_{3,i}\|^2 \\
&= \sum_{i=1}^n \left\| \frac{\eta^2}{\sqrt{md}} \sum_{r \in S_i} (\nabla_{a_r} L) (\nabla_{w_r} L)^\top x_i \mathbb{1}[w_r(k)^\top x_i \geq 0] \right\|^2 \\
&\leq \frac{\eta^4}{md} \left\| (\nabla_A L) ((\nabla_W L) D'_i)^\top x_i \right\|^2 \\
&= \frac{\eta^4 n}{md} \|\nabla_A L\|_F^2 \|\nabla_W L\|_F^2 \\
&\leq \frac{\eta^4 n \lambda_n^2}{m^2 d^2} \|X - U(k)\|_F^4 \|A(k)\|^2 \|W(k)\|^2 \\
&\leq \frac{16\eta^4 n \lambda_n^2}{d^2} \|X - U(k)\|_F^4
\end{aligned}$$

with probability at least $1 - 2 \exp(-m)$. Therefore,

$$C_3 \leq \frac{8\eta^2 \sqrt{n\lambda_n^2}}{d} \|X - U(k)\|_F^3 \leq \frac{8\eta^2 \sqrt{n\lambda_n^2}}{d} \|X - U(k)\|_F^2 \|X - U(0)\|_F$$

since $\|X - U(k)\|_F \leq \|X - U(0)\|_F$ by the induction hypothesis. ■

Claim C.8. Let $C_5 = \|U(k+1) - U(k)\|_F^2$. Then we have

$$C_5 \leq \frac{64\eta^2 \lambda_n^2}{md} \|X - U(k)\|_F^2.$$

with probability at least $1 - 3 \exp(-m)$.

Proof. We bound this by re-iterating the proof of Claim C.6:

$$\begin{aligned}
\|U(k+1) - U(k)\|_F^2 &= \frac{1}{m^2} \|A(k+1) \phi(W(k+1)^\top X) - A(k) \phi(W(k)^\top X)\|_F^2 \\
&\leq \frac{2\eta^2}{m^2} \left(\left\| (\nabla_A L) \phi(W(k+1)^\top X) \right\|^2 + \|A(k)\|^2 \left\| (\nabla_W L)^\top X \right\|^2 \right) \\
&\leq \frac{2\eta^2 \|X\|^2}{m^2} \left(\|\nabla_A L\|_F^2 \|W(k+1)\|^2 + \|A(k)\|^2 \|\nabla_W L\|_F^2 \right) \\
&\leq \frac{2\eta^2 \lambda_n^2}{m^3 d} \|X - U(k)\|_F^2 \left(\|W(k+1)\|^2 \|W(k)\|^2 + \|A(k)\|^4 \right) \\
&\leq \frac{64\eta^2 \lambda_n^2}{md} \|X - U(k)\|_F^2
\end{aligned}$$

with probability at least $1 - 3 \exp(-m)$. ■