Associative Memory in Iterated Overparameterized Sigmoid Autoencoders

Yibo Jiang 1  Cengiz Pehlevan 2

Abstract

Recent work showed that overparameterized autoencoders can be trained to implement associative memory via iterative maps, when the trained input-output Jacobian of the network has all of its eigenvalue norms strictly below one. Here, we theoretically analyze this phenomenon for sigmoid networks by leveraging recent developments in deep learning theory, especially the correspondence between training neural networks in the infinite-width limit and performing kernel regression with the Neural Tangent Kernel (NTK). We find that overparameterized sigmoid autoencoders can have attractors in the NTK limit for both training with a single example and multiple examples under certain conditions. In particular, for multiple training examples, we find that the norm of the largest Jacobian eigenvalue drops below one with increasing input norm, leading to associative memory.

1. Introduction

The mechanisms behind memory have been a long interest of neuroscientists. Hopfield’s seminal work proposed that associative memory can be implemented by attractor neural dynamics (Hopfield, 1982), and has been the dominant model that shapes thinking in this domain (Hertz, 2018). Recently, Radhakrishnan et al. (2019; 2018) proposed an alternative mechanism and showed that overparameterized autoencoders trained with gradient descent could also implement associative memory in an iterative fashion. These networks are reported to be easy to train and to not suffer from spurious attractors, unlike Hopfield networks (Amit & Treves, 1989; Hertz, 2018). The potential benefits of this approach make a theoretical account of it necessary, which we aim to provide here.

We study auto-encoders in a limit of neural networks that makes theoretical analysis possible. Specifically, as the width of the hidden layers of a feedforward neural network is taken to infinity with a particular initialization scheme, its training dynamics simplifies and can be described by ridgeless kernel interpolation with a kernel called the Neural Tangent Kernel (NTK) (Jacot et al., 2018). Working in the NTK limit, we examine the input-output Jacobian matrices of the trained networks as they control the stability of trained fixed points. We focus on networks with sigmoid activation functions, but our results can be extended to other sigmoidal functions such as erf and tanh using similar techniques. We make a distinction between the cases of a single training example and multiple training examples, which exhibit different memory behaviors.

Our main contributions and results are summarized below:

• First, we analyze autoencoders in the NTK limit trained on a single training example. We argue that the trained Jacobian will stay close to initialization under certain conditions. Therefore, if the initial Jacobian has all eigenvalue norms smaller than 1, this training example will be an attractor.

• Next, we specialize to 2-layer networks. We show that when the norms of training examples are small, attractor formation can fail due to the presence of eigenvalue 1 in the spectrum of trained Jacobian matrices.

• We show that a 2-layer network can have attractors when the norms of training examples are large as the induced NTK relies more on the non-linear, saturated region of the activation function. This suggests that the network transitions from a regime where attractor formation fails to a regime where it succeeds as the input norm grows.

• We verify the predictions of our theoretical results in simulations.

Many previous works on generalization (Allen-Zhu et al., 2019; Cao & Gu, 2019) and training (Jacot et al., 2018; Du et al., 2018; Allen-Zhu et al., 2018) in the NTK limit focus on input data on the unit sphere which is violated in practice (Krizhevsky et al., 2012). We highlight how the input norm can set trained neural networks in different learning regions with respect to the trained input-output Jacobian.

Associative memory behavior induced by training overparameterized autoencoders could provide insights into the
Weights: In a set of ideas that we will make use of later, we review relevant literature from these domains. The basic idea is that, if initialized properly, in the infinite width limit, parameters of the network stay close to initialization (Jacot et al., 2018). Thus, NTK stays relatively constant throughout training. Because NTK governs the training dynamics, positive-definite kernel ensures global convergence of optimization. Subsequent papers expanded this idea to finite network widths, gradient descent or stochastic gradient descent as opposed to gradient flow, and different loss functions for regression and classification (Du et al., 2018; Allen-Zhu et al., 2018; Zou & Gu, 2019; Lee et al., 2019). Also related is research pointing that neural networks at initialization in the infinite width limit behave as Gaussian Processes (Lee et al., 2017; Matthews et al., 2018).

Neural Tangent Kernel: We first review literature on neural networks optimization, especially the NTK theory which we will use extensively through this paper. Training of neural networks poses a challenging non-convex optimization problem. Analysis simplifies if focused on the linearized training dynamics of gradient flow using the NTK theory (Jacot et al., 2018). The basic idea is that, if initialized properly, in the infinite width limit, parameters of the network stay close to initialization (Chizat et al., 2019). Thus, NTK stays relatively constant throughout training. Because NTK governs the training dynamics, positive-definite kernel ensures global convergence of optimization. Subsequent papers expanded this idea to finite network widths, gradient descent or stochastic gradient descent as opposed to gradient flow, and different loss functions for regression and classification (Du et al., 2018; Allen-Zhu et al., 2018; Zou & Gu, 2019; Lee et al., 2019). Also related is research pointing that neural networks at initialization in the infinite width limit behave as Gaussian Processes (Lee et al., 2017; Matthews et al., 2018).

Signal Propagation in Deep Networks with Random Weights: In a set of ideas that we will make use of later, Poole et al. (2016) developed a mean-field formalism to study layer-to-layer propagation of activation variances and covariances in deep networks with random weights. This line of work (Poole et al., 2016; Schoenholz et al., 2016) identifies a phase transition between ordered and chaotic regime, where nearby input points converge or diverge as they propagate through the layers, induced by different variances of weight and bias initializations. Using random matrix theories developed for Gram matrices of neural networks (Pennington & Worah, 2017), Pennington et al. (2017; 2018) calculate singular value spectra for input-output Jacobians at initialization and identify its relation to ordered/chaotic training regime. These results cannot be applied to our problem, as they are in a different setting assuming a large depth limit such that the variances for all layers are at the fixed points of the layer-to-layer iterative maps.

Generalization and Implicit Bias: Associative memory behavior of neural networks can provide insight into generalization of deep learning and implicit bias of gradient descent. Here, we review some recent works in this area, focusing on overparameterized networks and NTK. A pair of recent papers (Hayou et al., 2019; Xiao et al., 2019) demonstrate how to use the theories developed for signal propagation to understand NTK regression better and thus trainability and generalization of neural networks. Methods derived from statistical physics also give insight (Cohen et al., 2019; Bordelon et al., 2020). However, there is a gap in terms of generalization between NTK regression and training neural networks (Allen-Zhu & Li, 2019; Arora et al., 2019) (see however (Lee et al., 2019)), which prompts research on generalization in deep learning beyond NTK (Allen-Zhu et al., 2019; Bai & Lee, 2019). Another line of research on generalization looks at the implicit bias of gradient descent. Gradient descent on logistic regression can lead to the max-margin solution (Soudry et al., 2018; Ji & Telgarsky, 2018), while optimization on mean squared regression has a shortest path solution (Oymak & Soltanolkotabi, 2018).

3. Preliminaries

In this section, we set up our notation and review background material.

3.1. Neural Networks

The output of a neural network is defined by $f(x) = \tilde{\alpha}^{(L)}(x)$, where the functions $\tilde{\alpha}^{(t)}(\cdot) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_t}$ (pre-activations) and $\alpha^{(t)}(\cdot) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_t}$ (activations) follow the recursive relation:

$$\alpha^{(0)}(x) \equiv x$$

$$\tilde{\alpha}^{(t+1)}(x) \equiv \frac{1}{\sqrt{n_t}} W^{(t)} \alpha^{(t)}(x), \quad \alpha^{(t)}(x) \equiv \sigma(\tilde{\alpha}^{(t)}(x)),$$

where $\sigma$ is an element-wise activation function and weights are initialized by sampling from an i.i.d. standard Gaussian. We are mostly interested in

$$\text{sigmoid} = \frac{1}{1 + e^{-x}}$$  \hspace{1cm} (1)$$

as the activation function. We drop the bias term for simplicity. We expect our results to be qualitatively the same with a bias term for sigmoid as the behavior is governed by the activation’s shape. Throughout the paper, we will use $f_0(x)$ and $f_\infty(x)$ to denote neural networks at initialization and training to zero loss respectively. As shown later, the
attractor behavior is mostly governed by the shape of the sigmoidal activations.

**Inputs:** Given \( n \) training points \( \{x_i\}_{i=1}^{n} \), we define the following two matrices.

\[
\hat{X} = \begin{pmatrix} x_1 & \cdots & x_n \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots \end{pmatrix}, \quad f(\hat{X}) = \begin{pmatrix} f(x_1) & \cdots & f(x_n) \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots \end{pmatrix},
\]

where \( \hat{X} \in \mathbb{R}^{n \times n} \) is the data matrix and each column of \( f(\hat{X}) \in \mathbb{R}^{n \times n} \) is the output of the corresponding training example. We further assume that all input examples share the same norm (i.e., \( \forall i \| x_i \|_2 = r \)).

**Jacobian Matrix:** Given the network setup, the input-output Jacobian matrix can be computed as:

\[
J(x) = \frac{1}{\sqrt{n_L}} W^{(L)} \prod_{k=1}^{L} \left( \frac{1}{\sqrt{n_{k-1}}} W^{(k-1)} \right)
\]

where

\[
D^{(k)} = \text{diag}(\sigma'(\sigma(x))).
\]

Here \( ' \) denotes first derivative, and \text{diag} takes in a vector and outputs a diagonal matrix with the vector at the diagonal. We will use \( J_0(x) \) and \( J_{\infty}(x) \) to denote Jacobian at initialization and training to zero loss respectively.

**Autoencoder:** An autoencoder network is trained via gradient flow to optimize the following loss function:

\[
\arg \min_f \frac{1}{2n} \sum_{i=1}^{n} \| f(x_i) - \hat{x}_i \|^2_2,
\]

where \( f \) is the network defined above.

### 3.2. Neural Tangent Kernel

In the large-width regime, the neural network \( f \) can be approximated by a linearization with respect to its parameters \( \theta \) (Lee et al., 2017):

\[
f(x; \theta_t) \approx f_0(x) + \theta f(x)|_{\theta = \theta_0} (\theta_t - \theta_0),
\]

where \( \theta_0 \) and \( \theta_t \) are vectors of the network parameters at initialization and time \( t \). The first term remains unchanged throughout training. Moreover, we can view \( \theta f(x)|_{\theta = \theta_0} \) as a feature map in Hilbert space and derive the following matrix kernel,

\[
(\Theta^{(L)}_0(\hat{x}, x))_{d'd'} = \left\langle \partial_x f d(x)|_{\theta = \theta_0}, \partial_x f d'(x)|_{\theta = \theta_0} \right\rangle,
\]

where \( \langle \cdot, \cdot \rangle \) denotes inner product. Indeed, in the infinite width limit (Jacot et al., 2018), \( \Theta^{(L)}_0 \) converges in probability (stochasticity induced by random initialization) to a deterministic limiting kernel, \( \Theta^{(L)}_\infty(\hat{x}, x) \rightarrow \Theta^{(L)}_\infty(\hat{x}, x)I_{n_L} \), where \( \Theta^{(L)}_\infty \) is a scalar kernel and the training dynamics is entirely governed by it.

On the other hand, it can also be shown that in the NTK limit (\( n_1, \ldots, n_L \rightarrow \infty \) sequentially) and when the weights are initialized by sampling from an i.i.d. standard Gaussian distribution, output functions at each layer tend to be i.i.d. centered Gaussian Processes at initialization (Lee et al., 2017), and the covariance matrix of layer \( \ell \) can be defined recursively by

\[
\Sigma^{(1)}(\hat{x}, x) = \frac{1}{n_0} \hat{x}^T x,
\]

\[
\Sigma^{(\ell+1)}(\hat{x}, x) = \mathbb{E}_{g \sim \mathcal{N}(0, \Sigma^{(\ell)})} [\sigma(g(x))\sigma(g(\hat{x}))].
\]

Under the same limit, \( \Theta^{(L)}_\infty \) can also be recursively defined (Jacot et al., 2018):

\[
\Theta^{(1)}(\hat{x}, x) = \Sigma^{(1)}(\hat{x}, x),
\]

\[
\Theta^{(\ell+1)}(\hat{x}, x) = \Theta^{(\ell)}(\hat{x}, x) \Sigma^{(\ell+1)}(\hat{x}, x) + \Sigma^{(\ell+1)}(\hat{x}, x),
\]

where

\[
\Sigma^{(\ell+1)}(\hat{x}, x) = \mathbb{E}_{g \sim \mathcal{N}(0, \Sigma^{(\ell)})} [\sigma(g(x))\sigma(g(\hat{x}))].
\]

For gradient flow training with a least squares loss to zero training error, there is a closed form solution for \( f_\infty(x) \) using NTK in the infinite width limit (Jacot et al., 2018). In the case of autoencoder, we get that,

\[
f_\infty(x) = \left( \hat{X} - f_0(\hat{X}) \right) \hat{K}^{-1} k_x + f_0(x),
\]

and

\[
J_\infty(x) = \left( \hat{X} - f_0(\hat{X}) \right) \hat{K}^{-1} \frac{\partial k_x}{\partial x} + J_0(x),
\]

where

\[
\hat{K}_{ij} = \Theta^{(L)}_\infty(x_i, x_j), \quad (k_x)_i = \Theta^{(L)}_\infty(x_i, x).
\]

### 3.3. Iterative Maps, Attractors, Associative Memory and Jacobian

We define attractors with respect to an iterative map. Notice that an autoencoder \( f \) is a map from \( \mathbb{R}^{n_0} \) to \( \mathbb{R}^{n_0} \). Therefore, we can apply \( f \) iteratively to input \( x \). Formally, we define this sequence for any input \( x \), \( \{f^k(x)\}_{k \in \mathbb{N}} \) where \( f^k = f(\ldots f(f(x))) \).

**Definition 1.** A fixed point \( x^* \) of map \( f(f(x^*) = x^*) \) is an attractor if there exists an open neighborhood of \( x^* \) such that for any \( x \) in this neighborhood, \( \{f^k(x)\}_{k \in \mathbb{N}} \) converges to \( x^* \) as \( k \rightarrow \infty \). The set of all such points is called basin of attraction of \( x^* \)
Fixed points attractors can be used to implement associative memory (Hopfield, 1982). A memory clue sets the initial condition of the network dynamics, positioning the network state in a basin of attraction, and the actual memory is recapitulated by the attractor dynamics converging to the corresponding fixed point.

There is a well-know condition for a fixed point to be an attractor (Rudin et al., 1964).

**Proposition 1.** A fixed point \( x^\ast \) is an attractor of a differentiable map \( f \) if all eigenvalues of the Jacobian of \( f \) at \( x^\ast \) are strictly less than 1 in absolute value.

Therefore, for a point \( x \) to be an attractor, we need two conditions: (1) \( f(x) = x \) and (2) all the eigenvalues of \( J(x) \) have norm strictly smaller than 1. Condition (1) can be justified in the NTK limit (Jacot et al., 2018; Du et al., 2018; Allen-Zhu et al., 2018) as long as \( K \) is positive definite. This is theoretically true if all data points live on a sphere, the network has non-polynomial Lipschitz activation function and \( L \geq 2 \) (c.f. Proposition 2 in (Jacot et al., 2018)). In practice, it is easy to achieve \( f(x) \approx x \) in overparameterized networks. Therefore, our focus is on the second condition.

4. Theoretical Results

In this section, we present our theoretical results about the attractor behavior of iterated overparameterized autoencoders. We focus on two distinct settings:

1. In the first setting, we consider a neural network of any depth in the NTK limit with a single training example.
2. In the second setting, we focus on a two-layer network in the NTK limit and multiple training examples.

We first give two key results about the Jacobian norm at network initialization that will be used later. We use the operator norm (induced \( l_2 \)-norm) of Jacobian as a proxy to control the eigenvalue, because the norm is given by the largest singular value, which upper bounds the norm of the largest eigenvalue.

**4.1. A Bound on the Norm of the Jacobian at Network Initialization**

In this section, we first show that with high probability, the operator norm of the initial Jacobian for sigmoid networks drops with increasing depth. To prove this, we use a mathematical technique called the \( \varepsilon \)-net argument (Tao, 2012) (Theorem 1) valid for any activation function. We then argue that for sigmoid networks the upper bound of the initial Jacobian norm is concentrated around 0.5 for large \( n_0 \). This explains the formation of attractors when the trained Jacobian norm stays close to initialization (c.f. Section 4.2).

We first prove a proposition to be used in the \( \varepsilon \)-net argument.

Given a unit vector \( z^{(0)} \) and input \( x \), we recursively define \( z^{(L)} = J(x)z^{(0)} \):

\[
\tilde{z}^{(L)} \equiv \frac{1}{\sqrt{n^{(L-1)}}} W^{(L-1)} z^{(L-1)}, \quad z^{(t)} = D^{(t)} \tilde{z}^{(t)}.
\]

The following proposition provides the distribution of \( z^{(t)} \).

**Proposition 2.** For a fixed unit vector \( z^{(0)} \), fixed input data \( x \) and a network of depth \( L \) at random initialization, with a Lipschitz nonlinearity \( \sigma \), and in the limit \( n_1, \ldots, n_{L-1} \to \infty \), \( J(x)z^{(0)} \) has the following recursion with \( z_i^{(t)} \) is an attractor of a differentiable map.

\[
\dot{z}^{(1)} = \sigma'(a) b \quad (a, b) \sim N(0, \frac{\|x\|^2}{n_0}, \frac{\|x\|^2}{n_0}),
\]

\[
\dot{z}^{(t+1)} = \sigma'(a) b \quad (a, b) \sim N(0, \frac{\|x\|^2}{n_0}),
\]

\[
\tilde{z}^{(L)} \equiv \frac{1}{\sqrt{n^{(L-1)}}} W^{(L-1)} z^{(L-1)}, \quad z^{(t)} = D^{(t)} \tilde{z}^{(t)},
\]

where

\[
\hat{a}^{(1)} = \sigma(a) \quad a \sim N(0, \|x\|^2),
\]

\[
\hat{a}^{(t+1)} = \sigma(a) \quad a \sim N(0, \|x\|^2).
\]

**Proof.** See Appendix A.

Using Proposition 2, we apply the \( \varepsilon \)-net argument (Tao, 2012) (Appendix A) to obtain a bound on the Jacobian operator norm.

**Theorem 1.** For any data point \( x_i \), \( i \in [1, \ldots, n] \), with probability at least \( 1 - O(n) e^{-O(n_0)} \),

\[
\|J(x_i)\|_{op} \leq c\sqrt{n_0} \tau
\]

where \( c \) is a constant and

\[
\tau = \sup_{x_i \in \mathcal{X}, \|z^{(0)}\|_2 = 1} \mathbb{E}[(\tilde{z}^{(L-1)})^2 | z^{(0)}, x_i]
\]

**Proof.** See Appendix A.

This bound on singular values of \( J(x) \) provides a way to understand how the Jacobian at network initialization changes with respect to activation functions and the number of layers. Specifically, for sigmoid activation function, we know that \( \sigma(x) \in (0, \frac{1}{2}] \) and for any arbitrary unit vector \( z^{(0)} \),

\[
\mathbb{E}[(\tilde{z}^{(L-1)})^2 | z^{(0)}, x] \leq \mathbb{E}[\sigma'(a)^2] \leq \frac{\mathbb{E}[(\tilde{z}^{(L-2)})^2 | z^{(0)}, x]}{16}.
\]

Thus,

\[
\mathbb{E}[(\tilde{z}^{(L-1)})^2 | z^{(0)}, x] \leq \frac{1}{n_0 16^{L-1}} \|z^{(0)}\|_2^2 = \frac{1}{n_0 16^{L-1}}.
\]
and with probability at least \(1 - O(n)e^{-O(n_0)}\),
\[
\|\mathbf{J}(\mathbf{x}_i)\|_{op} \leq \frac{c}{4^{L-1}} \quad \forall i \in [n].
\] (4)

Therefore, with high probability, the initial Jacobian norm decreases with increasing layers. Based on this result, we choose to study a 2-layer sigmoid network because it would give an upper bound on the initial Jacobian norm.

Next, we argue that the largest initial Jacobian norm for a two-layer sigmoid network is concentrated around 1/2. For a given training point \(\hat{x}\),
\[
\mathbf{J}(\hat{x}) = \frac{1}{\sqrt{n_1}} \mathbf{W}^{(1)} \mathbf{D}^{(1)} \frac{1}{\sqrt{n_0}} \mathbf{W}^{(0)}.
\]

From Proposition 2, we can reach maximum \(E[\hat{z}^{L-1} | z^{(0)}, \hat{x}]\) for any fixed \(z^{(0)}\) when \(\hat{x} = 0\) because every gradient of the hidden layer is at max \(\frac{1}{2}\) and the covariance between \(x\) and \(z^{(0)}\) is zero, which would formally suggest that the upper bound of initial Jacobian norm is likely achieved at \(x = 0\) (Another way to view this comes from Proposition 2 where for any given unit vector \(z^{(0)}\), the norm of output vector \(z_{b}^{(2)}\) has a higher mean when \(x = 0\)). This gives us the simplification:
\[
\mathbf{J}(\hat{x}) = \frac{1}{4^{\sqrt{n_1}}} \mathbf{W}^{(1)} \mathbf{W}^{(0)}.
\]

Observe that \(\mathbf{W} = \frac{1}{\sqrt{n_1}} \mathbf{W}^{(1)} \mathbf{W}^{(0)}\) is a Gaussian random matrix with \(n_1 \to \infty\) where each entry is i.i.d. and the largest singular value of \(\frac{1}{\sqrt{n_1}} \mathbf{W}\) is concentrated at 2 for large \(n_0\) (Vershynin, 2012). Consequently, the largest initial Jacobian norm for a sigmoid network is concentrated around 1/2.

### 4.2. Training a Multilayer Network with a Single Example

In this section, we consider the special case when there is only one training example, \(x_1\). We show that under certain conditions the Jacobian stays close to initialization, and, combined with the result from the previous section, the trained network can form attractors.

We start by analyzing the NTK solution. Note that in this case (2) can be simplified to:
\[
f^{(\infty)}(x) = \frac{\Theta^{(L)}_\infty(x, x_1)}{\Theta^{(L)}_\infty(x_1, x_1)} (x_1 - f_0(x_1)) + f_0(x).
\]

As we describe below, \(\Theta^{(L)}_\infty\) tends to a constant kernel as \(L \to \infty\) for some network initialization (Hayou et al., 2019) and therefore the trained Jacobian equals the initial one.

To see this, first, we pay close attention to the covariance matrix \(\Sigma^{(L)}\), which is the building block of \(\Theta^{(L)}_\infty\). We define \(q_{ab}^{(L)} = \Sigma^{(L)}(x_a, x_b)\) and \(c_{ab}^{(L)} = q_{ab}^{(L)} / \sqrt{q_{aa}^{(L)} q_{bb}^{(L)}}\). It can be shown that \(c^* = 1\) is a fixed point of \(c^{(L)}_{ab}\) as \(L \to \infty\) (Poole et al., 2016). For sigmoidal networks, the stability of \(c^*\) is governed by
\[
\chi_1 \equiv \frac{\partial c^{(L)}_{ab} / \partial c^{(L)}_{ab}}{c^{(L)}_{ab}} = E[\sigma(\sqrt{q} z)^2], \quad z \sim \mathcal{N}(0, 1).
\]

where \(q^*\) is what \(q^{(L)}_{ab}\) converges to. If \(\chi_1 < 1\), \(c^*\) is a stable fixed point, suggesting that all points become equally similar as they progress through layers. A network under such initialization is said to be in the ordered region (Schoenholz et al., 2016). For such networks, \(\Theta^{(L)}_\infty\) converges to a constant kernel with increasing layers (Xiao et al., 2019; Hayou et al., 2019). As \(L \to \infty\), \(\partial \Theta^{(L)}_\infty / \partial x_1 | x = x_1 \to 0\) and,
\[
\mathbf{J}^{(\infty)}(x_1) = \lim_{L \to \infty} \mathbf{J}_0(x_1),
\]

as long as \(\Theta^{L}_\infty(x_1, x_1)\) does not converge to 0 as \(L \to \infty\). We note that this argument applies to other weight and bias variance scaling factors at initialization than the special case (1 and 0 respectively) we focus on.

Specializing to sigmoid networks, we first observe that they are in the ordered region because \(\sigma(x) \in [0, \frac{1}{2}]\), giving an upper bound on \(\chi_1 \leq 1/16\). Further the lower bound of \(\Theta^{(L)}_\infty(x_1, x_1)\) is 1/4 (Lemma 5 in Appendix B). Therefore, we expect the Jacobian to be constant during training in the large depth limit. In practice, the trained Jacobian stays close to initialization for 2 - 3 layer sigmoid networks as shown in Section 5.2.

Our analysis can explain the empirical results of (Zhang et al., 2019) that fully connected networks trained with single example tend to learn constant functions, leading to memorization. To see this, remember that for a sigmoid network the norm of the initial Jacobian falls with increasing layer (Hayou et al., 2016). For sigmoidal networks, the stability of \(c^*\) behaves like a linear network resulting in a Jacobian with eigenvalue 1, and thus may not form associative memory.
We start examining the Jacobian with the easiest case, and we will return to the discussion of variance for the Gaussian process and the expectation is linearly approximated. This justifies a linear approximation to activation function.

Lemma 3. Suppose there is a 2-layer network with activation function \( \sigma(x) = \alpha x \) and given initial weights \( W^{(1)} \in \mathbb{R}^{n_0 \times n_1}, W^{(0)} \in \mathbb{R}^{n_1 \times n_0} \) and every data point has the same norm \( r \). If the data matrix is full rank with \( n \leq n_0 \), then, at the NTK limit \( n_1 \rightarrow \infty \), \( J_\infty (x) \) has eigenvalues 1 with multiplicity at least \( n \). If at the NTK limit, \( \alpha \) is chosen such that \( \|J_0(x)\|_{op} < 1 \), then the multiplicity is exactly \( n \) and 1 is the largest eigenvalue norm.

Proof. See Appendix C.1.

Remark 1. Lemma 3 suggests that a network trained with a single example at convergence can have Jacobian eigenvalue 1 for \( \sigma(x) = \alpha x \) regardless of initial Jacobians. This result may seem to contradict Section 4.2. However, in this case, the argument in Section 4.2 is violated because the diagonal value of \( \Theta_\infty^{(2)} \) is also converging to zero as when \( \alpha < 1 \), the linear activation has a shrinking effect on the layer outputs.

The result can be naturally extended to \( \sigma(x) = \alpha x + \beta \) with \( \beta > 0 \).

Lemma 4. Suppose there is a 2-layer network with activation function \( \sigma(x) = \alpha x + \beta \), given initial weights \( W^{(1)} \in \mathbb{R}^{n_0 \times n_1}, W^{(0)} \in \mathbb{R}^{n_1 \times n_0} \) and every data point has the same norm \( r \) (i.e., \( \forall i \in [n], \|x_i\|_2 = r \)). If the data matrix is full rank with \( n \leq n_0 \), then, at the NTK limit \( n_1 \rightarrow \infty \), \( J_\infty (x) \) has eigenvalues 1 with multiplicity at least \( n - 1 \). If at the NTK limit, \( \alpha \) and \( \beta \) are chosen such that

\[
\|J_0(x)\|_{op} = 1 - \Delta, \quad \left\| \frac{1}{\sqrt{n_1}} W^{(1)} 1_{n_1} \right\|_2 < \frac{\beta n_0 \Delta}{2r \alpha^2},
\]

where \( 0 < \Delta \leq 1 \), then the multiplicity is exactly \( n - 1 \) and 1 is the largest eigenvalue norm.

Proof. See Appendix C.1.

Collectively we proved some sufficiency conditions for the largest trained Jacobian eigenvalue to be 1. We emphasize that if the largest eigenvalue is 1, and not strictly below 1, the network can fail to form attractors.

Now, we return to sigmoid networks and discuss the implications of Lemma 4. In the linear region, we have \( \alpha = \frac{1}{2} \) and \( \beta = \frac{1}{2} \). As mentioned in Section 4.1, in the NTK limit, \( \|J_0(x)\|_2 \approx \frac{1}{2} \), implying \( \Delta \approx \frac{1}{2} \). On the other hand, \( \|1_{n_1} W^{(1)} 1_{n_1} \|_2 \) is the norm of a standard Gaussian vector which follows the chi distribution, and it is concentrated around \( \sqrt{n_0} \). Then, the conditions in Lemma 4 hold with high probability if \( r < 2 \sqrt{n_0} \). Attractor formation can fail in sigmoid networks for small input norms, which we observe in simulations (Section 5).

4.4. Training a 2-Layer Network with Multiple Examples: Beyond Linear Region and a Transition to Attractor Formation

Section 4.3 suggests that small \( r \) leads to linear behavior and networks may fail to form associative memory. In this case,
section, we explore larger norm inputs where a network in the NTK regime utilizes the non-linear region of the sigmoid. We will argue that in the large norm limit all Jacobian eigenvalues’ norms are below 1. Taking into account our result that attractor formation may fail for small values of \( r \), we identify a transition with increasing \( r \) from a regime where memory formation does not occur to a regime where it occurs. Our results can be adapted for other sigmoidal functions.

For simplicity, we further assume that there are no parallel inputs in the training data (no \( x \) and \(-x\) at the same time). This is not a hard constraint and an analysis with parallel inputs is given in Appendix D.3.

Our strategy is to calculate the Jacobian in the large norm limit using Equation (3). We first note for large \( r \), \( X_{ij} \gg f_0(X)_{ij} \) because all the hidden units in the network are between 0 and 1. Therefore Equation 3 can be approximated by

\[
J_\infty(x) \approx \hat{X} \hat{K}^{-1} \frac{\partial k_x}{\partial x} + J_0(x) \tag{5}
\]

The items of interest here are \( \hat{K} \) and \( \frac{\partial k_x}{\partial x} \), for which we need to calculate the NTK and its gradient. We first define an approximation of the NTK and then use it to estimate the Jacobian.

**Approximation of the NTK:** Approximating sigmoid function \( \sigma(x) \) by \( \text{erf} \) function there \( \sigma(x) \approx \frac{1}{2} \text{erf} \left( \frac{x}{\sqrt{2}} \right) + \frac{1}{2} \) (shown in Figure 1b) allows us to use a known closed form solution for NTK (Lee et al., 2019; Williams, 1997). With this approximation, 2-layer NTK can be written as follows (full derivation can be found in Appendix D.1):

\[
\Theta^{(2)}_\infty(\hat{x}, x) \\
= \frac{1}{2\pi} \sqrt{\frac{2n_0 + \hat{x}^T \hat{x}}{(2n_0 + \hat{x}^T \hat{x} + n_0)^2}} + \frac{1}{2\pi} \arcsin \left( \frac{\hat{x}^T \hat{x}}{\sqrt{(\hat{x}^T \hat{x} + n_0)(\hat{x}^T \hat{x} + 2n_0)}} \right) + \frac{1}{4} \tag{6}
\]

In order to calculate the gradient of NTK, we define:

\[
\mathcal{T}(\Sigma, \sigma_1, \sigma_2)(x_1, x_2) \equiv \mathbb{E}_{x \sim \mathcal{N}(\Sigma)}[\sigma_1(f(x_1))\sigma_2(f(x_2))]
\]

Using this definition, the NTK’s gradient for a 2-layer network with arbitrary activation function \( \sigma \) is given by,

\[
\frac{\partial \Theta^{(2)}_\infty(\hat{x}, x)}{\partial x} = \frac{1}{n_0^2} \hat{x}^T \mathcal{T}(\Sigma^1, \sigma, \sigma) \hat{x} + \frac{2}{n_0} \mathcal{T}(\Sigma^1, \sigma, \sigma) \hat{x} + \frac{1}{n_0^2} \hat{x}^T \mathcal{T}(\Sigma^1, \sigma, \sigma) x + \frac{1}{n_0} \mathcal{T}(\Sigma^1, \sigma, \sigma) x.
\]

Now we can use these expressions to calculate \( \hat{K} \) and \( \frac{\partial k_x}{\partial x} \) in Equation 5.

### Calculation of \( \hat{K} \) and \( \frac{\partial k_x}{\partial x} \):

We first look at Equation (6), using the fact that because all the data points have the same norm \( x^T x = r^2 \rho_{i,j} \) with \( |\rho_{i,j}| < 1 \) (since there are no parallel inputs):

\[
\hat{K}_{ij} = \frac{1}{4} + \frac{1}{2\pi} \arcsin \left( \frac{r^2 \rho_{i,j}}{r^2 + 2n_0} \right) + \frac{1}{2\pi} \sqrt{\frac{r^2}{(2n_0 + r^2)^2} - r^4 \rho_{i,j}^2}.
\]

To get insight into the behavior of \( \hat{K}_{ij} \), let’s consider diagonal and off-diagonal entries separately in the large-\( r \) limit. First the off-diagonals:

\[
\hat{K}_{ij} \approx \frac{1}{4} + \frac{1}{2\pi} \arcsin (\rho_{i,j}) + \frac{1}{2\pi} \sqrt{1 - \rho_{i,j}^2}.
\]

Next we look at the diagonals:

\[
\hat{K}_{ii} \approx \frac{r}{4\pi \sqrt{n_0}}.
\]

The diagonal grows linearly with \( r \) and dominates over the off-diagonals. The kernel tends to a scaled identity matrix in the \( r \rightarrow \infty \) limit. Thus,

\[
\left\| \hat{K}^{-1} \right\|_{op} \approx \frac{4\pi}{r} \sqrt{n_0}.
\]

Next, we look at \( \frac{\partial k_x}{\partial x} \). As we are interested in trained Jacobian at training examples, without loss of generality, we focus on \( J_\infty(x_1) \). We already know that \( \hat{K} \) tends to converge to a diagonal matrix with large \( r \). One could use this result to suggest that \( \frac{\partial k_x}{\partial x} \big|_{x_1} \approx 0 \) if \( i \neq j \). In fact, with large \( r \), it can be shown that (Appendix D.2):

\[
\left\| \frac{\partial k_x}{\partial x} \right\|_{op} \approx \frac{1}{8\pi \sqrt{n_0}} \tag{7}
\]

and

\[
\left\| \frac{\partial k_x}{\partial x} \right\|_{op} \approx \frac{1}{8\pi \sqrt{n_0}} \tag{8}
\]

### Trained Jacobian and attractor formation:

Finally, we can use our results in Equations (7) and (9) to calculate the trained Jacobian using Equation (5) with large \( r \):

\[
\left\| J_\infty(x) \right\|_{op} \approx \left\| \hat{X} \hat{K}^{-1} \frac{\partial k_x}{\partial x} + J_0(x) \right\|_{op}
\]

\[
\leq \left\| \hat{X} \hat{K}^{-1} \frac{\partial k_x}{\partial x} \right\|_{op} + \left\| J_0(x) \right\|_{op}
\]

\[
\approx r \frac{4\pi \sqrt{n_0}}{r} \frac{1}{8\pi \sqrt{n_0}} + \left\| J_0(x) \right\|_{op}
\]

\[
= \frac{1}{2} + \left\| J_0(x) \right\|_{op}.
\]
To go from the second to the third row of the equation above, we used the fact that each column of $X$ has norm $r$, and in the limit $K^{-1}$ is a scaled identity matrix and $\frac{\partial x}{\partial x}$ has only one non-zero row and that row is a scaled $x_1$ (Appendix D.2).

To see the implications of this result, we observe that as $r \to \infty$, $J_0(x) \approx 0$ because the pre-activation of the two layer network will be infinitely big and the units saturate leading to zero gradients in the Jacobian matrix, we can conclude in this limit

$$\|J_\infty(x)\|_{op} \leq 1/2.$$

Combined with our previous result that attractor formation may fail for small values of $r$, this suggests that there is a transition from a region where associative memory formation fails to a region where it succeeds with increasing $r$. We show this transition in simulations in Section 5 and verify that the largest eigenvalue of the Jacobian falls towards $1/2$ asymptotically.

4.5. Beyond NTK

In practice, large values of $r$ often require a larger width to stay in NTK. If the large width condition is violated, trained weights will not stay close to initialization. However, in this case, a significant deviation of weights from initialization caused by gradient descent may saturate hidden units, leading to zero gradients in the Jacobian matrix. As shown in Figure 3, for fixed hidden size, larger input radius leads to near zero eigenvalues when the NTK conditions no longer hold. Therefore, the network can behave as if only the last layer is trained with saturated hidden features and close to zero Jacobian norm since all the $D^{(t)}$ matrices have mostly zero diagonal entries. Note that, in overparameterized networks, this type of optimization can often result in zero training loss as well due to the over-parameterization of the last layer.

5. Simulations

5.1. Experiment Setup

Training and iterative convergence criteria: Training is stopped when the training loss of the auto-encoder drops below a threshold, which we chose to be $10^{-7}$. Iterative convergence happens when a non-fixed point converges to a fixed point measured by mean-squared-error after passing through the trained autoencoder iteratively. This threshold was $10^{-2}$.

Implementation details: We used vanilla gradient descent with learning rate $1$, similar to (Jacot et al., 2018). The code is implemented with Pytorch. For each setting, we ran experiments 100 times to get 100 sets of samples. For all experiments except experiments on MNIST, the samples are randomly generated.

5.2. Single Training Example

We first present experiments for a single training example. The sigmoid network is chosen to have hidden dimension 1000 with input dimension 32. Figure 2 shows that the difference between the largest eigenvalue norms at initialization and after training decreases with more number of layers.

5.3. Multiple Training Examples

Linear Region: In this section, we first illustrate the eigenvalue distribution in the linear region by sampling unit vectors as data ($r = 1$). Here, we trained 2 layer sigmoid networks with input dimension 10 and hidden size 1000 for 2, 5 and 8 training points. As suggested by Lemma 4, there should be $n - 1$ eigenvalues with norm around 1. This is supported by Figure E.1 in the Appendix, where 10%, 40% and 70% of the eigenvalues are near 1. Here, the presence of eigenvalue 1 indicates that the network operates in the linear region.

Beyond Linear Region: We demonstrate how the largest eigenvalue norm varies with the input radius. We test with various hidden dimensions to achieve the NTK limit on input dimension 32 with the number of training data 5, 20, and 40. Only experiments that can be trained to have loss below $10^{-7}$ or fit into single Titan V GPU are included.

The general trend, as shown in Figure 3, is that as we move away from the linear region, the largest eigenvalue norm will drop, and as we keep increasing the hidden layer size, it will move close to the 1/2 limit as suggested by our analysis. It is also worth noting that the input radius needs to be increased with large number of training examples to get training loss below $10^{-7}$ under reasonable iterations, implying the capacity of the networks is controlled by input radius and agreeing with the intuition and theoretical results from (Allen-Zhu et al., 2018), that there needs to be some level of separation between data points to train the network.
5.4. Basin of Attraction

We test basin of attraction by adding Gaussian noise to training examples and check if the modified examples can converge to the original ones via iterative maps under 50 iterations. And the convergence rate is the number of samples that could be successfully recovered. The standard deviation of the Gaussian noise is called the noise radius. The network has two layers with hidden size 10000 and input dimension 32. Not surprisingly, Figure 4a shows that larger input norm gives greater basin of attraction. More experiments can be found in Appendix E.2.

5.5. MNIST Data

We also test basin of attraction experiments on MNIST dataset to check if we can recover real training examples. The images are prepossessed by subtracting means and rescaled to have different input norms for testing. Similar to the setting before, Figure 4b also shows that larger input norm gives greater basin of attraction. Notice that because MNIST images have large input dimension, they need larger radius to move out of the linear region. More experiments can be found in Appendix E.3.

5.6. Sigmoidal Activation

Finally, we show that our results can be extended to different sigmoidal activation functions as well. We chose two-layer network with hidden size 10000 and input dimension 32 and 20 training examples. As before, only settings that led to a training loss below $10^{-7}$ are included. Figure 5 clearly shows that all the activation functions share similar curves. Notice that both tanh and erf have large eigenvalue when $r$ is small. This is not a contradiction to our Lemma 3 as their $\alpha = \dot{\sigma}(0)$ is too large to satisfy the conditions in Lemma 3. To further verify this result, we also include how the histogram of eigenvalue norm changes for those activations in the Appendix E.4.

6. Conclusion

In this paper, we theoretically and empirically show that training overparameterized sigmoid autoencoders can lead to attractors for a single training example and multiple training examples in the non-linear region, with the help of theories developed in the NTK limit. We identified a behavior change governed by the input radius. Some future directions include generalizing our results to other activations, identifying other factors that can determine whether autoencoders can learn to have attractors or not, and how the formations of attractors are related to generalization in deep learning.

Acknowledgements

C. Pehlevan thanks the Harvard Data Science Initiative, Google and Intel for support.
References


