A Theory of Non-Linear Feature Learning with One Gradient Step in Two-Layer Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Feature learning is thought to be one of the fundamental reasons for the success of deep neural networks. It is rigorously known that in two-layer fully-connected neural networks under certain conditions, one step of gradient descent on the first layer followed by ridge regression on the second layer can lead to feature learning; characterized by the appearance of a separated rank-one component—spike—in the spectrum of the feature matrix. However, with a constant gradient descent step size, this spike only carries information from the linear component of the target function and therefore learning non-linear components is impossible. We show that with a learning rate that grows with the sample size, such training in fact introduces multiple rank-one components, each corresponding to a specific polynomial feature. We further prove that the limiting large-dimensional and large sample training and test errors of the updated neural networks are fully characterized by these spikes. By precisely analyzing the improvement in the loss, we demonstrate that these non-linear features can enhance learning.

1 Introduction

Learning non-linear features—or representations—from data is thought to be one of the fundamental reasons for the success of deep neural networks (e.g., Bengio et al., 2013; Donahue et al., 2016; Yang & Hu, 2021; Shi et al., 2022; Radhakrishnan et al., 2022, etc.). At the same time, the current theoretical understanding of feature learning is incomplete. In particular, among many theoretical approaches to study neural nets, much work has focused on two-layer fully-connected neural networks with a randomly generated, untrained first layer and a trained second layer—or *random features models* (Rahimi & Recht, 2007). Despite their simplicity, random features models can capture various empirical properties of deep neural networks. Nevertheless, feature learning is absent in random features models, because the first layer weights are assumed to be randomly generated, and then fixed. Thus, random features models fall short of providing a comprehensive explanation for the success of deep learning. While other models such as the neural tangent kernel (Jacot et al., 2018; Du et al., 2019) can be more expressive, they also lack feature learning.

To bridge the gap between random features models and feature learning, several recent approaches have shown provable feature learning for neural nets under certain conditions. In particular, the recent pioneering work of Ba et al. (2022) analyzed two-layer neural networks, trained with one gradient step on the first layer. They showed that when the step size is small, after one gradient step, the resulting two-layer neural network can learn linear features. However, it still behaves as a noisy linear model and does not capture non-linear components of a teacher function. Moreover, they showed that for a sufficiently large step size, under certain conditions, the one-step updated random features model can outperform linear and kernel predictors. However, the effects of a large gradient step size on the features is unknown. What happens in the intermediate step size regime also remains unexplored. In this paper, we focus on the following key questions in this area: What nonlinear features are learned by a two-layer neural network after one gradient update? How are these features reflected in the singular values and vectors of the feature matrix, and how does this depend on the scaling of the step size? What exactly is the improvement in the loss due to the nonlinear features learned?

2 Preliminaries

In this paper, we study a supervised learning problem with training data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, for $i \in [2n]$, where d is the feature dimension and $n \geq 2$ is the sample size. We assume that the data is generated according to

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \mathbf{I}_d), \text{ and } y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \tag{1}$$

in which $f_\star$ is the ground truth or *teacher function*, and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2)$ is additive noise.

We fit a model to the data in order to predict outcomes for unlabeled examples at test time; using a two-layer neural network. We let the width of the internal layer be $N \in \mathbb{N}$. For a weight matrix $\mathbf{W} \in \mathbb{R}^{N \times d}$, an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ applied element-wise, and the weights $\boldsymbol{a} \in \mathbb{R}^N$ of a linear layer, we define the two-layer neural network as $f_{\mathbf{W},\boldsymbol{a}}(\boldsymbol{x}) = \boldsymbol{a}^\top \sigma(\mathbf{W}\boldsymbol{x})$.

Following Ba et al. (2022), for the convenience of the theoretical analysis, we split the training data into two parts: $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}, \boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\tilde{\mathbf{X}} = [\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{2n}]^\top \in \mathbb{R}^{n \times d}, \tilde{\boldsymbol{y}} = (y_{n+1}, \ldots, y_{2n})^\top \in \mathbb{R}^n$. We train the two layer neural network as follows. First, we initialize $\boldsymbol{a} = (a_1, \ldots, a_N)^\top$ with $a_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, 1/N)$ and initialize $\mathbf{W}$ with $\mathbf{W}_0 = [\boldsymbol{w}_{0,1}, \ldots, \boldsymbol{w}_{0,N}]^\top \in \mathbb{R}^{N \times d}, \quad \boldsymbol{w}_{0,i} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$,where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ and $\mathrm{Unif}(\mathbb{S}^{d-1})$ is the uniform measure over it. Fixing $\boldsymbol{a}$ at initialization, we perform *one step of gradient descent* on $\mathbf{W}$ with respect to the squared loss computed on $(\mathbf{X}, \boldsymbol{y})$. Recalling that $\circ$ denotes element-wise multiplication, the negative gradient can be written as

$$G := -\frac{\partial}{\partial \mathbf{W}} \left[\frac{1}{2n} \left\| \boldsymbol{y} - \sigma(\mathbf{X}\mathbf{W}^\top)\boldsymbol{a} \right\|_2^2 \right]_{\mathbf{W}=\mathbf{W}_0} = \frac{1}{n} \left[(\boldsymbol{a}\boldsymbol{y}^\top - \boldsymbol{a}\boldsymbol{a}^\top \sigma(\mathbf{W}_0\mathbf{X}^\top)) \circ \sigma'(\mathbf{W}_0\mathbf{X}^\top) \right] \mathbf{X},$$

and the one-step update is $\mathbf{W} = \mathbf{W}_0 + \eta\, \mathbf{G}$ for a *learning rate* or *step size* η.

After the update on $\mathbf{W}$, we perform ridge regression on $\boldsymbol{a}$ using $(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}})$. Let $\mathbf{F} = \sigma(\tilde{\mathbf{X}}\mathbf{W}^\top) \in \mathbb{R}^{n \times N}$ be the feature matrix after the one-step update. For a regularization parameter $\lambda > 0$, we set

$$\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}(\mathbf{F}) = \arg\min_{\boldsymbol{a} \in \mathbb{R}^N} \frac{1}{n} \|\tilde{\boldsymbol{y}} - \mathbf{F}\boldsymbol{a}\|_2^2 + \lambda \|\boldsymbol{a}\|_2^2 = \left(\mathbf{F}^\top\mathbf{F} + \lambda n \mathbf{I}_N\right)^{-1} \mathbf{F}^\top \tilde{\boldsymbol{y}}. \tag{2}$$

Then, for a test datapoint with features $\boldsymbol{x}$, we predict the outcome $\hat{y} = f_{\mathbf{W},\hat{\boldsymbol{a}}}(\boldsymbol{x}) = \hat{\boldsymbol{a}}^\top \sigma(\mathbf{W}\boldsymbol{x})$.

2.1 Conditions

Our theoretical analysis applies under the following conditions:

Condition 2.1 (Asymptotic setting) *We assume that the sample size n, dimension d, and width of hidden layer N all tend to infinity with $d/n \to \phi > 0$ and $d/N \to \psi > 0$.*

Condition 2.2 *We let $f_\star : \mathbb{R}^d \to \mathbb{R}$ be a single-neuron model $f_\star(\boldsymbol{x}) = \sigma_\star(\boldsymbol{x}^\top \boldsymbol{\beta}_\star)$, where $\boldsymbol{\beta}_\star \in \mathbb{R}^d$ is an unknown parameter with $\boldsymbol{\beta}_\star \sim \mathsf{N}(0, \frac{1}{d}\mathbf{I}_d)$ and $\sigma_\star : \mathbb{R} \to \mathbb{R}$ is a* teacher activation *function. We further assume that $\sigma_\star : \mathbb{R} \to \mathbb{R}$ is $\Theta(1)$-Lipschitz.*

We let H_k, $k \geq 1$ be the (probabilist's) Hermite polynomials on $\mathbb{R}$.

Condition 2.3 *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ has the following Hermite expansion in L^2: $\sigma(z) = \sum_{k=1}^\infty c_k H_k(z), \quad c_k = \frac{1}{k!}\mathbb{E}_{Z \sim \mathsf{N}(0,1)}[\sigma(Z)H_k(Z)]$, where $c_1 \neq 0$. Moreover, the first three derivatives of σ exist and are bounded.*

Condition 2.4 *The teacher activation $\sigma_\star : \mathbb{R} \to \mathbb{R}$ has the following Hermite expansion in L^2: $\sigma_\star(z) = \sum_{k=0}^M c_{\star,k} H_k(z), \quad c_{\star,k} = \frac{1}{k!}\mathbb{E}_{Z \sim \mathsf{N}(0,1)}[\sigma_\star(Z)H_k(Z)]$ for some $M \in \mathbb{N}$. Also, we define $c_\star = (\sum_{k=0}^M k! c_{\star,k}^2)^{\frac{1}{2}}$.*

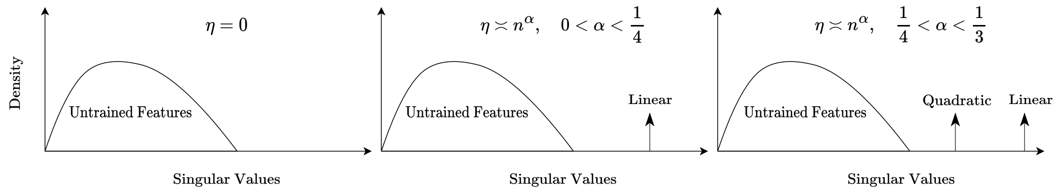

Figure 1: Spectrum of the updated feature matrix for different regimes of the gradient step size η. Spikes corresponding to monomial features are added to the spectrum of the initial matrix. The number of spikes depends on the range α. See Theorems 3.2 and 3.3 for more details.

3 Analysis of the Feature Matrix

As the following proposition suggests, $\boldsymbol{\beta} = \frac{1}{n}\mathbf{X}^\top \boldsymbol{y}$ can be viewed as a noisy estimate of $\boldsymbol{\beta}_\star$.

Proposition 3.1 *If Conditions 2.1-2.4 hold, then* $\frac{|\boldsymbol{\beta}_\star^\top \boldsymbol{\beta}|}{\|\boldsymbol{\beta}_\star\|_2 \|\boldsymbol{\beta}\|_2} \to_P \frac{|c_{\star,1}|}{\sqrt{c_{\star,1}^2 + \phi(c_\star^2 + \sigma_\varepsilon^2)}}$.

Next, we will show that after the gradient step, the spectrum of the feature matrix $\mathbf{F}$ will consist of a bulk of singular values that stick close together—given by the spectrum of the initial feature matrix $\mathbf{F}_0 = \sigma(\tilde{\mathbf{X}}\mathbf{W}_0^\top)$—and ℓ separated spikes[1], where ℓ is an integer that depends on the step size used in the gradient update. Specifically, when the step size is $\eta \asymp n^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$ for some $\ell \in \mathbb{N}$, the feature matrix $\mathbf{F}$ can be approximated in operator norm by the untrained features $\mathbf{F}_0 = \sigma(\tilde{\mathbf{X}}\mathbf{W}_0^\top)$ plus ℓ rank-one terms, where the left singular vectors of the rank-one terms are aligned with the non-linear features $\tilde{\mathbf{X}} \mapsto (\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k}$, for $k \in [\ell]$. Recall that the shifted ReLU activation $\sigma : \mathbb{R} \to \mathbb{R}$ is defined for all $x \in \mathbb{R}$ by $\sigma(x) = \max(x, 0) - \frac{1}{\sqrt{2\pi}}$.

Theorem 3.2 (Spectrum of feature matrix) *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a polynomial or the shifted ReLU activation. Let $\eta \asymp n^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$ for some $\ell \in \mathbb{N}$. If Conditions 2.1-2.4 hold, then for c_k from Condition 2.3 and $\mathbf{F}_0 = \sigma(\tilde{\mathbf{X}}\mathbf{W}_0^\top)$,*

$$\mathbf{F} = \mathbf{F}_\ell + \boldsymbol{\Delta}, \qquad \text{with} \qquad \mathbf{F}_\ell := \mathbf{F}_0 + \sum_{k=1}^{\ell} c_1^k c_k \eta^k (\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k} (\boldsymbol{a}^{\circ k})^\top, \tag{3}$$

where $\|\boldsymbol{\Delta}\|_{\text{op}} = o(\sqrt{n})$ with probability $1 - o(1)$.

To understand $(\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k}(\boldsymbol{a}^{\circ k})^\top$, notice that for a datapoint with features $\tilde{\boldsymbol{x}}_i$, the activation of each neuron is proportional to the polynomial feature $(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^k$, with coefficients given by $\boldsymbol{a}^{\circ k}$ for the neurons. The spectrum of the initial feature matrix $\mathbf{F}_0$ is fully characterized in Pennington & Worah (2017); Benigni & Péché (2021, 2022), and its operator norm is known to be $\Theta_{\mathbb{P}}(\sqrt{n})$. Moreover, it follows from the proof that the operator norm of each of the terms $c_1^k c_k \eta^k (\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k}(\boldsymbol{a}^{\circ k})^\top$, $k \in [l]$ is with high probability of order larger than $\sqrt{n}$. Thus, Theorem 3.2 identifies the spikes in the spectrum of the feature matrix.

In the following theorem, we argue that the subspace spanned by the non-linear features $\{\sigma(\tilde{\mathbf{X}}\boldsymbol{w}_i)\}_{i \in [N]}$ can be approximated by the subspace spanned by the monomials $\{(\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k}\}_{k \in [\ell]}$.

Theorem 3.3 *Let $\mathcal{F}_\ell$ be the ℓ-dimensional subspace of $\mathbb{R}^n$ spanned by top-ℓ left singular vectors (principal components) of $\mathbf{F}$. Under the conditions of Theorem 3.2, we have $d(\mathcal{F}_\ell, \operatorname{span}\{(\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k}\}_{k \in [\ell]}) \to_P 0$, where d is the principal angular distance.*

This result shows that after one step of gradient descent with step size $\eta \asymp n^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$, the subspace of the top-ℓ left singular vectors carries information from the polynomials

[1]Using terminology from random matrix theory (Bai & Silverstein, 2010; Yao et al., 2015).

$\{(\tilde{\mathbf{X}}\boldsymbol{\beta})^{\circ k}\}_{k\in[\ell]}$. Also, recall that by Proposition 3.1, the vector $\boldsymbol{\beta}$ is aligned with $\boldsymbol{\beta}_\star$. Hence, it is shown that $\mathcal{F}_\ell$ carries information from the first ℓ polynomial components of the teacher function.

4 Learning Higher-Degree Polynomials

4.1 Equivalence Theorems

Given a regularization parameter $\lambda > 0$, recalling the ridge estimator $\hat{\boldsymbol{a}}(\mathbf{F})$ from equation 2, we define the training loss $\mathcal{L}_{\text{tr}}(\mathbf{F}) = \frac{1}{n}\|\tilde{\boldsymbol{y}} - \mathbf{F}\hat{\boldsymbol{a}}(\mathbf{F})\|_2^2 + \lambda\|\hat{\boldsymbol{a}}(\mathbf{F})\|_2^2$. In the next theorem, we show that when $\eta \asymp n^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$, the training loss $\mathcal{L}_{\text{tr}}(\mathbf{F})$ can be approximated with negligible error by $\mathcal{L}_{\text{tr}}(\mathbf{F}_\ell)$. In other words, the approximation of the feature matrix in Theorem 3.2 can be used to derive the asymptotics of the training loss.

Theorem 4.1 (Training loss equivalence) *Let $\eta \asymp n^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$ for some $\ell \in \mathbb{N}$ and recall $\mathbf{F}_\ell$ from equation 3. If Conditions 2.1-2.4 hold, then for any fixed $\lambda > 0$, we have $\mathcal{L}_{\text{tr}}(\mathbf{F}) - \mathcal{L}_{\text{tr}}(\mathbf{F}_\ell) = o(1)$, with probability $1 - o(1)$.*

Similar equivalence results can also be proved for the test risk, i.e., the average test loss. For any $\boldsymbol{a} \in \mathbb{R}^N$, we define the test risk of $\boldsymbol{a}$ as $\mathcal{L}_{\text{te}}(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{f},y}(y - \boldsymbol{f}^\top\boldsymbol{a})^2$, in which the expectation is taken over $(\boldsymbol{x}, y)$ where $\boldsymbol{f} = \sigma(\mathbf{W}\boldsymbol{x})$ with $\boldsymbol{x} \sim \mathsf{N}(0, \mathbf{I}_d)$ and $y = f_\star(\boldsymbol{x}) + \varepsilon$ with $\varepsilon \sim \mathsf{N}(0, \sigma_\varepsilon^2)$. The next theorem shows that one can also use the approximation of the feature matrix from Theorem 3.2 to derive the asymptotics of the test risk.

Theorem 4.2 (Test risk equivalence) *Let $\eta \asymp n^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$ for some $\ell \in \mathbb{N}$ and $\mathbf{F}_\ell$ be defined as in equation 3. If Conditions 2.1-2.4 hold, then for any $\lambda > 0$, if $\mathcal{L}_{\text{te}}(\hat{\boldsymbol{a}}(\mathbf{F})) \to_P \mathcal{L}_{\mathbf{F}}$ and $\mathcal{L}_{\text{te}}(\hat{\boldsymbol{a}}(\mathbf{F}_\ell)) \to_P \mathcal{L}_{\mathbf{F}_\ell}$, we have $\mathcal{L}_{\mathbf{F}} = \mathcal{L}_{\mathbf{F}_\ell}$.*

4.2 Analysis of Training Loss

The following results depend on the limits of traces of the matrices $(\mathbf{F}_0\mathbf{F}_0^\top + \lambda n\mathbf{I}_n)^{-1}$ and $\tilde{\mathbf{X}}^\top(\mathbf{F}_0\mathbf{F}_0^\top + \lambda n\mathbf{I}_n)^{-1}\tilde{\mathbf{X}}$. These limits have been determined in Adlam et al. (2022); Adlam & Pennington (2020), see also Pennington & Worah (2017); Péché (2019). We leverage that $\lim_{d,n,N\to\infty}\operatorname{tr}(\tilde{\mathbf{X}}^\top(\mathbf{F}_0\mathbf{F}_0^\top + \lambda n\mathbf{I}_n)^{-1}\tilde{)}/d = \psi m_2/\phi > 0$ and $\lim_{d,n,N\to\infty}\operatorname{tr}((\mathbf{F}_0\mathbf{F}_0^\top + \lambda n\mathbf{I}_n)^{-1}) = \psi m_1/\phi > 0$.

Theorem 4.3 *If Conditions 2.1-2.4 hold, and if $\eta \asymp n^\alpha$ with $0 < \alpha < \frac{1}{4}$ so that $\ell = 1$, then for the learned feature map $\mathbf{F}$ and the untrained feature map $\mathbf{F}_0$ we have $\mathcal{L}_{\text{tr}}(\mathbf{F}_0) - \mathcal{L}_{\text{tr}}(\mathbf{F}) \to_P \Delta_1$, where*

$$\Delta_1 = \frac{\psi\lambda c_{\star,1}^4 m_2}{\phi[c_{\star,1}^2 + \phi(c_\star^2 + \sigma_\varepsilon^2)]} > 0. \tag{4}$$

The above theorem confirms our intuition that training the first-layer parameters improves the performance of the trained model. From this theorem, it can be seen that when $\ell = 1$, the improvement in the loss is increasing in the strength of the linear component $c_{\star,1}$ keeping the signal strength $c_\star$ fixed; and not so for the strength of the non-linear component $c_{\star,>1}^2 = c_\star^2 - c_{\star,1}^2$. Our next theorem shows that when we further increase the step size to the $\ell = 2$ regime, the loss of the trained model will drop by an additional positive value Δ_2 depending on the strength $c_{\star,2}$ of the quadratic signal, which supports our claim that the quadratic component of the target function is also being learned.

Theorem 4.4 *If Conditions 2.1-2.4 hold, while we also have $c_2 \neq 0$, and $\eta \asymp n^\alpha$ with $\frac{1}{4} < \alpha < \frac{1}{3}$ so that $\ell = 2$, then for the learned feature map $\mathbf{F}$ and the untrained feature map $\mathbf{F}_0$, we have $\mathcal{L}_{\text{tr}}(\mathbf{F}_0) - \mathcal{L}_{\text{tr}}(\mathbf{F}) \to_P \Delta_1 + \Delta_2$, where Δ_1 was defined in Theorem 4.3 and*

$$\Delta_2 = \frac{4\psi\lambda c_{\star,1}^4 c_{\star,2}^2 m_1}{3\phi[\phi(c_\star^2 + \sigma_\varepsilon^2) + c_{\star,1}^2]^4} > 0. \tag{5}$$

Given $\ell \in \{1, 2\}$, the loss of the trained model is asymptotically constant for all $\eta = cn^\alpha$ with $\frac{\ell-1}{2\ell} < \alpha < \frac{\ell}{2\ell+2}$ and $c \in \mathbb{R}$. There are sharp jumps at the edges between regimes of α, whose size is precisely characterized in the theorems above.

References

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, 2020.

Ben Adlam, Jake A Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022.

Zhidong Bai and Jack W Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer, 2010.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.

Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint arXiv:2201.04753*, 2022.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2016.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Sandrine Péché. A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, 2017.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2022.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, 2021.

Jianfeng Yao, Zhidong Bai, and Shurong Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.