Asymptotics of Ridge Regression in Convolutional Models

Mojtaba Sahraee-Ardakan [1,2] **Tung Mai** [3] **Anup Rao** [3] **Ryan Rossi** [3] **Sundeep Rangan** [4] **Alyson K. Fletcher** [1,2]

Abstract

Understanding generalization and estimation error of estimators for simple models such as linear and generalized linear models has attracted a lot of attention recently. This is in part due to an interesting observation made in machine learning community that highly over-parameterized neural networks achieve zero training error, and yet they are able to generalize well over the test samples. This phenomenon is captured by the so called double descent curve, where the generalization error starts decreasing again after the interpolation threshold. A series of recent works tried to explain such phenomenon for simple models. In this work, we analyze the asymptotics of estimation error in ridge estimators for convolutional linear models. These convolutional inverse problems, also known as deconvolution, naturally arise in different fields such as seismology, imaging, and acoustics among others. Our results hold for a large class of input distributions that include i.i.d. features as a special case. We derive exact formulae for estimation error of ridge estimators that hold in a certain high-dimensional regime. We show the double descent phenomenon in our experiments for convolutional models and show that our theoretical results match the experiments.

1. Introduction

Increasingly powerful hardware along with deep learning libraries that efficiently use these computational resources have allowed us to train ever larger neural networks. Modern neural networks are so over-parameterized that they can perfectly fit random noise (Zhang et al., 2016; Li et al., 2020). With enough over-parameterization, they can also achieve zero loss over training data with their parameters moving

[1]Department of Electrical and Computer Engineering, University of California, Los Angeles [2]Department of Statistics, University of California, Los Angeles [3]Adobe Research [4]Department of Electrical and Computer Engineering, New York University. Correspondence to: Mojtaba Sahraee-Ardakan <msahraee@ucla.edu>.

only slightly away from the initialization (Allen-Zhu et al., 2018; Soltanolkotabi et al., 2018; Du et al., 2018a;b; Li and Liang, 2018; Zou et al., 2020). Yet, these models generalize well on test data and are widely used in practice (Zhang et al., 2016). In fact, some recent work suggest that it is best practice to use as large a model as possible for the tasks in hand (Huang et al., 2018). This seems contrary to our classical understanding of generalization where increasing the complexity of the model space to the extent that the training data can be easily interpolated indicates poor generalization. Most statistical approaches explain generalization by controlling some notion of capacity of the hypothesis space, such as VC dimension, Rademacher complexity, or metric entropy (Anthony and Bartlett, 2009). Such approaches that do not incorporate the implicit regularization effect of the optimization algorithm fail to explain generalization of over-parameterized deep networks (Kalimeris et al., 2019; Oymak and Soltanolkotabi, 2019).

The so-called *double descent* curve where the test risk starts decreasing again by over-parameterizing neural networks beyond the interpolation threshold is widely known by now (Belkin et al., 2019a; Nakkiran et al., 2019). Interestingly, such phenomenon is not unique to neural networks and have been observed even in linear models (Le Cun et al., 1991; Dobriban et al., 2018; Advani et al., 2020). Recently, another line of work has also connected neural networks to linear models. In (Jacot et al., 2018), the authors show that infinitely wide neural networks trained by gradient descent behave like their linearization with respect to the parameters around their initialization. The problem of training such wide neural networks with square loss then turns into a kernel regression problem in a RKHS associated to a fixed kernel called the *neural tangent kernel* (NTK). The NTK results were later extended to many other architectures such as convolutional networks and recurrent neural networks (Li et al., 2019; Alemohammad et al., 2020; Yang, 2020). Trying to understand the generalization in deep networks and explaining such phenomenon as the double descent curve has attracted a lot of attention to theoretical properties of kernel methods as well as simple machine learning models. Such models, despite their simplicity, can help us gain a better understanding of machine learning models and algorithms that might be hard to achieve just by looking at deep neural networks due their complex nature.

Double descent has been shown in linear models (Dobriban et al., 2018; Belkin et al., 2019b; Hastie et al., 2019), logistic regression (Deng et al., 2019), support vector machines (Montanari et al., 2019), generalized linear models (Emami et al., 2020), kernel regression (Liang et al., 2020), random features regression (Mei and Montanari, 2019; Hastie et al., 2019), and random Fourier feature regression (Liao et al., 2020) among others. Most of these works consider the problem in a doubly asymptotic regime where both the number of parameters and the number of observations go to infinity at a fixed ratio. This is in contrast to classical statistics where either the number of parameters is assumed to be fixed and the samples go to infinity or vice versa. In practice, the number of parameters and number of samples are usually comparable and therefore the doubly asymptotic regime provides more value about the performance of different models and algorithms. In this work we study the performance of ridge estimators in convolutional linear inverse problems in this asymptotic regime. Unlike ordinary linear inverse problems, theoretical properties of the estimation problem in convolutional models has not been studied, despite their wide use in practice, e.g. in solving inverse problems with deep generative priors (Ulyanov et al., 2018).

Beyond machine learning, inverse problems involving convolutional measurement models are often called deconvolution and are encountered in many different fields. In astronomy, deconvolution is used for example to deblur, sharpen, and correct for optical aberrations in imaging (Starck et al., 2002). In seismology, deconvolution is used to separate seismic traces into a source wavelet and an impulse response that corresponds to the layered structure of the earth (Treitel and Lines, 1982; Mueller, 1985). In imaging, it is used to correct for blurs caused by the point spread function, sharpen out of focus areas in 3D microscopy (McNally et al., 1999), and to separate neuronal spikes for calcium traces in calcium imaging (Friedrich et al., 2017) among others. In practical applications, the convolution kernel might not be known and should either be estimated from the physics of the problem or jointly with the unknown signal using the data.

Summary of Contributions. We analyze the performance of ridge estimators for convolutional models in the proportional asymptotics regime. Our main result (Theorem 1) characterizes the limiting joint distribution of the true signal and its ridge estimate in terms of the spectral properties of the data. As a result of this theorem, we can provide an exact formula to compute the mean squared error of ridge estimator in the form of a scalar integral (Corollary 1). Our assumptions on the data are fairly general and include many random processes as an example as opposed to i.i.d. features only. Even though our theoretical results hold only in a certain high dimensional limit, our experiments show that

its prediction matches the observed error even for problems of moderate size. We show that our result can predict the double descent curve of the estimation error as we change the ratio of number of measurements to unknowns.

Prior Work. Asymptotic error of ridge regression for ordinary linear inverse problems (as opposed to convolutional linear inverse problem considered in this work) is studied in (Dicker et al., 2016) for isotropic features. Asymptotics of ridge regression for correlated features is studied in (Dobriban et al., 2018). Error of ridgeless (minimum ℓ_2-norm interpolant) regression for data generated from a linear or nonlinear model is obtained in (Hastie et al., 2019). These works use results from random matrix theory to derive closed form formulae for the estimation or generalization error. For features with general i.i.d. prior other than Gaussian distribution, approximate message passing (AMP) (Donoho et al., 2010; Bayati and Montanari, 2011) or vector approximate message passing (VAMP) (Rangan et al., 2019) can be used to obtain asymptotics of different types of error. Instead of a closed form formula, these works show that the asymptotic error can be obtained via a recursive equation that is called the *state evolution*. In (Deng et al., 2019), the authors use convex Gaussian min-max theorem to characterize the performance of maximum likelihood as well as SVM classifiers with i.i.d. Gaussian covariates. In (Emami et al., 2020), the problem of learning generalized linear models is reduced to an inference problem in deep networks and the results of (Pandit et al., 2020) are used to obtain the generalization error.

Notation. We use uppercase boldface letters for matrices and tensors, and lowercase boldface letters for vectors. For a matrix $\mathbf{A}$, its ith row and column is denoted by $\mathbf{A}_{i*}$ and $\mathbf{A}_{*i}$ respectively. A similar notation is used to show slices of tensors. The submatrix formed by columns i through $j-1$ of $\mathbf{A}$ is shown by $\mathbf{A}_{*,i:j}$. Standard inner product for vectors, matrices, and tensors is represented by $\langle \cdot, \cdot \rangle$. $\mathcal{N}(0,1)$ and $\mathcal{CN}(0,1)$ denote standard normal and complex normal distributions respectively. Finally, $[n] = \{1, \ldots, n\}$.

2. Problem Formulation

We consider the inverse problem of estimating $\mathbf{X}$ from $\mathbf{Y}$ in the convolutional model

$$\mathbf{Y} = \mathbf{K} * \mathbf{X} + \mathbf{\Xi}, \tag{1}$$

where $\mathbf{X} \in \mathbb{C}^{n_x \times T}$, $\mathbf{Y} \in \mathbb{C}^{n_y \times T}$, $\mathbf{K} \in \mathbb{C}^{n_y \times n_x \times k}$ with $\mathbf{K}_{i**} \in \mathbb{C}^{n_x \times k}$ being the ith convolutional kernel of width k, and $\mathbf{\Xi}$ is a noise matrix with the same shape as $\mathbf{Y}$ and i.i.d. zero-mean complex normal entries $\mathcal{CN}(0, \sigma^2)$. See Appendix A for a brief overview of complex normal distribution. The circular convolution in equation (1) is defined

as

$$\mathbf{Y}_{i*} = \mathbf{K}_{i**} * \mathbf{X} + \mathbf{\Xi}_{i*} \qquad (2)$$

$$\mathbf{Y}_{it} = \langle \mathbf{K}_{i**}, \mathbf{X}_{*,t:t+k} \rangle + \mathbf{\Xi}_{it}, \quad i \in [n_y], t \in [T] \quad (3)$$

Note that in (3) we are not using the correct definition of the convolution operation where the kernel (or the signal $\mathbf{X}$) is reflected along the time axis, but rather we are using the common definition used in machine learning.

We consider the inference problem in the Bayesian setting where the signal $\mathbf{X}$ is assumed to have a prior that admits a density (with respect to Lebesgue measure) $p(\mathbf{X})$. Further, we assume that rows of $\mathbf{X}_i$ are i.i.d. such that this density factorizes as

$$p(\mathbf{X}) = \prod_{i=1}^{n_x} p(\mathbf{X}_{i*}). \qquad (4)$$

The convolution kernel $\mathbf{K}$ is assumed to be known with i.i.d. entries drawn from $\mathcal{CN}(0, 1/(n_y k))$. Given this statistical model, the posterior is

$$p(\mathbf{X}|\mathbf{Y}) \propto p(\mathbf{X}, \mathbf{Y}) = p(\mathbf{X})p(\mathbf{Y}|\mathbf{X}), \qquad (5)$$

where with some abuse of notation, we are using $p(\cdot)$ to represent the densities of all random variables to simplify the notation. From the Gaussianity assumption on noise we have

$$\mathbf{Y}_{i*}|\mathbf{X} \sim \mathcal{CN}(\mathbf{K}_{i**} * \mathbf{X}, \sigma^2 \mathbf{I}), \qquad (6)$$

where $\mathbf{I}$ is the identity matrix of size $T \times T$.

Given the model in (5), one can consider different types of estimators for $\mathbf{X}$. Of particular interest are regularized M-estimators

$$\widehat{\mathbf{X}}_{\mathsf{m\text{-}est}} = \arg\min_{\mathbf{X}} \mathcal{L}(\mathbf{X}, \mathbf{Y}) + \mathcal{R}(\mathbf{X}), \qquad (7)$$

where $\mathcal{L}$ is a loss function and $\mathcal{R}$ corresponds to the regularization term. Taking negative log-likelihood as the loss and negative log-prior as the regularization we get the *maximum a posteriori* (MAP) estimator

$$\widehat{\mathbf{X}}_{\mathsf{MAP}} := \arg\min_{\mathbf{X}} -\log p(\mathbf{Y}|\mathbf{X}) - \log p(\mathbf{X}), \qquad (8)$$

which selects mode of the posterior as the estimate.

In this work we are interested in analyzing the performance of ridge-regularized least squares estimator which is another special case of the regularized M-estimator in (7) with square loss and ℓ_2-norm regularization

$$\widehat{\mathbf{X}}_{\mathsf{ridge}} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{K} * \mathbf{X}\|_{\mathrm{F}}^2 + \lambda \|\mathbf{X}\|_{\mathrm{F}}^2, \qquad (9)$$

where λ is the regularization parameter. By Gaussianity of the noise, this can also be thought of as ℓ_2-regularized maximum likelihood estimtor. Given an estimate $\widehat{\mathbf{X}}$, one is usually interested in performance of the estimator based on some metric. In Bayesian setting, the metric is usually an average risk (averaged over the prior and the randomness of data)

$$R = \mathbb{E}\ell(\mathbf{X}, \widehat{\mathbf{X}}), \qquad (10)$$

where ℓ is some loss function between the true parameters and the estimates. The most widely used loss is the squared error where $\ell(\mathbf{X}, \widehat{\mathbf{X}}) = \|\mathbf{X} - \widehat{\mathbf{X}}\|_{\mathrm{F}}^2$. Our theoretical results exactly characterize the mean squared error (MSE) of ridge estimator (9) in a certain high-dimensional regime described below.

3. Main Result

Similar to other works in this area (Bayati and Montanari, 2011; Rangan et al., 2019; Pandit et al., 2020), our goal is to analyze the average case performance of the ridge estimator in (9) for the convolutional model in (1) in a certain high dimensional regime that is called *large system limit* (LSL).

3.1. Large System Limit

We consider a sequence of problems indexed by T and n_x. We assume that $k := k(T)$ and $n_y := n_y(n_x)$ are functions of T and n_x respectively and

$$\lim_{n_x \to \infty} \frac{n_y(n_x)}{n_x} = \delta \in (0, \infty), \quad \lim_{T \to \infty} \frac{k(T)}{T} = \beta \in (0, 1].$$

This doubly asymptotic regime where both the number of parameters and unknowns are going to infinity at a fixed ratio is sometimes called proportional asymptotic regime in the literature. Note that even though we make the assumption that k and T are going to infinity at a fixed ratio throughout this paper, the linear scaling of k with respect to T is not necessary. We only need $T \to \infty$ and k could scale at an arbitrary rate with respect to T (or even remain a constant) so long as certain the variance of kernel entries is scaled appropriately (see Remark 5 for more details).

3.1.1. ASSUMPTIONS ON THE NOISE AND KERNEL

We assume that the entries of the convolution kernel $\mathbf{K}$ and the noise $\mathbf{\Xi}$ converge empirically with second order to random variables K and Ξ, with distributions $\mathcal{CN}(0, \sigma_K^2/(kn_y))$ and $\mathcal{CN}(0, \sigma^2)$ respectively. See Appendix B for definition of empirical convergence of random variables and some useful results. Informally, empirical convergence means that as we take the limit of $T \to \infty$ and $x_x \to \infty$ the empirical distribution of entries of convolution kernel tensor and entries of noise matrix in (1) converge to the distribution of complex normal random variables. As stated formally in the appendix, the convergence of the distributions is with respect to Wasserstein-2 metric on the space of distributions (see Proposition 1 for the precise statement).

3.1.2. ASSUMPTIONS ON $\mathbf{X}_{i*}$

Next, based on (Peligrad et al., 2010), we state the distributional assumptions on the rows of $\mathbf{X}$ that we require for our theory to hold. Let $\{\xi_t\}_{t\in\mathbb{Z}}$ be a stationary ergodic Markov chain defined on a probability space $(S, \mathcal{F}, P)$ and let $\pi(\cdot)$ be the distribution of ξ_0. Note that stationarity implies all ξ_ts have the same marginal distribution. Define the space of functions

$$\mathcal{L}_0^2(\pi) := \left\{ h | \mathbb{E}_\pi[h(\xi_0)] = 0, \mathbb{E}_\pi[h^2(\xi_0)] < \infty \right\}. \quad (11)$$

Define $\mathcal{F}_k := \sigma(\{\xi_t\}_{t\leq k})$, the σ-algebra generated by ξ_t up to time k and let $x_t = h(\xi_t)$ for some $h \in \mathcal{L}_0^2(\pi)$. We assume that the process $\{x_t\}_{t\in\mathbb{Z}}$ satisfies the regularity condition

$$\mathbb{E}[x_0|\mathcal{F}_{-\infty}] = 0, \quad P - \text{almost surely.} \quad (12)$$

The class of processes that satisfy these conditions is quite large and includes i.i.d. random processes as an example. It also includes causal functions of i.i.d. random variables of the form $X_n = f(\xi_k, k \leq n)$ where ξ_k is i.i.d. such as autoregressive (AR) processes and many Markov chains. See (Peligrad et al., 2010) for more examples satisfying these conditions.

We assume that each row $\mathbf{X}_{i*}$ of $\mathbf{X}$, is an i.i.d. sample of a process that satisfies the conditions mentioned. Let $\widetilde{\mathbf{X}}_i(\omega)$ be its (normalized) Fourier transform (defined in (18)). and define $g(\omega) := \lim_{T\to\infty} \mathbb{E}|\widetilde{\mathbf{X}}_i(\omega)|^2$. As shown in (Peligrad et al., 2010), since the rows are i.i.d., this limit is the same for all the rows. Also, $g(\omega)$ is proportional to the spectral density of the process that generates the rows of $\mathbf{X}$

$$g(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} c_n \exp(-\mathrm{i}\omega t), \quad c_t = \mathbb{E}[\mathbf{X}_{i0}\mathbf{X}_{it}].$$
$$(13)$$

As we will see in the next section, $g(\omega)$ plays a key role in characterization of estimation error of ridge estimator in convolutional linear inverse problems that have such processes as inputs.

3.2. Asymptotics of Ridge Estimator

The main result of our paper characterizes the limiting distribution to which the the Fourier transform of the true signal $\widetilde{\mathbf{X}}_0(\omega)$ and Fourier transform of the estimated signal $\widehat{\widetilde{\mathbf{X}}}_{\mathrm{ridge}}(\omega)$ converge. The proof can be found in Section 4. In the following $\mathcal{B}$ and μ represent Borel σ-algebra and Lebesgue measure respectively.

Theorem 1. Under the assumptions in Section 3.1, as $n_x, n_y, T, k \to \infty$, over the product space $([0, 2\pi] \times S^{n_x}, \mathcal{B} \times \mathcal{F}^{n_x}, \mu \times P^{n_x})$ the Ridge estimator satisfies

$$\begin{bmatrix} \widetilde{\mathbf{X}}_0(\omega) \\ \widehat{\widetilde{\mathbf{X}}}_{\mathrm{ridge}}(\omega) \end{bmatrix} \overset{d, PL(2)}{=} \begin{bmatrix} \sqrt{g(U)}Z_0 \\ \alpha(\sqrt{g(U)}Z_0 + \tau(g(U))Z_1) \end{bmatrix},$$

where $U \sim \mathrm{unif}([0, 2\pi])$, $Z_0, Z_1 \sim \mathcal{CN}(0, 1)$ where $\mathcal{CN}(0, 1)$ is the standard complex normal distribution, U, Z_0 and Z_1 are independent of each other, α is the smaller root of the quadratic equation

$$\frac{\lambda}{\sigma_K^2} = \frac{(1-\alpha)(1-\alpha/\delta)}{\alpha}, \quad (14)$$

and

$$\tau^2(g(U)) = \frac{\sigma^2/\sigma_K^2 + (1-\alpha^2)g(U)/\delta}{1 - \alpha^2/\delta}. \quad (15)$$

The convergence in this theorem is weak convergence for ω and PL(2) for $\widetilde{\mathbf{X}}_0(\omega)$ and $\widehat{\widetilde{\mathbf{X}}}_{\mathrm{ridge}}(\omega)$. See appendix B for the definition of this convergence. This convergence result allows us to find the asymptotic mean squared error of ridge estimator for the convolutional model as an integral.

Corollary 1. Under the same assumptions as in Theorem 1, ridge estimator satisfies

$$\lim_{n_x\to\infty} \lim_{T\to\infty} \frac{1}{n_x T}\|\widehat{\widetilde{\mathbf{X}}}_{\mathrm{ridge}} - \widetilde{\mathbf{X}}\|_{\mathrm{F}}^2$$
$$= \int_0^{2\pi} \left((\alpha - 1)^2 g(\omega) + \alpha^2 \tau(g(\omega)) \right) d\omega. \quad (16)$$

Remark 1. As shown in the proof Lemma 6, for $\lambda \geq 0$, the quadratic equation (14) always has two real positive solutions the smaller of which determines the error.

Remark 2. The $1/\sqrt{T}$ scaling in our definition of Fourier operator in Section 4.2 makes it a unitary operator, i.e. ℓ_2 norm is preserved under the Fourier transform and its inverse. This implies

$$\|\widehat{\mathbf{X}}_{\mathrm{ridge}} - \mathbf{X}\|_{\mathrm{F}} = \|\widehat{\widetilde{\mathbf{X}}}_{\mathrm{ridge}} - \widetilde{\mathbf{X}}\|_{\mathrm{F}}. \quad (17)$$

Therefore, the result of Corollary 1 also holds in time domain.

Remark 3. If rows $\mathbf{X}_{i*}$ have zero mean i.i.d. entries, then the correlation coefficients c_t in (13) are all zero except for c_0. Therefore, $g(\omega) = g$ where g is constant. Hence, in this case, the integrand in (16) would be a constant and the estimation error across all the frequencies would be the same as the estimation error in the ordinary ridge regression as in Lemma 6, i.e. the error vs. δ would be exactly the same as the double descent curve in ordinary ridge regression with i.i.d. priors. In other words, the double descent curve for ordinary ridge regression carries over to the convolutional ridge regression for i.i.d. priors (see Figure 1).

4. Proof

In this section we present the proof of Theorem 1. Before presenting the details of the proof, it is helpful to see an overview of the proof.

4.1. Proof overview

We first show that convolutional models turn into ordinary linear models for each frequency in Fourier domain. We then show that ridge estimators in time domain can also be written as ridge estimators in frequency domain. This uses the fact that Fourier transform matrix, with appropriate normalization is a unitary matrix, and ℓ_2 norm is preserved under unitary transformation, i.e. it is an isometry. Next we use properties of Fourier transform of random processes to show that under certain conditions, they asymptotically converge to a Gaussian process in frequency domain that is independent across different frequencies for almost every frequency. These together allow us to turn the ridge estimation in time domain into multiple ridge estimators in frequency domain, one for each frequency. We then use theoretical properties of ridge estimators to derive estimation error for each of these ridge estimators and integrate them over frequencies to derive our main result.

Our proof is based on previous results for asymptotic error of ridge estimators for ordinary linear inverse problems. This has been studied in many works (Dicker et al., 2016; Dobriban et al., 2018; Hastie et al., 2019) where the authors take advantage of the fact that ridge estimators have a closed form solution that can be analyzed, e.g. using results from random matrix theory. In this work we use approximate message passing (Donoho et al., 2010; Bayati and Montanari, 2011) to derive the asymptotic error of ridge estimators.

4.2. Main Technical Lemmas

In order to prove Theorem 1, we need several lemmas. We first characterize the convolutional model in (1) in Fourier domain. For $\omega \in \{0, 1 \cdot \frac{2\pi}{T}, \ldots, (T-1) \cdot \frac{2\pi}{T}\} =: \Omega$, let $\widetilde{\mathbf{X}}_j(\omega)$ be the discrete (circular) Fourier transform (DFT) of $\mathbf{X}_{j*}$

$$\widetilde{\mathbf{X}}_j(w) = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} \mathbf{X}_{jt} \exp(-\mathrm{i}\omega t). \tag{18}$$

If we let $\mathbf{F}$ to denote the T-point unitary DFT matrix, i.e. $\mathbf{F}_{mn} = 1/\sqrt{T} \exp(-\mathrm{i}2\pi mn/T)$, then this equation can be written in matrix form as $\widetilde{\mathbf{X}}_i(*) = \mathbf{X}_{i*}\mathbf{F}$. Define $\widetilde{\mathbf{Y}}_i(\omega)$, $\widetilde{\mathbf{K}}_{ij}(\omega)$, and $\widetilde{\mathbf{\Xi}}(\omega)$ similarly. Note that in these definitions, we have included a $1/\sqrt{T}$ factor which is usually not included in the definition of Fourier transform, but since this makes the Fourier matrix unitary, it eases our notation slightly. With this definition we have $\mathbf{F}^{\mathsf{T}} = \mathbf{F}$ and $\mathbf{F}^*\mathbf{F} = \mathbf{I}$.

Lemma 1 (Convolutional models in Fourier domain)**.** The convolutional model in (1) can be written in Fourier domain as

$$\widetilde{\mathbf{Y}}(\omega) = \sqrt{T}\widetilde{\mathbf{K}}_{**}(\omega)\widetilde{\mathbf{X}}(\omega) + \widetilde{\mathbf{\Xi}}(\omega), \quad \forall \omega \in \Omega. \tag{19}$$

Proof. Taking Fourier transform of Equation (2) and using the convolution theorem we get

$$\widetilde{\mathbf{Y}}_i(\omega) = \sqrt{T}\sum_{j=1}^{n_x} \widetilde{\mathbf{K}}_{ij}(\omega)\widetilde{\mathbf{X}}_j(\omega) + \widetilde{\mathbf{\Xi}}_i(\omega). \tag{20}$$

Note that the $\sqrt{T}$ factor on the right hand is due to our definition of Fourier transform where we have used a $1/\sqrt{T}$ factor to make the Fourier operator unitary. Rewriting this equation in matrix form gives us the desired result. □

The next lemma characterizes the ridge estimator in (9) in frequency domain.

Lemma 2. The ridge estimator in (9) in frequency domain is equivalent to solving separate ordinary ridge regressions for each $\omega \in \Omega$:

$$\widehat{\widetilde{\mathbf{X}}}_{\mathsf{ridge}}(\omega) = \arg\min_{\widetilde{\mathbf{X}}(\omega)} \left\| \widetilde{\mathbf{Y}}(\omega) - \sqrt{T}\widetilde{\mathbf{K}}(\omega)\widetilde{\mathbf{X}}(\omega) \right\|_2^2$$
$$+ \lambda \left\| \widetilde{\mathbf{X}}(\omega) \right\|_2^2. \tag{21}$$

Proof. Since the Fourier matrix is unitary, we have

$$\|\mathbf{X}\|_{\mathrm{F}} = \|\mathbf{X}\mathbf{F}\|_{\mathrm{F}} = \|\widetilde{\mathbf{X}}\|_{\mathrm{F}}$$
$$\|\mathbf{Y} - \mathbf{K} * \mathbf{X}\|_{\mathrm{F}} = \|(\mathbf{Y} - \mathbf{K} * \mathbf{X})\mathbf{F}\|_{\mathrm{F}} = \|\widetilde{\mathbf{Y}} - \sqrt{T}\widetilde{\mathbf{K}}\widetilde{\mathbf{X}}\|_{\mathrm{F}},$$

where $\widetilde{\mathbf{K}}\widetilde{\mathbf{X}}$ is a tensor-matrix product defined as $(\widetilde{\mathbf{K}}\widetilde{\mathbf{X}})(\omega) = \widetilde{\mathbf{K}}(\omega)\widetilde{\mathbf{X}}(\omega)$. Then, a change of variable $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{F}$ in (9) proves the lemma. □

Lemma 3. If the kernel $\mathbf{K}$ has i.i.d. $\mathcal{CN}(0, \sigma_K^2/(kn_y))$ entries, then for each $\omega \in \Omega$, $\sqrt{T}\widetilde{\mathbf{K}}(\omega)$ has i.i.d. complex normal entries $\mathcal{CN}(0, \sigma_K^2/n_y)$.

Proof. The DFT of the kernel is

$$\widetilde{\mathbf{K}}_{ij}(\omega) = \frac{1}{\sqrt{T}} \sum_{t=0}^{k-1} \mathbf{K}_{ijt} \exp(-\mathrm{i}\omega t). \tag{22}$$

This is a linear combination of complex Gaussian random variables and therefore, $\widetilde{\mathbf{K}}$ is a tensor with jointly complex Gaussian entries. Clearly, $\mathbb{E}(\widetilde{\mathbf{K}}) = \mathbf{0}$ and for $(i, j) \neq (i', j')$, using independence of $\mathbf{K}_{ij*}$ and $\mathbf{K}_{i'j'*}$ we have

$$\mathbb{E}\widetilde{\mathbf{K}}_{ij}(\omega)\widetilde{\mathbf{K}}_{i'j'}(\omega') = 0, \quad \mathbb{E}\widetilde{\mathbf{K}}_{ij}(\omega)\widetilde{\mathbf{K}}_{i'j'}^*(\omega') = 0 \quad \forall \omega, \omega',$$

which proves that for any ω, $\widetilde{\mathbf{K}}(\omega)$ has independent entries and all the dependence in $\widetilde{\mathbf{K}}$ is across different frequencies.

It remains to find the variance and relation of each entry of $\widetilde{\mathbf{K}}(\omega)$. Let $\mathbf{k} := \mathbf{K}_{ij*}$ for some i and j be a row vector, and

let $\widetilde{\mathbf{k}} := \widetilde{\mathbf{K}}_{ij}(*)$. Then we have $\widetilde{\mathbf{k}} = \mathbf{k}\mathbf{F}$ and the variance of $\sqrt{T}\widetilde{\mathbf{k}}_m$ is

$$\gamma = T\mathbb{E}[\widetilde{\mathbf{k}}_m \widetilde{\mathbf{k}}_m^*] = T\mathbb{E}[\mathbf{k}\mathbf{F}_{*m}\mathbf{F}_{*m}^* \mathbf{k}^\mathsf{T}] \tag{23}$$

$$= T\mathbf{F}_{m*}\mathbb{E}[\mathbf{k}^\mathsf{T}\mathbf{k}]\mathbf{F}_{m*}^* = \frac{T\sigma_K^2}{kn_y}\mathbf{F}_{m*}\begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{F}_{m*}^* \tag{24}$$

$$= \frac{\sigma_K^2}{kn_y}\sum_{t=0}^{k-1}\exp(\frac{2\pi \mathrm{i}tm}{T})\exp(\frac{-2\pi \mathrm{i}tm}{T}) \tag{25}$$

$$= \frac{\sigma_K^2}{n_y}. \tag{26}$$

Similarly, the relation is

$$c(m) = T\mathbb{E}[\widetilde{\mathbf{k}}_m \widetilde{\mathbf{k}}_m^\mathsf{T}] = T\mathbb{E}[\mathbf{k}\mathbf{F}_{*m}\mathbf{F}_{m*}\mathbf{k}^\mathsf{T}]$$
$$= T\mathbf{F}_{m*}\mathbb{E}[\mathbf{k}^\mathsf{T}\mathbf{k}]\mathbf{F}_{*m} = \frac{\sigma_K^2}{k}\mathbf{F}_{m*}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\mathbf{F}_{*m}$$
$$= 0.$$

Therefore, for each ω, $\widetilde{\mathbf{K}}_{ij}(\omega) \sim \mathcal{CN}(0, \sigma_K^2/n_y)$ and they are i.i.d. for all i, j. □

Remark 4. Observe that the scaling of variance of entries of $\mathbf{K}$ with $1/k$ is crucial to get a non-trivial distribution for entries of $\sqrt{T}\widetilde{\mathbf{K}}(\omega)$ as we take the limit $T \to \infty$.

Remark 5. Here, we should emphasize that the linear scaling of k with respect to T is not necessary for this Lemma to hold. All we need is for the k and variance of the convolution kernel to scale in such a way that in Fourier domain the entries of $\widehat{\mathbf{K}}$ have i.i.d. $\mathcal{CN}(0, \sigma_K^2/n_y)$ entries. This can be achieved by appropriate scaling of variance in time domain and k even for the case that k is constant.

Lemma 4. If noise Ξ has i.i.d. $\mathcal{CN}(0, \sigma^2)$ entries, then $\widetilde{\Xi}$ has i.i.d. complex normal entries $\widetilde{\Xi}_{ij}(\omega) \sim \mathcal{CN}(0, \sigma^2)$, i.e.

$$\Xi_{ij}(\omega) \stackrel{d}{=} \frac{\sigma^2}{2}Z_1 + \frac{\sigma^2}{2}Z_2, \quad Z_1, Z_2 \sim \mathcal{N}(0,1), Z_1 \perp\!\!\!\perp Z_2.$$

Proof. The proof is similar to the proof of Lemma 3 with $k = T$. □

Lemma 4 is the complex analogue of the fact that distribution of vectors with i.i.d. Gaussian entries is invariant under orthogonal transformations.

As stated in Appendix B, for a Gaussian random sequence, convergence in the first and second moments implies convergence in Wasserstein distance which is equivalent to $PL(2)$ convergence. Therefore, Lemma 3 and 4 also imply that the entries of kernel and noise for each frequency are converging empirically with second order to i.i.d. complex Gaussian random variables. As shown in the appendix, this convergence is stronger than convergence in distribution.

Next, we mention a result about Fourier transform of random processes from (Peligrad et al., 2010).

Lemma 5 (Fourier transform of random processes (Peligrad et al., 2010))**.** Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stationary ergodic process that satisfies the assumptions in Section 3.1.2. Let $\widetilde{X}(\omega)$ be its (normalized) Fourier transform and $g(\omega) := \lim_{T\to\infty}\mathbb{E}|\widetilde{X}(\omega)|^2$. Then on the product space $([0, 2\pi] \times S, \mathcal{B} \times \mathcal{F}, \mu \times P)$ we have

$$\widetilde{X}(\omega) \stackrel{d}{=} \sqrt{g(U)}\mathcal{CN}(0,1), \tag{27}$$

where $U \sim \mathrm{unif}([0, 2\pi])$ is independent of $\mathcal{CN}(0,1)$.

Lemma 5 allows us to characterize the limiting distribution of each row $\mathbf{X}_{i*}$ of the input in the frequency domain, i.e. distribution of $\widetilde{\mathbf{X}}_i(\omega)$ as $T \to \infty$.

All the lemmas so far allow us to look at the convolutional model in the frequency domain. We need one last lemma to characterize the asymptotics of ridge regression in high dimensions.

Lemma 6 (Asymptotics of ridge regression)**.** Consider the linear model $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \boldsymbol{\xi}$, where $\mathbf{x} \in \mathbb{R}^{n_x}, \boldsymbol{\xi} \in \mathbb{R}^{n_y}$, and $\mathbf{A} \in \mathbb{R}^{n_y \times n_x}$ all have i.i.d. components with $\mathbf{x}_i \sim \mathcal{N}(0, \sigma_x^2), \boldsymbol{\xi}_i \sim \mathcal{N}(0, \sigma^2)$, and $\mathbf{A}_{ij} \sim \mathcal{N}(0, \sigma_A^2/n_y)$. Then the ridge estimator

$$\widehat{\mathbf{x}}_{\mathsf{ridge}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2 \tag{28}$$

as $n_x, n_y \to \infty$ with $n_y/n_x \to \delta$ satisfy

$$\begin{bmatrix} \mathbf{x}_0 \\ \widehat{\mathbf{x}}_{\mathsf{ridge}} \end{bmatrix} \stackrel{PL(2)}{=} \begin{bmatrix} X_0 \\ \alpha(X_0 + \tau(\sigma_x)Z) \end{bmatrix}, \tag{29}$$

where $X_0 \sim p_X, Z \sim \mathcal{N}(0,1)$ independent of X_0, α is the smaller root of the quadratic equation

$$\frac{\lambda}{\sigma_A^2} = \frac{(1-\alpha)(1-\alpha/\delta)}{\alpha}, \tag{30}$$

and

$$\tau^2(\sigma_x^2) = \frac{\sigma^2/\sigma_A^2 + (1-\alpha^2)\sigma_x^2/\delta}{1 - \alpha^2/\delta}. \tag{31}$$

Therefore, we almost surely have

$$\lim_{n_x \to \infty} 1/n_x \|\mathbf{x}_{\mathsf{ridge}} - \mathbf{x}_0\|_2^2 = (\alpha - 1)^2\sigma_X^2 + \alpha^2\tau^2(\sigma_x).$$

Proof. This is a consequence of using *approximate message passing* (AMP) algorithm to solve (28). We have mentioned in the overview of the proof our main theorem (4.1) that AMP is only one of the ways to obtain the asymptotic error of ridge estimators in ordinary linear problems. There are several other methods mostly based on random matrix theory that also derive such asymptotic errors, but in this work we adopt the AMP approach. See Appendix D for a brief

introduction to AMP and Appendix D.1 for the proof of this lemma. Here we briefly mention the sketch of the proof. AMP is a recursive algorithm that originally was proposed to solve linear inverse problems but since has been applied to many different problems. The key property of AMP algorithms is that a set of equations called the state evolution (SE) exactly characterize the performance of AMP at each iteration. As shown in the appendix, the AMP recursions to solve ridge regression in (28) are

$$\mathbf{x}^{t+1} = \alpha(\mathbf{A}^\mathsf{T}\mathbf{z}^t + \mathbf{x}^t), \tag{32}$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\alpha}{\delta}\mathbf{z}^{t-1}, \tag{33}$$

where α is the smaller root of the quadratic equation in (30). In the appendix, we prove that for $\lambda \geq 0$, the roots of this equation are real and positive, and only for the smaller root the AMP algorithm converges. Finally, using the state evolution, we obtain that the ridge estimator and true values of $\mathbf{x}$ jointly converge as in (29), and $\tau^2(\sigma_x)$ in (31) is the fixed point value of state evolution recursion in (69). $\quad\square$

This lemma allows us to find asymptotics of ridge regression for real linear inverse problems. Even though the τ in (31) depends also on α, δ, and noise variance, we have only made the dependence on σ_x explicit, as all the other parameters will be fixed for the ridge regression problem for each frequency, but σ_x could change as a function of frequency.

Remark 6. The exact same result holds for complex valued ridge regression *mutatis mutandis*, i.e. by changing normal distributions $\mathcal{N}(0, \cdot)$ to complex normal distributions $\mathcal{CN}(0, \cdot)$.

Remark 7. The requirements of Lemma 6 can be relaxed. Rather than requiring $\mathbf{x}$ or $\boldsymbol{\xi}$ to have i.i.d. Gaussian entries, we only need them to converge PL(2) to random variables with these distributions.

4.3. Proof of Theorem 1

We now have all the ingredients to prove Theorem 1. Lemma 2 allows us to find Fourier transform of Ridge estimator in 9 using a series of ridge regressions in Fourier domain. Lemmas 3 and 4 show that the Fourier transform of the convolution kernel and the noise and the signal asymptotically have complex Gaussian distributions with i.i.d. components for each frequency. Then, Lemma 5 shows that

$$\widetilde{\mathbf{X}}_i(\omega) \stackrel{d}{=} \sqrt{g(U)}\mathcal{CN}(0,1), \tag{34}$$

where $g(\cdot)$ is defined in the Lemma.

Next, the complex version of Lemma 6 would give us the asymptotic error of ridge estimator on the product space

$([0, 2\pi] \times S, \mathcal{B} \times \mathcal{F}, \lambda \times P)$ in the limit

$$\begin{bmatrix} \widetilde{\mathbf{X}}_0(\omega) \\ \widetilde{\mathbf{X}}(\omega) \end{bmatrix} \stackrel{PL(2)}{=} \begin{bmatrix} \sqrt{g(U)}Z_0 \\ \alpha(\sqrt{g(U)}Z_0 + \tau(g(U))Z_1) \end{bmatrix}, \tag{35}$$

where $U \sim \mathrm{unif}([0, 2\pi])$, and $Z_0, Z_1 \sim \mathcal{CN}(0, 1)$, and $\tau(\cdot)$ is the function in (31). Note that the variance of $\widetilde{\mathbf{X}}_0(\omega)$ is $g(\omega)$, hence the term $\tau(g(U))Z_1$. As mentioned earlier, this variance is the only variable that changes with frequency while all the other parameters are the same for all frequencies. Using this convergence, in the limit, the error is

$$\lim_{n_x \to \infty} \lim_{T \to \infty} \frac{1}{n_x T} \|\widehat{\widetilde{\mathbf{X}}}_{\mathsf{ridge}} - \widetilde{\mathbf{X}}\|_{\mathrm{F}}^2$$
$$= \int_0^{2\pi} \left((\alpha - 1)^2 g(\omega) + \alpha^2 \tau(g(\omega)) \right) d\omega. \tag{36}$$

As mentioned in Section 4.2, our scaling of the Fourier operator makes it a unitary operator. Therefore, ℓ_2 norm is preserved under our definition of Fourier transform and its inverse. This implies that the same result as in (36) also holds in *time* domain.

5. Experiments

In this section we validate our theoretical results on simulated data. We generate data using a ground truth convolutional model of the form (3). We use i.i.d. complex normal convolution kernel and noise with different variances. For the data matrix $\mathbf{X}$, we consider two different models: i) i.i.d. complex normal data; and ii) a non-Gaussian autoregressive process of order 1 (an AR(1) process). In both cases we take $T = 256$, $n_y = 500$ and use different values of n_x to create plots of estimation error with respect to $\delta = n_y/n_x$.

AR(1) process is a process that evolves in time as

$$x_t = ax_{t-1} + \xi_t, \tag{37}$$

where ξ_t is some zero mean random noise and a is a fixed constant. Note that our assumptions on the process require it to be a stationary and ergodic process. These assumptions are only satisfied when the noise is i.i.d. and $|a| < 1$. The parameter a controls how fast the process is mixing. The case were $a = 0$ results in an i.i.d. process, whereas values with magnitude close to 1 would result in a process that has strong correlations over a long period of time. This process has the form of one of the examples given in Section 3.1.2 and hence satisfies all the assumptions required for our theory to hold.

If the noise ξ_t has zero mean Gaussian distribution, the process will be a centered Gaussian process which is completely characterized by an auto-correlation function

$$R[t] = \mathbb{E}[x_{t'+t}x_{t'}], \tag{38}$$

where we have used the fact that stationarity of the process implies this auto-correlation only depends on the time difference and not on the actual time. Since for Gaussian processes, the Fourier transform is another Gaussian process, we do not need to use the results of (Peligrad et al., 2010) to analyze them. As such, a more interesting example would be to use a non-Gaussian noise ξ. Therefore, besides the Gaussian AR process, we also take $\xi_t \sim \mathrm{unif}(\{-s, s\})$ where s is an step size that controls the variance of the process. The variance and auto-correlation of a univariate AR(1) process can be found as follows. Squaring both sides of (37) and taking the expectation we obtain

$$\mathbb{E}[x^2] = \frac{\mathbb{E}[\xi^2]}{1 - a^2}. \tag{39}$$

Similarly, it is also easy to show that the auto-correlation of this process is

$$R[t] = \frac{\mathbb{E}[\xi^2]}{1 - a^2} a^{|t|}. \tag{40}$$

The auto-correlation function of the AR process only depends on the variance of the noise and not its distribution. Our main result, Theorem 1, shows that the asymptotic error of ridge estimator depends on the the underlying process only through the function $g(\omega)$ which as stated earlier, is proportional to the spectral density of the process, i.e. norm of Fourier transform of the auto-correlation function. Therefore, so long as the zero mean noise has the same variance, irrespective of its distribution, we expect to see the same asymptotic error in the convolutional ridge regression when the rows of $\mathbf{X}$ are i.i.d. samples of such processes. To show this, we use both a Gaussian AR(1) process with $\mathrm{var}(\xi_t^2) = 0.1$ as well as $\xi_t \, \mathrm{unif}(\{-s, s\})$ with $s = \sqrt{0.1}$ to match the variances. In both cases we take $a = 0.9$ and measurement noise variance $\sigma^2 = 0.1$.

We first present the results for the i.i.d. Gaussian covariates. In this case the variance of signal and noise are 0.004 and 1 respectively. Figure 1 shows the log of normalized estimation error with respect to $\delta = n_y/n_x$ for three different values of the regularization parameter λ. Normalized estimation error is defined as

$$\mathrm{NMSE} = \frac{\mathbb{E}\|\widehat{\mathbf{X}}_{\mathsf{ridge}} - \mathbf{X}_0\|_2^2}{\mathbb{E}\|\mathbf{X}_0\|_2^2}. \tag{41}$$

The solid curves correspond to what our theory predicts and the dots correspond to what we observe on synthetic data. Even though our results hold in the limit of $n_x, n_y, k, T \to \infty$ at proportional ratio, we can see that already at this problem size, there is an almost perfect match between our predictions and the error that is observed in practice. This suggests that the errors concentrate around these asymptotic values. The figure also shows the double descent phenomenon where as the number of parameters increases beyond the interpolation threshold, the error starts decreasing
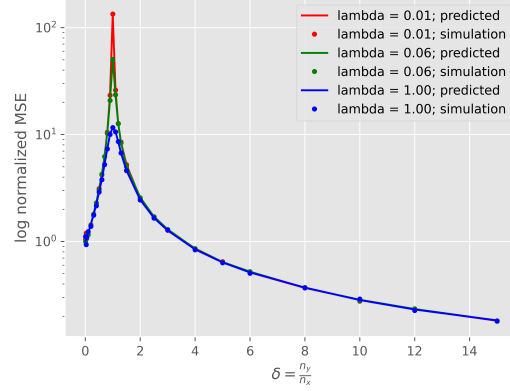

Figure 1: Log of normalized error for i.i.d. Gaussian features with respect to $\delta = n_y/n_x$ for three different values of λ. Solid lines show the predictions of our theory whereas the dots show the observed error on synthetic data.

again. It can be seen that regularization helps with pulling the estimation error down in vicinity of the interpolation threshold. The interpolation threshold is where we have just enough parameters to fit the observations perfectly. This happens at $\delta = 1$, i.e. $n_x = n_y$.

As mentioned in Remark 3 an i.i.d. process has a white spectrum in frequency domain, meaning that $\boldsymbol{g}(\omega)$ is a constant. Therefore, for these processes, the integral in Corollary 1 would be proportional to the integrand. The integrand in turn is the asymptotic error of an ordinary ridge regression problem with i.i.d. Gaussian features. As such, this figure is essentially the same as the figures in papers that have looked at asymptotics of ridge regression, some of which we have mentioned in the Introduction and prior work.

Figure 2 shows the same plot for the case where the rows of $\mathbf{X}$ are i.i.d. samples of an AR(1) process with process noise $\mathrm{unif}(\{-s, s\})$. The asymptotic error for processes that have dependencies over time can be significantly different from i.i.d. random features. The red curve is similar to the red curve in Figure 1, but the other curves show very different behavior. The double descent phenomenon is still present here. The plot for AR(1) process with Gaussian noise is essentially identical to Figure 2 and we have moved it to the appendix (Figure 3). This supports our theoretical result that error in this asymptotic regime only depends on the spectral density of the process.

6. Conclusion

Summary. We characterized the performance of ridge estimator for convolutional models in proportional asymptotics regime. By looking at the problem in Fourier domain, we showed that the asymptotic mean squared error of ridge
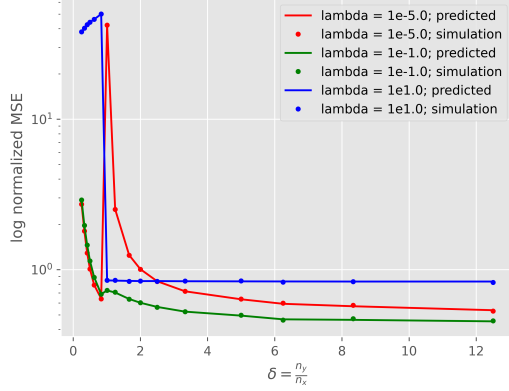
Figure 2: Log of normalized error for the AR(1) features with the process noise $\text{unif}(\{-s, s\})$, with respect to $\delta = n_y/n_x$ for three different values of λ. Solid lines show the predictions of our theory and the dots show the observed error on synthetic data. The plots for Gaussian AR process is essentially identical. to this plot.

estimator can be found from a scalar integral that depends on the spectral properties of the true signal. Our experiments show that our theoretical predictions match what we observe in practice even for problems of moderate size.

Future work. The results of this work only apply to ridge regression estimator for convolutional linear inverse problem. The key property of ridge regularization is that it is invariant under unitary transforms, and hence we could instead solve the problem in frequency domain. Proving such result for general estimators and regularizers allows us to extend this work to inference in deep convolutional neural networks similar to (Pandit et al., 2020). Such work would allow us to obtain the estimation error inverse problems use of deep convolutional generative priors.

Acknowledgements

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446.

Alemohammad, S., Wang, Z., Balestriero, R., and Baraniuk, R. (2020). The recurrent neural tangent kernel. *arXiv preprint arXiv:2006.10246*.

Allen-Zhu, Z., Li, Y., and Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019a). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Belkin, M., Hsu, D., and Xu, J. (2019b). Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*.

Chafaı, D., Chafä, D., Guédon, O., Lecue, G., and Pajor, A. (2009). Singular values of random matrices. *Lecture Notes*.

Deng, Z., Kammoun, A., and Thrampoulidis, C. (2019). A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*.

Dicker, L. H. et al. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37.

Dobriban, E., Wager, S., et al. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

Donoho, D. L., Maleki, A., and Montanari, A. (2010). Message passing algorithms for compressed sensing: I. motivation and construction. In *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*, pages 1–5. IEEE.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. K. (2020). Generalization error of generalized linear models in high dimensions. *arXiv preprint arXiv:2005.00180*.

Friedrich, J., Zhou, P., and Paninski, L. (2017). Fast online deconvolution of calcium imaging data. *PLoS computational biology*, 13(3):e1005423.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Huang, Y., Cheng, Y., Chen, D., Lee, H., Ngiam, J., Le, Q. V., and Chen, Z. (2018). Gpipe: Efficient training of giant neural networks using pipeline parallelism. *CoRR*, abs/1811.06965.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580.

Jury, E. I. (1963). On the roots of a real polynomial inside the unit circle and a stability criterion for linear discrete systems. *IFAC Proceedings Volumes*, 1(2):142–153.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., and Zhang, H. (2019). Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, pages 3496–3506.

Le Cun, Y., Kanter, I., and Solla, S. A. (1991). Second order properties of error surfaces learning time and generalization. *Advances in neural information processing systems*, 3:918–924.

Li, M., Soltanolkotabi, M., and Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166.

Li, Z., Wang, R., Yu, D., Du, S. S., Hu, W., Salakhutdinov, R., and Arora, S. (2019). Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*.

Liang, T., Rakhlin, A., et al. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347.

Liao, Z., Couillet, R., and Mahoney, M. W. (2020). A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *arXiv preprint arXiv:2006.05013*.

Maleki, A., Anitori, L., Yang, Z., and Baraniuk, R. G. (2013). Asymptotic analysis of complex lasso via complex approximate message passing (camp). *IEEE Transactions on Information Theory*, 59(7):4290–4308.

McNally, J. G., Karpova, T., Cooper, J., and Conchello, J. A. (1999). Three-dimensional imaging by deconvolution microscopy. *Methods*, 19(3):373–385.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.

Mueller, C. S. (1985). Source pulse enhancement by deconvolution of an empirical green's function. *Geophysical Research Letters*, 12(1):33–36.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*.

Oymak, S. and Soltanolkotabi, M. (2019). Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960.

Pandit, P., Sahraee-Ardakan, M., Rangan, S., Schniter, P., and Fletcher, A. K. (2020). Inference with deep generative priors in high dimensions. *IEEE Journal on Selected Areas in Information Theory*.

Peligrad, M., Wu, W. B., et al. (2010). Central limit theorem for fourier transforms of stationary processes. *The Annals of Probability*, 38(5):2009–2022.

Rangan, S., Schniter, P., and Fletcher, A. K. (2019). Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. (2018). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769.

Starck, J.-L., Pantin, E., and Murtagh, F. (2002). Deconvolution in astronomy: A review. *Publications of the Astronomical Society of the Pacific*, 114(800):1051.

Treitel, S. and Lines, L. (1982). Linear inverse theory and deconvolution. *Geophysics*, 47(8):1153–1159.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Yang, G. (2020). Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2020). Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492.