

High-Dimensional Analysis for Generalized Nonlinear Regression: From Asymptotics to Algorithm

Jian Li¹, Yong Liu^{2*}, Weiping Wang¹

¹Institute of Information Engineering, Chinese Academy of Sciences

²Gaoling School of Artificial Intelligence, Renmin University of China
lijian9026@iie.ac.cn, liuyonggsai@ruc.edu.cn, wangweiping@iie.ac.cn

Abstract

Overparameterization often leads to benign overfitting, where deep neural networks can be trained to overfit the training data but still generalize well on unseen data. However, it lacks a generalized asymptotic framework for nonlinear regressions and connections to conventional complexity notions. In this paper, we propose a generalized high-dimensional analysis for nonlinear regression models, including various nonlinear feature mapping methods and subsampling. Specifically, we first provide an implicit regularization parameter and asymptotic equivalents related to a classical complexity notion, i.e., effective dimension. We then present a high-dimensional analysis for nonlinear ridge regression and extend it to ridgeless regression in the under-parameterized and over-parameterized regimes, respectively. We find that the limiting risks decrease with the effective dimension. Motivated by these theoretical findings, we propose an algorithm, namely RFRed, to improve generalization ability. Finally, we validate our theoretical findings and the proposed algorithm through several experiments.

Introduction

In conventional machine learning (Vapnik 1999), an explicit regularization term should be added to the learning objective to avoid overfitting, where the model fits the training data well but generalizes poorly on unseen data. From the perspective of statistical learning, the regularization parameter λ balances the bias and variance (Li, Liu, and Wang 2023b). However, recent studies on overparameterized models, including neural networks and kernel methods, have shown that even without explicit regularization, these models often achieve benign overfitting, interpolating the training data while still generalizing well (Belkin, Ma, and Mandal 2018; Liang and Rakhlin 2020; Bartlett et al. 2020; Zhang et al. 2021). Furthermore, the "double descent" performance curve has been observed beyond neural networks (Nakkiran et al. 2021). Based on random matrix theory, subsequent works have theoretically analyzed this phenomenon using high-dimensional asymptotics for various models, including linear models (Belkin, Hsu, and Xu 2020; Hastie et al. 2022), random Fourier features (Liao, Couillet, and Mahoney 2020; Mei and Montanari 2022; Li, Liu, and Zhang

2022), neural networks (Ba et al. 2020; Frei, Chatterji, and Bartlett 2022; Somepalli et al. 2022), sketching (Chen et al. 2023), and random projections (Bach 2023). These studies have shown that in the under-parameterized regime, there is a U-shaped performance curve, with models achieving benign overfitting when over-parameterized. Despite the extensive literature devoted to understanding the double descent phenomenon, there are still several open problems: 1) The lack of a general asymptotic analysis framework for generalized nonlinear regression models that covers various models. 2) Existing asymptotic results often remain as self-consistency equations that are hard to estimate, and there is a need for a connection to traditional model complexity measures, such as effective dimension and Rademacher complexity, to aid in understanding. 3) Benign overfitting can be caused by overparameterization, and subsampling may also achieve better performance from a dual view.

In this paper, we address these challenges by developing a generalized high-dimensional analysis framework and an improved algorithm for nonlinear regression models, thanks to the asymptotic equivalents involving effective dimension in (Bach 2023) and the insights from downsampling in (Chen et al. 2023). We first devise a generalized nonlinear model that covers linear regression, random features, neural networks, random projections, and sketching. Next, we establish the implicit regularization and asymptotic equivalents that are implicitly related to effective dimension in high-dimensional analysis for generalized nonlinear regression. Using these tools, we derive asymptotic risks for nonlinear ridge regression and ridgeless regression models. Motivated by the theoretical finding that the excess risk decreases with the effective dimension, we design a random feature regression model with effective dimension (RFRed) to minimize the training loss and effective dimension by jointly optimizing the feature mapping and model parameters. We conduct experiments to explore the impacts of nonlinear feature mappings and subsampling, respectively. We leave the proofs and more experiments in the appendix¹. Our contributions can be summarized as follows:

- We provide a generalized asymptotic analysis framework for general nonlinear regression models, where the limit-

*Corresponding author

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Full version: <https://lijian.ac.cn/files/2024/NonlinearHDA.pdf>
Code: <https://github.com/superlj666/NonlinearHDA>

ing risks are related to the effective dimension rather than self-consistency equations.

- Motivated by the theoretical findings, we devise a trainable nonlinear regression algorithm that minimizes the effective dimension by optimizing the feature mapping, regularization parameter, and the subsampling matrix.
- We discover interesting byproducts of the asymptotic results, such as the use of nonlinear feature mapping to reduce effective dimension and the potential benefits of subsampling for generalization.

Preliminaries

In the random design setting of linear regression, the covariates $x_1, \dots, x_n \in \mathbb{R}^d$ are sampled independently from a fixed distribution P_x such that the covariates are zero mean $\mathbb{E}(x_i) = 0$ and have a covariance matrix $\text{Cov}(x_i) = \Sigma \in \mathbb{R}^{d \times d}$. We consider the linear model $f(x; \eta) = \eta^\top x$ where $\eta \in \mathbb{R}^d$ is the parameter vector. Specifically, the response y_i is determined by $y_i = \eta_*^\top x_i + \varepsilon_i, \forall i \in [n]$, with x_i and ε_i are independent, η_* is the underlying parameter vector, $\mathbb{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. The optimal parameter vector $\eta_* \in \mathbb{R}^d$ satisfies $\mathbb{E}(y - \eta_*^\top x) = \min_{\eta} \mathbb{E}(y - \eta^\top x)$.

We denote $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ the response vector, $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$ the feature matrix, and $\varepsilon \in \mathbb{R}^n$ the noise vector. Thus, we have $y = X\eta_* + \varepsilon$. We also define $\hat{\Sigma} = \frac{1}{n}X^\top X \in \mathbb{R}^{d \times d}$ the empirical covariance matrix, of which the expected counterpart is the population covariance matrix $\mathbb{E}(\hat{\Sigma}) = \mathbb{E}(\frac{1}{n}x_i x_i^\top) = \Sigma$.

Linear Ridge Regression

The linear ridge regression aims to solve the minimization problem:

$$\hat{\eta} = \arg \min_{\eta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (\eta^\top x_i - y_i)^2 + \lambda \|\eta\|_2^2 \right\}, \quad (1)$$

which admits the closed-form solution

$$\hat{\eta} = (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} \eta_* + (\hat{\Sigma} + \lambda I)^{-1} \frac{X^\top \varepsilon}{n}. \quad (2)$$

Generalized Nonlinear Regression Model

Although the ridge linear regression has been well-studied in the high-dimensional setting (Dobriban and Wager 2018; Hastie et al. 2022), the linear models are rather simple while the modern models are usually equipped with nonlinear feature mappings. In this section, we first introduce a generalized nonlinear feature mapping for ridge regression and then present subsampling for nonlinear models to reduce the number of samples.

We consider the nonlinear model $f(x; \theta) = \theta^\top \phi(x)$ where $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ is the nonlinear feature mapping and $\theta \in \mathbb{R}^p$ is the parameter vector in the feature space \mathbb{R}^p .

Assumption 1. (Existence of θ_* in the feature space) We assume the response y_i are generated in the feature space \mathbb{R}^p after the feature mapping $\phi(x_i)$, admitting $y_i = f(x; \theta_*) + \varepsilon_i$ where $\theta_* \in \mathbb{R}^p$ is the ideal estimator in the feature space

and $\varepsilon \in \mathbb{R}^n$ is the noise vector. The label noise ε is independent of $\phi(x_i)$ and follows a distribution on \mathbb{R} such that $\mathbb{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. We also assume the norm of θ_* is bounded.

The above assumption implies $\mathbb{E}(y|x) = x^\top \theta_*$ and was widely used in the generalization analysis of kernel ridge regression (Caponnetto and De Vito 2007; Smale and Zhou 2007; Li, Liu, and Wang 2023a, 2024). Therefore, instead of (1), nonlinear ridge regression aims to solve

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|\phi(X)\theta - y\|_2^2 + \lambda \|\theta\|_2^2 \right\}, \quad (3)$$

with the closed-form solution

$$\hat{\theta} = (\hat{\Sigma}_\phi + \lambda I)^{-1} \hat{\Sigma}_\phi \theta_* + (\hat{\Sigma}_\phi + \lambda n I)^{-1} \frac{\phi(X)^\top \varepsilon}{n}, \quad (4)$$

where $\phi(X) = [\phi(x_1), \dots, \phi(x_n)]^\top \in \mathbb{R}^{n \times p}$ is the feature matrix and $\hat{\Sigma}_\phi = \frac{1}{n} \phi(X)^\top \phi(X) \in \mathbb{R}^{p \times p}$ is the covariance matrix after the nonlinear feature mappings. We consider some special cases for the nonlinear feature mapping:

- Linear method: $\phi(x) = x$.
- Random projection: $\phi(x) = Wx$, where $W \in \mathbb{R}^{p \times d}$ has sub-Gaussian components with mean zero and unit variance.
- Random Fourier features (Li, Liu, and Wang 2022): $\phi(x) = \sqrt{\frac{2}{p}} \cos(Wx + b)$, where $W = [w_1, \dots, w_p]^\top \in \mathbb{R}^{p \times d}$ are sampled from the the Fourier transform of the kernel and the bias b is uniformly sampled from $[0, 2\pi]^p$.
- Neural network with a single-hidden layer, e.g. ReLU $\phi(x) = \max\{Wx, 0\}$, and Sigmoid $\phi(x) = \frac{1}{1 + \exp^{-Wx}}$.
- Deep neural networks: $\phi(x) = \phi_L(\phi_{L-1}(\dots \phi_1(x)))$, where L is the depth of the network and the feature mappings ϕ_1, \dots, ϕ_L may be different.

Generalized Nonlinear Regression Model with Subsampling

We consider the subsampling methods for nonlinear ridge regression with the subsampling feature matrix $S\phi(X) \in \mathbb{R}^{m \times p}$ where $S \in \mathbb{R}^{m \times n}$ is subsampling matrix. Note that the regression problem are based on subsampled examples

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{m} \|S\phi(X)\theta - Sy\|_2^2 + \lambda \|\theta\|_2^2 \right\}, \quad (5)$$

where the closed form solution is

$$\hat{\theta} = (\hat{\Sigma}_{S\phi} + \lambda I)^{-1} \hat{\Sigma}_{S\phi} \theta_* + (\hat{\Sigma}_{S\phi} + \lambda I)^{-1} \frac{\phi(X)^\top S^\top S \varepsilon}{m}, \quad (6)$$

where $\hat{\Sigma}_{S\phi} = \frac{1}{m} \phi(X)^\top S^\top S \phi(X) \in \mathbb{R}^{p \times p}$. There are some special cases for subsampling:

- Full sampling: $S = I_n$ and $m = n$, such that $S\phi(X) = \phi(X)$.
- Subset selection: only one 1 and other zeros in each row of S . For example, $[0, 1, 0, \dots, 0]$ represents to subsample the second example x_2 .

- Sketching: S is the sketching matrix, e.g. Gaussian sketching requires that the sketching matrix S is generated from the Gaussian distribution.

When $S = I_n$ and $\phi(X) = X$, we can recover the traditional linear ridge regression (2) from (6).

Asymptotics for Generalized Nonlinear Regression

Throughout this paper, we study the behaviors of the out-of-sample excess risk in the proportional asymptotic limit where the sample size n , the dimension d , the dimension of the feature mapping p , and the subsampling size m go to infinity, i.e. $n, d, p, m \rightarrow \infty$, in such a way that $p/m \rightarrow \gamma \in (0, \infty)$. We call $\gamma < 1$ the under-parameterized model and $\gamma > 1$ the over-parameterized model.

Assumptions

We use the notations in random features (Rudi and Rosasco 2017) to obtain the population covariance matrix.

Assumption 2 (Continuous and bounded feature mapping). *Assume the feature mapping $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ be continuous in terms of both the input and hyperparameters in the mapping and bounded.*

Note that the assumption above is satisfied when the activation function is continuous and bounded, for example, random Fourier features (RFF) $\phi(x) = \sqrt{\frac{2}{p}} \cos(Wx + b)$ provides continuous and bounded feature mapping.

Proposition 1. *Under Assumption 2, the inputs in X are sampled i.i.d. from P_x , and then the empirical covariance matrix of the feature mapping $\widehat{\Sigma}_\phi = \frac{1}{p}\phi(X)^\top\phi(X)$ converges to a deterministic covariance matrix $\Sigma_\phi = \mathbb{E}[\phi(x)\phi(x)^\top]$ when $p \rightarrow \infty$.*

Using the operatorial definitions in random features, we define the above population covariance matrix $\Sigma_\phi = \mathbb{E}[\phi(x)\phi(x)^\top] \in \mathbb{R}^{p \times p}$, which also apply to neural network. Using the central limit theorem, the empirical covariance matrix converges to the population covariance matrix $\widehat{\Sigma} = \frac{1}{p}\phi(X)^\top\phi(X) = \frac{1}{p}\sum_{i=1}^p \phi(x_i)\phi(x_i)^\top \rightarrow \mathbb{E}[\phi(x)\phi(x)^\top] = \Sigma_\phi$ when $p \rightarrow \infty$.

We modify high-dimensional assumptions in linear regression (Dobriban and Wager 2018; Hastie et al. 2022; Bach 2023) to nonlinear regression.

Assumption 3 (Covariance condition for nonlinear feature mapping). *Suppose Σ_ϕ is invertible and bounded, and the eigenvalues of Σ_ϕ are positive and bounded. $\phi(X) = Z\Sigma_\phi^{1/2}$ where Z has i.i.d. entries with zero mean, and unit variance.*

The above assumption specifies the covariance structure for feature mappings, where Z can be standard Gaussian components or Rademacher random variables $P(z = -1) = P(z = 1) = 1/2$.

Assumption 4 (Orthogonal subsampling matrix). *Suppose the rows of subsampling matrix is orthogonal, such that $SS^\top = I_m$. Meanwhile, $S^\top S$ converges to a deterministic matrix Σ_S .*

Remark 1. *The above assumption is relatively strict that cannot be satisfied by i.i.d. sketching matrices. However, this assumption holds for orthogonal sketching matrix and subset selection since the subsampling matrix S is fixed for the subset selection without replacement.*

Using the above assumptions, we can prove that the subsampled nonlinear feature mappings have a deterministic covariance and make the following assumption.

Assumption 5 (Covariance condition for subsampled nonlinear models). *The empirical covariance matrix of $\widehat{\Sigma}_{S\phi} = \frac{1}{m}\phi(X)^\top S^\top S\phi(X)$ converges to a deterministic covariance matrix $\Sigma_{S\phi} = \Sigma_\phi^{1/2} Z^\top \Sigma_S Z \Sigma_\phi^{1/2}$. The spectral distribution $F_{\Sigma_{S\phi}}$ of $\Sigma_{S\phi}$ converges to a limit probability distribution μ supported on $[0, +\infty)$ and Σ is invertible and bounded in operator norm.*

The above assumption implies there is no vanishing eigenvalues in the limiting μ .

Implicit Regularization and Asymptotic Equivalents

For any measure G on $[0, \infty)$, we define the Stieltjes transform by $m_G(z) = \int \frac{1}{l-z} dG(l)$, where $z \in \mathbb{C} \setminus \mathbb{R}^+$. Meanwhile, the companion Stieltjes transform v_G is defined by $v_G(z) + 1/z = \gamma(m_G(z) + 1/z)$.

The Stieltjes transform $v(z)$ is the limit of the Stieltjes transform of the spectral measure of the kernel matrix $\widehat{v}(z) = \frac{1}{m} \text{tr} \left[(\widehat{\mathbf{K}}_{S\phi} - zI)^{-1} \right]$, where $\widehat{\mathbf{K}}_{S\phi} = \frac{1}{m} S\phi(X)\phi(X)^\top S^\top \in \mathbb{R}^{m \times m}$. More examples about Stieltjes transform refer to (Dobriban and Wager 2018; Dobriban and Liu 2019; Hastie et al. 2022).

There is an unique positive solution for $v(z)$ in the self-consistency equation (Bai and Silverstein 2010; Jacot et al. 2020; Bach 2023):

$$m + mzv(z) = p \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{1/v(z) + \sigma}, \quad (7)$$

where μ is the limit probability distribution of the spectral measure of $\Sigma_{S\phi}$. However, it's hard to describe the limiting density for general $\Sigma_{S\phi}$.

Setting $z = -\lambda$ for any $\lambda > 0$, we have $\widehat{v}(-\lambda) = \frac{1}{m} \text{tr}[(\widehat{\mathbf{K}}_{S\phi} + \lambda I)^{-1}]$, which converges to $v(-\lambda)$ almost surely. From Section A.1 in (Bach 2023), using the self-consistency equation (7), there holds the asymptotic equivalence for the effective dimension (also called degree of freedom) for $\lambda > 0$, $\widehat{df}_1(\lambda) = \text{tr}(\widehat{\Sigma}_{S\phi}(\widehat{\Sigma}_{S\phi} + \lambda I)^{-1}) \sim df_1(\kappa) = \text{tr}(\Sigma_{S\phi}(\Sigma_{S\phi} + \kappa I)^{-1})$, where $\kappa = \frac{1}{v(-\lambda)}$ is the implicit regularization parameter. Here, we denote $a \sim b$ as the asymptotic equivalence, such that the ratio a/b tends to one when $n, d, p, m \rightarrow +\infty$. Note that, when $\lambda = 0$, κ may not be zero and leads to implicit regularization in the over-parameterized regime $\gamma > 1$.

Using the definition of μ in Assumption 5, we have $\frac{1}{p}df_1(\kappa) = \frac{1}{p} \sum_{i=1}^p \frac{\sigma_i}{\sigma_i + \kappa} \rightarrow \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{\sigma + \kappa}$, which is decreasing in κ and converges to one when $\kappa = 0$, while

$\text{df}_1(0) = \text{rank}(\Sigma_{S\phi})$. Therefore, we have $\text{df}_1\left(\frac{1}{v(z)}\right) \rightarrow p \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{1/v(z)+\sigma}$ when $\kappa = \frac{1}{v(z)}$. Then, using $z = -\lambda$ and $\kappa = \frac{1}{v(-\lambda)}$, we rewrite (7) as

$$\lambda \sim \kappa \left(1 - \frac{1}{m} \text{df}_1(\kappa)\right). \quad (8)$$

Note that, the implicit regularization parameter $\kappa = \frac{1}{v(-\lambda)} \in \mathbb{R}_+$, which is the limit of $1/\text{tr}[(S\phi(X)\phi(X)^\top S^\top + m\lambda I)^{-1}]$. Since $\text{df}_1(\kappa)$ is decreasing in κ , we know that λ and γ is positive correlated.

Remark 2 (Implicit regularization). *We consider the ridgeless settings when $\lambda \rightarrow 0$ in terms of γ . 1) Underparameterization ($\gamma < 1$): Since $\text{df}_1(\kappa) \leq p$ and $(1 - \frac{1}{m} \text{df}_1(\kappa)) > 0$ from (8), we have $\kappa = 0$ when $\lambda = 0$. Meanwhile, $\Sigma_{S\phi}$ is invertible from Assumption 5, such that $\text{df}_1(\kappa) = p$ and $\lambda \sim \kappa(1 - \gamma)$ when λ goes to zero. 2) Overparameterization ($\gamma > 1$): If $\kappa = 0$ when $\lambda = 0$, from (8) we have $\lambda \sim \kappa(1 - \gamma) < 0$, violating the fact $\lambda > 0$. Therefore, when $\lambda \rightarrow 0_+$, we have $\kappa > 0$ and $\text{df}_1(\kappa) \rightarrow m_+$. There is an implicit regularization parameter $\kappa > 0$ with $\text{df}_1(\kappa) = m$ for the ridgeless regression $\lambda = 0$.*

Bach (Bach 2023) provided asymptotic equivalents for spectral functions of the empirical covariance operator, which established the relations between spectral functions and expected effective dimensions. Following the asymptotic equivalents in (Bach 2023), we provide asymptotic equivalents for spectral functions of nonlinear regression models.

Proposition 2. *Under Assumptions 2 - 5, the following asymptotic equivalents holds:*

$$\text{tr}[\Sigma_{S\phi}(\widehat{\Sigma}_{S\phi} + \lambda I)^{-1}] \sim \frac{\kappa}{\lambda} \text{df}_1(\kappa), \quad (9)$$

$$\text{tr}[\Sigma_{S\phi}(\widehat{\Sigma}_{S\phi} + \lambda I)^{-2}] \quad (10)$$

$$\sim \frac{\kappa^2}{\lambda^2} \text{tr}[\Sigma_{S\phi}(\Sigma_{S\phi} + \kappa I)^{-2}] \cdot \frac{m}{m - \text{df}_2(\kappa)},$$

$$\theta_*^\top (\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \Sigma_{S\phi} (\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \theta_* \quad (11)$$

$$\sim \frac{\kappa^2}{\lambda^2} \theta_*^\top (\Sigma_{S\phi} + \kappa I)^{-2} \Sigma_{S\phi} \theta_* \cdot \frac{m}{m - \text{df}_2(\kappa)}.$$

where $\text{df}_1(\kappa) = \text{tr}(\Sigma_{S\phi}(\Sigma_{S\phi} + \kappa I)^{-1})$ and $\text{df}_2(\kappa) = \text{tr}(\Sigma_{S\phi}^2(\Sigma_{S\phi} + \kappa I)^{-2})$.

Asymptotic Analysis of Ridge Regression

Let M be a self-adjoint positive semidefinite matrix and the vector norm $\|v\|_M^2 := v^\top M v$. In the fixed design setting, the covariates x_1, \dots, x_n are assumed deterministic and the expected excess risk measures the "in-sample" error $\mathbb{E}_\varepsilon [\|\hat{\theta} - \theta_*\|_{\widehat{\Sigma}_{S\phi}}^2]$ defined by $\mathbb{E}_\varepsilon \left[\frac{1}{m} \|S\phi(X)(\hat{\theta} - \theta_*)\|_2^2 \right]$. In contrast, the random design setting assumed the covariates to be sampled i.i.d. with the covariance matrix $\Sigma_{S\phi}$. We can obtain the excess risk for the random design setting with $\widehat{\Sigma}_{S\phi}$ replaced by $\Sigma_{S\phi}$, i.e. $\mathbb{E}_\varepsilon [\|\hat{\theta} - \theta_*\|_{\Sigma_{S\phi}}^2]$, to measure

the "out-of-sample" error. We only provide main results in this section, while leaving the proofs and comparison with related work in the appendix.

We recover the bias-variance decomposition for the nonlinear regression with subsampling (6).

Lemma 1 (Bias-variance decomposition). *Under Assumptions 4, the excess risk of the nonlinear ridge regression with subsampling (6) exhibits the following bias-variance decomposition*

$$\mathbb{E}_\varepsilon \left[\left\| \hat{\theta} - \theta_* \right\|_{\Sigma_{S\phi}}^2 \right] = \underbrace{\mathbb{E}_\varepsilon \left[\left\| \hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta}) \right\|_{\Sigma_{S\phi}}^2 \right]}_{\text{Variance}} + \underbrace{\left\| \mathbb{E}_\varepsilon(\hat{\theta}) - \theta_* \right\|_{\Sigma_{S\phi}}^2}_{(\text{Bias})^2},$$

where

$$\begin{aligned} \text{Variance} &= \frac{\sigma^2}{m} \text{tr} \left[(\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \widehat{\Sigma}_{S\phi} (\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \Sigma_{S\phi} \right], \\ (\text{Bias})^2 &= \lambda^2 \theta_*^\top (\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \Sigma_{S\phi} (\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \theta_*. \end{aligned}$$

Note that, in the proof of the variance term, there is a sketched covariance $\frac{\phi(X)^\top S^\top SS^\top S\phi(X)}{m}$, which is difficult to estimate by effective dimension. Dobriban et al. (Dobriban and Liu 2019) utilized the orthogonal invariance of Gaussian matrices and properties of Wishart matrices to provide asymptotic limit for the i.i.d. sketched covariance matrix. However, these proof techniques only applied to the under-parameterized regime $n \geq p$ and ignored the over-parameterized regime. Recent work used random matrix theory tools for estimating the limiting variance and the results are with self-consistent equations, which is hard to be estimated and related to the effective dimension.

Theorem 1 (Asymptotic risk for ridge regression). *Under Assumptions 2 - 5, the nonlinear ridge regression with subsampling estimator in (6) admits the following limiting variance and bias:*

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\left\| \hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta}) \right\|_{\Sigma_{S\phi}}^2 \right] &\sim \sigma^2 \frac{\text{df}_2(\kappa)}{m - \text{df}_2(\kappa)}, \\ \left\| \mathbb{E}_\varepsilon(\hat{\theta}) - \theta_* \right\|_{\Sigma_{S\phi}}^2 &\sim \frac{m\kappa^2 \theta_*^\top (\Sigma_{S\phi} + \kappa I)^{-2} \Sigma_{S\phi} \theta_*}{m - \text{df}_2(\kappa)}. \end{aligned} \quad (12)$$

From (12), we find that both the variance and bias terms are increasing for larger $\text{df}_2(\kappa)$ and the excess risk explodes when $\text{df}_2(\kappa) \rightarrow m$.

Remark 3. *The value of $\text{df}_2(\kappa)$ is influenced by three factors: 1) The explicit regularization parameter λ : According to the self-consistency equation in (8), κ is positively correlated with λ . In (Bach 2023), it is suggested that choosing an appropriate λ can ensure $\text{df}_2(\kappa) \ll m$ to prevent risk explosion. 2) The feature mapping: By considering the covariance on the feature mapping and subsampling of the inputs $S\phi(X)$ instead of the primal inputs X , the value of $\text{df}_2(\kappa)$ still depends on the choice of the feature mapping. Using a suitable feature mapping with hyperparameters or trainable feature mapping can further decrease $\text{df}_2(\kappa)$ and improve generalization performance. This explains why ridge regression models employing nonlinear feature mappings such as*

kernel ridge regression (KRR), random features, and neural networks outperform linear regression. 3) The subsampling matrix: The value of $\text{df}_2(\kappa)$ is also affected by the subsampling matrix. By employing an appropriate subsampling matrix, even through downsampling, it is possible to reduce $\text{df}_2(\kappa)$.

Asymptotic Analysis of Ridgeless Regression

We consider the limit when $\lambda = 0$ where the ridge regression estimator (6) becomes a minimum ℓ_2 -norm (ridgeless) estimator $\hat{\theta} = \theta_* + \frac{\widehat{\Sigma}_{S\phi}^{-1} \phi(X)^\top S^\top S \varepsilon}{m}$.

We first consider the under-parameterized regime $\gamma < 1$ where $\kappa = 0$ when $\lambda = 0$ as discussed in Remark 2. Since $\widehat{\Sigma}_{S\phi}$ is invertible, we have $\text{df}_2(\kappa) = \text{rank}(\widehat{\Sigma}_{S\phi}) = p$. Substituting $\kappa = 0$ and $\text{df}_2(\kappa) = p$ to (12), we obtain the following results.

Corollary 1 (Under-parameterized regime). *Under Assumptions 2 - 5, if $\lambda = 0$ and $\gamma < 1$, the nonlinear ridgeless regression with subsampling estimator in (6) admits the following limiting variance and bias:*

$$\mathbb{E}_\varepsilon \left[\left\| \hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta}) \right\|_{\Sigma_{S\phi}}^2 \right] \sim \sigma^2 \frac{p}{m-p}, \quad \left\| \mathbb{E}_\varepsilon(\hat{\theta}) - \theta_* \right\|_{\Sigma_{S\phi}}^2 = 0.$$

In above results, we recover the classical results in Theorem 1 of (Hastie et al. 2022) for underparameterization. When $\gamma \rightarrow 1$, i.e. $p \rightarrow m$, the variance term explodes. Note that, since we assume $\theta_* \in \mathbb{R}^p$ in Assumption 1 and responses are generated by $\theta_*^\top \phi(x)$ in the feature space, the variance term is zero and the risk is increasing in p , i.e. there is no U-shape excess risk in the under-parameterized regime. However, as shown in Proposition 4 of (Jacot et al. 2020; Bach 2023), if we assume $\theta_* \in \mathbb{R}^d$ in the input space and the feature mapping $\phi(x) = W^\top x$ where $W \in \mathbb{R}^{d \times p}$, we can obtain a nonzero bias term that is decreasing in p and observe an U-shape excess risk.

We then consider the over-parameterized regime $\gamma > 1$ where κ is defined by $\text{df}_1(\kappa) = m$ from Remark 2 and $\text{df}_2(\kappa) \leq \text{df}_1(\kappa) = m$.

Corollary 2 (Over-parameterized regime). *Under Assumptions 2 - 5, if $\lambda = 0$ and $\gamma > 1$, with κ_0 defined by $\text{df}_1(\kappa_0) = m$ the nonlinear ridgeless regression with subsampling estimator in (6) admits the following limiting variance and bias:*

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\left\| \hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta}) \right\|_{\Sigma_{S\phi}}^2 \right] &\sim \sigma^2 \frac{\text{df}_2(\kappa_0)}{m - \text{df}_2(\kappa_0)}, \\ \left\| \mathbb{E}_\varepsilon(\hat{\theta}) - \theta_* \right\|_{\Sigma_{S\phi}}^2 &= \frac{m \kappa^2 \theta_*^\top (\Sigma_{S\phi} + \kappa I)^{-2} \Sigma_{S\phi} \theta_*}{m - \text{df}_2(\kappa_0)}. \end{aligned} \quad (13)$$

The excess risk in over-parameterized regime depends on the differences between two effective dimensions $\text{df}_1(\kappa_0) - \text{df}_2(\kappa_0)$. If $\text{df}_2(\kappa_0) \ll \text{df}_1(\kappa_0)$, the variance term tends to zero and the bias term tends to $\kappa^2 \theta_*^\top (\Sigma_{S\phi} + \kappa I)^{-2} \Sigma_{S\phi} \theta_*$ when $m \rightarrow \infty$. If $\text{df}_2(\kappa_0) = \text{rmdf}_1(\kappa_0)$, there is catastrophic overfitting where both the variance and bias terms explode. In other situations, these two effective dimensions are constants away from each other, and there is no catastrophic overfitting but the variance term is a constant.

Since $\kappa_0 \sim \frac{1}{\text{tr}((S\phi(X)\phi(X)^\top S^\top)^{-1})}$ and $\text{df}_2(\kappa_0)$ is decreasing in κ_0 , we can optimize $S\phi(X)$ to increase κ_0 and reduce $\text{df}_2(\kappa_0)$ at the same time.

Trainable Nonlinear Regression Model

From Theorem 1, we notice that both the limiting variance and bias are increasing in $\text{df}_2(\kappa)$, and thus we lower the excess risk by reducing the effective dimension $\text{df}_2(\kappa)$. However, $\text{df}_2(\kappa)$ is hard to compute and thus we use the empirical effective dimension $\widehat{\text{df}}_2(\lambda)$ instead. That coincides with reducing $\widehat{\text{df}}_2(\lambda)$ can improve the fixed design risk $\mathbb{E}_\varepsilon[\|\hat{\theta} - \theta_*\|_{\widehat{\Sigma}_{S\phi}}^2] = \lambda^2 \text{tr}[\theta_* \theta^\top (\widehat{\Sigma}_{S\phi} + \lambda I)^{-1} \widehat{\Sigma}_{S\phi}] + \frac{\sigma^2}{n} \widehat{\text{df}}_2(\lambda)$. Therefore, smaller $\widehat{\text{df}}_2(\lambda)$ can lead to better performance. To obtain the smallest $\widehat{\text{df}}_2(\lambda)$, we devise a generalized nonlinear regression model with a bi-level problem

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{n} \|S\phi(X)\theta - Sy\|_2^2 + \lambda \|\theta\|_2^2 \\ \text{s.t.} \quad & \{\lambda, \phi, S\} = \arg \min_{\lambda, \phi, S} \widehat{\text{df}}_2(\lambda). \end{aligned} \quad (14)$$

The value of $\text{df}_2(\kappa)$ depends on λ , the feature mapping ϕ , and the subsampling matrix S , as discussed in Remark 3. To solve (14), we alternate between optimizing the model parameter θ and the hyperparameters λ, ϕ, S : 1) With fixed hyperparameters λ, ϕ, S , the algorithm trains the nonlinear model θ . 2) With a fixed nonlinear estimator θ , the algorithm optimizes the hyperparameters λ, ϕ, S .

To compute $\widehat{\text{df}}_2(\lambda)$, the computational complexity is typically impractical for the over-parameterized regime ($p > m$) at $\mathcal{O}(p^3 + mp^2)$ time. For the under-parameterized regime, an alternative form is provided as

$$\widehat{\text{df}}_2(\lambda) = \text{tr}(\widetilde{X}^\top (\widetilde{X}\widetilde{X}^\top + \lambda m I)^{-1} \widetilde{X}^\top (\widetilde{X}\widetilde{X}^\top + \lambda m I)^{-1} \widetilde{X}), \quad (15)$$

where $\widetilde{X} = S\phi(X)$. The time complexity for the above form is $\mathcal{O}(p^2 n)$ since $p > n$. To accelerate the computational efficiency, we only compute $\widehat{\text{df}}_2(\lambda)$ for every α -iterations rather than each iteration, where $\alpha \in \mathbb{N}_+$.

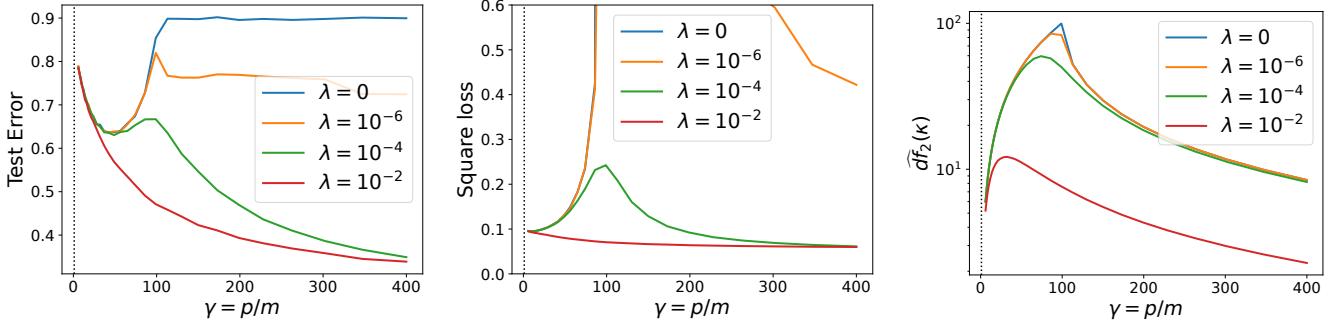
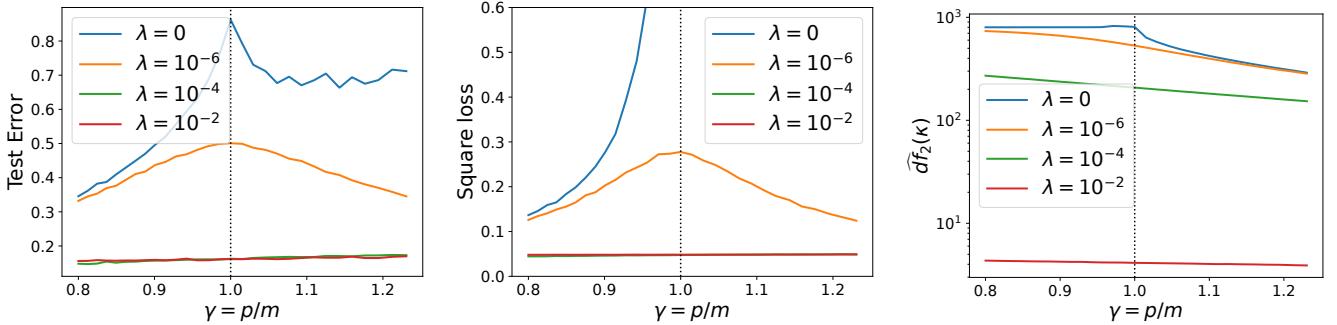
Example: Random Feature Regression with Effective Dimension (RFRed)

From (14), there are too many hyperparameters to optimize and the computation of (15) is still very time-consuming. Based on random Fourier features (Rahimi and Recht 2007), we devise Random Feature Regression model with Effective Dimension (RFRed) to solve (14) with the feature mapping

$$\phi(x) = \sqrt{\frac{2}{p}} \cos(W^\top x + b), \quad (16)$$

where the frequency matrix $W = [w_1, \dots, w_p] \in \mathbb{R}^{d \times p}$ composed of p vectors drawn i.i.d. from a Gaussian distribution $\mathcal{N}(0, \frac{1}{\sigma^2} \mathbf{I}) \in \mathbb{R}^d$. The phase vectors $b = [b_1, \dots, b_p] \in \mathbb{R}^p$ are drawn uniformly from $[0, 2\pi]$.

To improve the computational efficiency, we tune hyperparameters λ, S before the training and optimize the feature

Figure 1: Testing error, testing loss, and effective dimension $\widehat{df}_2(\widehat{\kappa})$ versus the increase of the feature dimension p .Figure 2: Testing error, testing loss, and effective dimension $\widehat{df}_2(\widehat{\kappa})$ versus the increase of the subsampling size m .

mapping ϕ during the training. To accelerate the solve of (14), we optimize θ and W jointly by minimize the the following objective

$$\mathcal{L}(\theta; W) = \frac{1}{n} \|S\phi(X)\theta - Sy\|_2^2 + \lambda\|\theta\|_2^2 + \beta \widehat{df}_2(\lambda), \quad (17)$$

where $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ is defined in (16), $\widehat{df}_2(\lambda)$ is defined in (15), $\lambda = 0$ for ridgeless regression, and β is a hyperparameter to balance the effect between squared loss and the effective dimension.

Complexity. Using batch stochastic gradient method, we have $\nabla_\theta \mathcal{L} = \frac{1}{n} \tilde{X}_b^\top (\tilde{X}_b \theta - \tilde{y}_b)$ where $\{\tilde{X}_b \in \mathbb{R}^{b \times p}, \tilde{y}_b \in \mathbb{R}^b\}$ is a batch of $\{S\phi(X), Sy\}$ with the batch size b . We also use the batch data to approximate $\widehat{df}_2(\lambda)$ where \tilde{X} in (15) is replaced by \tilde{X}_b . The compute of $S\phi(X)$ consumes $\mathcal{O}(mnp + ndp)$. With T iterations, the update of θ takes $\mathcal{O}(pbT)$ time, the update of W consumes $\mathcal{O}(pb^2T)$, and the compute of $\widehat{df}_2(\lambda)$ requires $\mathcal{O}(\frac{p^2 n T}{n\alpha})$.

Experiments

We utilize the random Fourier feature, as defined in equation (16), to provide an approximation of the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\sigma^2 \|\mathbf{x} - \mathbf{x}'\|^2 / 2)$. It is important to note that the random Fourier features, specified in equation (16), are associated with the frequency matrix $W \sim \mathcal{N}(0, \sigma^2)$. Our implementation is based on PyTorch, and we fine-tune

the hyperparameters through a grid search approach, exploring values for σ^2 in the range of $\{0.01, \dots, 1000\}$ and $\lambda \in \{0.1, \dots, 10^{-5}\}$. We leave more experiments in the appendix, including the impact of the trainable feature mapping and the comparison experiments.

Impact of the Dimension of Nonlinear Regression

We fix $n = 100, S = I_n, m = n$ and change the random features dimension $p \in [10, 400]$. The training examples $n = 100$ are randomly drawn from the MNIST dataset (LeCun et al. 1998). We set the same hyperparameter $\sigma^2 = 0.1$. We estimate the implicit regularization parameter $\kappa \sim \widehat{\kappa} = 1/\text{tr}[(\phi(X)\phi(X)^\top + \lambda n I)^{-1}]$ and use $\widehat{df}_2(\widehat{\kappa})$ to approximate the key quantity $df_2(\kappa)$. The results are reported in Figure 1, which illustrated: 1) Mild regularization ($\lambda = 10^{-2}$ and $\lambda = 10^{-4}$) exhibits double descent phenomena, while stronger regularization restricts $df_2(\kappa)$ and eliminates double descent. In the underparameterized setting ($\gamma < 1$), test errors initially decrease due to reduced bias, but increase later as variance dominates. When $\gamma = 1$, the excess risk explodes due to a small $m - df_2(\kappa)$ (Theorem 1). In the overparameterized setting ($\gamma > 1$), test errors decrease again as $df_2(\kappa)$ decreases. 2) Ridgeless regression experiences exploding loss in the overparameterized regime, while ridge regression losses decrease with p due to numerical issues with matrix inversion when $\text{rank}(\widehat{\Sigma}_{S\phi}) \ll p$. 3) Increasing p leads to a larger $\widehat{\kappa}$ approaching λ , and $\widehat{\kappa}_0$ is

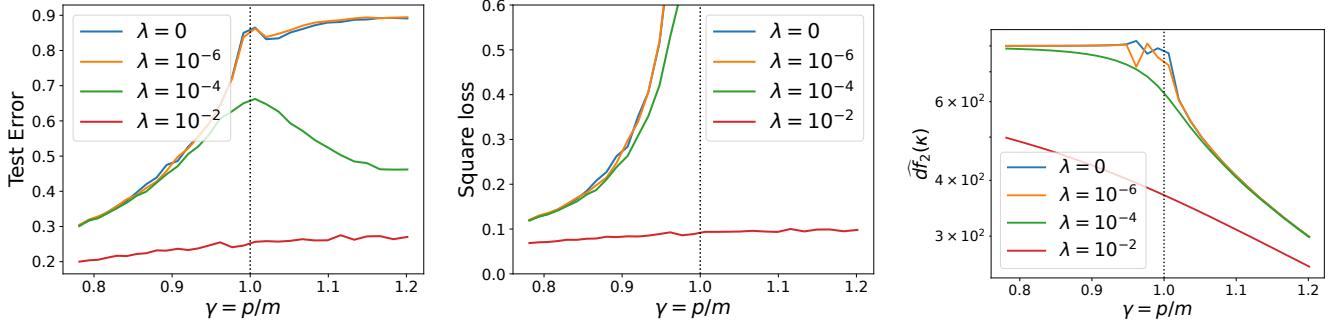


Figure 3: The testing error, testing loss, and effective dimension $\hat{df}_2(\kappa)$ versus the increase of the ROS size.

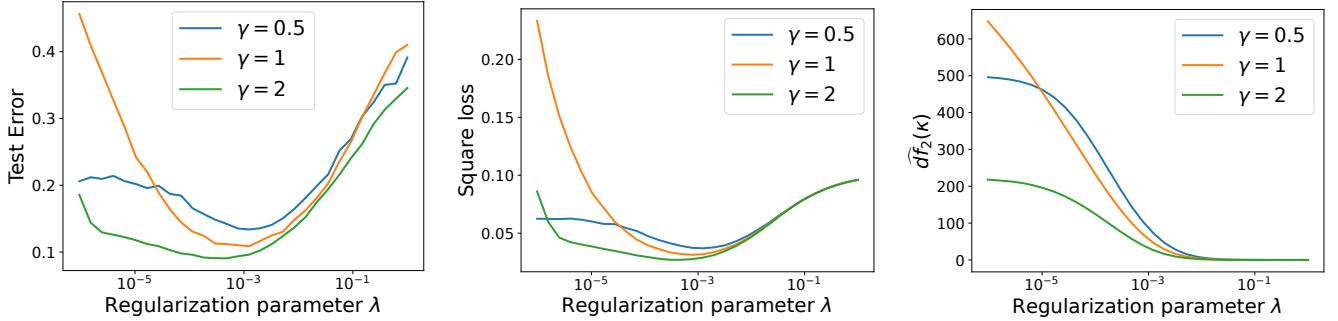


Figure 4: Testing error, testing loss, and effective dimension $\hat{df}_2(\kappa)$ w.r.t. different values of λ .

similar to $\hat{\kappa}$ with smaller λ when $\lambda = 0$. Smaller $df_2(\kappa)$ and appropriate regularization yield better generalization.

Impact of the Subsampling Size

We fix $m = 1000, p = 800$ and vary the subsampling size $m \in [640, 1000]$. We directly use subset selection matrix in this experiment. Using the same hyperparameters and performance indicators, we report the results in Figure 2. We find that: 1) Downsampling does not always hurt the generalization ability, for example, the test error and square loss of ridge regression with $\lambda = 10^{-16}$ decreases in the over-parameterized regime $\gamma > 1$ where $m < p < n$. 2) The square loss and test error of ridgeless regression explodes after $\gamma > 1$. Meanwhile, nonlinear models with larger regularization parameters lose generalization ability slowly, such that one can improve efficiency by sacrificing some accuracies. 3) Subsampling size has little influence on the implicit regularization. Although the effective dimension $\hat{df}_2(\lambda)$ decrease when $\gamma > 1$ for $\lambda = 10^{-2}, \lambda = 10^{-4}$, there is no decreasing errors since the decreasing m offsets the benefits from smaller effective dimension.

Impact of the ROS Sketches

Here, we use the orthogonal sketch matrices, e.g. randomized orthonormal system (ROS) sketches (Pilanci and Wainwright 2015; Yang et al. 2017). Under same settings as above experiments, we fix $n = 1024, p = 800$ and use different sketch size $m \in [666, 1024]$. As shown in Figure 3, we find

that: 1) Downsampling sketching $m < n$ may also benefits the generalization performance, i.e. the case $\lambda = 10^{-4}$, where the test errors increases first but drops again after $p > m$. 2) With the setting $\lambda = 10^{-4}$, the test accuracies coincides with our theoretical findings in Theorem 1, where the test error is highest when $m = p$ due to both variance explosion and bias explosion. 3) Strong regularization, for example $\lambda = 10^{-2}$, leads to better performance compared to milder regularization terms. 4) As shown in the first and last figures, ridge regression estimators with lower test errors correspond to smaller $\hat{df}_2(\lambda)$.

Impact of the Different Regularization Parameter

Under same settings as above experiments, we fix $n = 1000$ and vary the regularization parameter $\lambda \in [10^{-6}, 1]$ and compare the performance in different settings, i.e. $\gamma = 0.5$, $\gamma = 1$, and $\gamma = 2$. We report the results in Figure 4, which illustrates: 1) the optimal regularization parameters λ are similar even in different settings $\gamma = 0.5, 1, 2$, i.e. near $\lambda = 10^{-3}$. 2) The implicit regularization parameter κ mainly depends on λ rather than different γ . 3) Test error, squared loss, and the effective dimension $\hat{df}_2(\kappa)$ are positive correlated. 4) When the regularization parameter is small, the threshold $\gamma = 1$ have highest test error, while the under-parameterized estimator $\gamma = 0.5$ performs worse than the others when the regularization parameter is near optimal.

Acknowledgments

The work of Jian Li is supported partially by National Natural Science Foundation of China (No. 62106257), and Project funded by China Postdoctoral Science Foundation (No. 2023T160680). The work of Yong Liu is supported partially by National Natural Science Foundation of China (No.62076234), Beijing Outstanding Young Scientist Program (No.BJJWZYJH012019100020098), the Unicorn Innovation Ecological Cooperation Plan, and the CCF-Huawei Populus Grove Fund.

References

- Ba, J.; Erdogdu, M.; Suzuki, T.; Wu, D.; and Zhang, T. 2020. Generalization of two-layer neural networks: An asymptotic viewpoint. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.
- Bach, F. 2023. High-dimensional analysis of double descent for linear regression with random projections. *arXiv preprint arXiv:2303.01372*.
- Bai, Z.; and Silverstein, J. W. 2010. *Spectral analysis of large dimensional random matrices*, volume 20. Springer.
- Bartlett, P. L.; Long, P. M.; Lugosi, G.; and Tsigler, A. 2020. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences (PNAS)*, 117(48): 30063–30070.
- Belkin, M.; Hsu, D.; and Xu, J. 2020. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4): 1167–1180.
- Belkin, M.; Ma, S.; and Mandal, S. 2018. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, 540–548.
- Caponnetto, A.; and De Vito, E. 2007. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368.
- Chen, X.; Zeng, Y.; Yang, S.; and Sun, Q. 2023. Sketched Ridgeless Linear Regression: The Role of Downampling. *arXiv preprint arXiv:2302.01088*.
- Dobriban, E.; and Liu, S. 2019. Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 3670–3680.
- Dobriban, E.; and Wager, S. 2018. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1): 247–279.
- Frei, S.; Chatterji, N. S.; and Bartlett, P. 2022. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT)*, 2668–2703. PMLR.
- Hastie, T.; Montanari, A.; Rosset, S.; and Tibshirani, R. J. 2022. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2): 949–986.
- Jacot, A.; Simsek, B.; Spadaro, F.; Hongler, C.; and Gabriel, F. 2020. Implicit regularization of random feature models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 4631–4640. PMLR.
- LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.
- Li, J.; Liu, Y.; and Wang, W. 2022. Convolutional spectral kernel learning with generalization guarantees. *Artificial Intelligence (AI)*, 313: 103803.
- Li, J.; Liu, Y.; and Wang, W. 2023a. Optimal Convergence Rates for Agnostic Nyström Kernel Learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
- Li, J.; Liu, Y.; and Wang, W. 2023b. Optimal Convergence Rates for Distributed Nyström Approximation. *Journal of Machine Learning Research (JMLR)*, 24: 141.
- Li, J.; Liu, Y.; and Wang, W. 2024. Optimal Rates for Agnostic Distributed Learning. *IEEE Transactions on Information Theory (TIT)*.
- Li, J.; Liu, Y.; and Zhang, Y. 2022. Ridgeless Regression with Random Features. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 3208–3214.
- Liang, T.; and Rakhlin, A. 2020. Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *The 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. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, volume 33, 13939–13950.
- Mei, S.; and Montanari, A. 2022. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4): 667–766.
- Nakkiran, P.; Kaplun, G.; Bansal, Y.; Yang, T.; Barak, B.; and Sutskever, I. 2021. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12): 124003.
- Pilanci, M.; and Wainwright, M. J. 2015. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory (TIT)*, 61(9): 5096–5115.
- Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 1177–1184.
- Rudi, A.; and Rosasco, L. 2017. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 3215–3225.
- Smale, S.; and Zhou, D.-X. 2007. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2): 153–172.
- Somepalli, G.; Fowl, L.; Bansal, A.; Yeh-Chiang, P.; Dar, Y.; Baraniuk, R.; Goldblum, M.; and Goldstein, T. 2022. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13699–13708.

Vapnik, V. 1999. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Yang, Y.; Pilanci, M.; Wainwright, M. J.; et al. 2017. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3): 991–1023.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.