High-dimensional analysis for Generalized Nonlinear Regression: From Asymptotics to Algorithm

Jian Li[1], **Yong Liu**[2*], **Weiping Wang**[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]Gaoling School of Artificial Intelligence, Renmin University of Chinas
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 appendix[1]. Our contributions can be summarized as follows:

- We provide a generalized asymptotic analysis framework for general nonlinear regression models, where the limiting risks are related to the effective dimension rather than

*Corresponding author

[1]https://lijian.ac.cn/files/2024/NonlinearHDA.pdf

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 $x_1, \cdots, x_n \in \mathbb{R}^d$ 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 $\mathrm{Cov}(\mathrm{x_i}) = \Sigma \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$. 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 $\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$. 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 $y = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ the response vector, $X = (x_1, \cdots, x_n)^\top \in \mathbb{R}^{n\times d}$ the feature matrix, and $\varepsilon \in \mathbb{R}^n$ the noise vector. Thus, we have $y = X\eta_* + \varepsilon$. We also define $\widehat{\Sigma} = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$ the empirical covariance matrix, of which the expected counterpart is the population covariance matrix $\mathbb{E}(\widehat{\Sigma}) = \mathbb{E}(\frac{1}{n}x_i x_i^\top) = \Sigma$.

Linear Ridge Regression

The linear ridge regression aims to solve the minimization problem:

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

which admits the closed-form solution

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

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 $\phi : \mathbb{R}^d \to \mathbb{R}^p$ 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 $\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$. 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

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

with the closed-form solution

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

where $\phi(X) = [\phi(x_1), \cdots, \phi(x_n)]^\top \in \mathbb{R}^{n\times p}$ is the feature matrix and $\widehat{\Sigma}_\phi = \frac{1}{n}\phi(X)^\top\phi(X) \in \mathbb{R}^{p\times p}$ 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 $W = [w_1, \cdots, w_p]^\top \in \mathbb{R}^{p\times d}$ are sampled from the the Fourier transform of the kernel and the bias vector 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: $\phi(x) = \phi_L(\phi_{L-1}(\cdots\phi_1(x)))$, where L is the depth of the network and the feature mappings $\phi_1, \cdots, \phi_L$ 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

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

where the closed form solution is

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

where $\widehat{\Sigma}_{S\phi} = \frac{1}{m}\phi(X)^\top S^\top S\phi(X) \in \mathbb{R}^{p\times p}$. 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, $[0, 1, 0, \cdots, 0]$ 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. $n, d, p, m \to \infty$, in such a way that $p/m \to \gamma \in (0, \infty)$. 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 $\phi : \mathbb{R}^d \to \mathbb{R}^p$ 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 $\Sigma_\phi = \mathbb{E}[\phi(\boldsymbol{x})\phi(\boldsymbol{x})^\top]$ when $p \to \infty$.*

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

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)**.** *Supoose 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 Gaussain 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 $\widehat{v}(z) = \frac{1}{m}\mathrm{tr}\left[\left(\widehat{\mathbf{K}}_{S\phi} - z\mathrm{I}\right)^{-1}\right]$, 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):

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

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 $\hat{v}(-\lambda) = \frac{1}{m}\mathrm{tr}[(\widehat{\mathbf{K}}_{S\phi} + \lambda\mathrm{I})^{-1}]$, which converges to $v(-\lambda)$ almost surely. From Section A.1 in (Bach 2023), using the the self-consistency equation (7), there holds the asymptotic equivalence for the effective dimension (also called degree of freedom) for $\lambda > 0$,

$$\begin{aligned} \widehat{\mathrm{df}}_1(\lambda) &= \mathrm{tr}\left(\widehat{\Sigma}_{S\phi}(\widehat{\Sigma}_{S\phi} + \lambda\mathrm{I})^{-1}\right) \\ \sim \quad \mathrm{df}_1(\kappa) &= \mathrm{tr}\left(\Sigma_{S\phi}(\Sigma_{S\phi} + \kappa\mathrm{I})^{-1}\right), \end{aligned} \tag{8}$$

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 $n, d, p, m \to +\infty$. 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 $\frac{1}{p}\mathrm{df}_1(\kappa) = \frac{1}{\mathrm{p}}\sum_{\mathrm{i}=1}^{\mathrm{p}} \frac{\sigma_\mathrm{i}}{\sigma_\mathrm{i}+\kappa} \to \int_0^{+\infty} \frac{\sigma \mathrm{d}\mu(\sigma)}{\sigma+\kappa}$, which is decreasing in κ and converges to one when $\kappa = 0$, while $\mathrm{df}_1(0) = \mathrm{rank}(\Sigma_{\mathrm{S}\phi})$. Therefore, we have $\mathrm{df}_1\left(\frac{1}{\mathrm{v(z)}}\right) \to \mathrm{p}\int_0^{+\infty} \frac{\sigma \mathrm{d}\mu(\sigma)}{1/\mathrm{v(z)}+\sigma}$ when $\kappa = \frac{1}{v(z)}$. Then, using $z = -\lambda$ and $\kappa = \frac{1}{v(-\lambda)}$, we rewrite (7) as

$$\lambda \sim \kappa\left(1 - \frac{1}{m}\mathrm{df}_1(\kappa)\right). \tag{9}$$

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

Remark 2 (Implicit regularization)**.** *We consider the ridgeless settings when $\lambda \to 0$ in terms of γ.*

1) Underparameterization ($\gamma < 1$): Since $\mathrm{df}_1(\kappa) \le \mathrm{p}$ and $(1 - \frac{1}{m}\mathrm{df}_1(\kappa)) > 0$ from (9)*, we have $\kappa = 0$ when $\lambda = 0$. Meanwhile, $\Sigma_{S\phi}$ is invertible from Assumption 5, such that $\mathrm{df}_1(\kappa) = \mathrm{p}$ and $\lambda \sim \kappa(1-\gamma)$ when λ goes to zero.*

2) Overparameterization ($\gamma > 1$): If $\kappa = 0$ when $\lambda = 0$, from (9) *we have $\lambda \sim \kappa(1-\gamma) < 0$, violating the fact $\lambda > 0$. Therefore, when $\lambda \to 0_+$, we have $\kappa > 0$ and $\mathrm{df}_1(\kappa) \to \mathrm{m}_+$. There is an implicit regularization parameter $\kappa > 0$ with $\mathrm{df}_1(\kappa) = \mathrm{m}$ 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:*

$$\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda \mathrm{I})^{-1}\right] \sim \frac{\kappa}{\lambda}\mathrm{df}_1(\kappa), \tag{10}$$

$$\begin{aligned}&\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda \mathrm{I})^{-2}\right] \\ &\quad\sim \frac{\kappa^2}{\lambda^2}\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa \mathrm{I})^{-2}\right] \cdot \frac{\mathrm{m}}{\mathrm{m} - \mathrm{df}_2(\kappa)},\end{aligned} \tag{11}$$

$$\begin{aligned}&\theta_*^\top(\widehat{\Sigma}_{S\phi} + \lambda I)^{-1}\Sigma_{S\phi}(\widehat{\Sigma}_{S\phi} + \lambda I)^{-1}\theta_* \\ &\quad\sim \frac{\kappa^2}{\lambda^2}\theta_*^\top(\Sigma_{S\phi} + \kappa I)^{-2}\Sigma_{S\phi}\theta_* \cdot \frac{m}{m - \mathrm{df}_2(\kappa)}.\end{aligned} \tag{12}$$

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

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 $x_1, \cdots, x_n$ are assumed deterministic and the expected excess risk measures the "in-sample" error $\mathbb{E}_\varepsilon\left[\|\hat{\theta} - \theta_*\|_{\widehat{\Sigma}_{S\phi}}^2\right]$ 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. $\mathbb{E}_\varepsilon\left[\|\hat{\theta} - \theta_*\|_{\Sigma_{S\phi}}^2\right]$, 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*

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

where

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

Note that, in the proof of the variance term, there is a sketched covariance $\frac{\phi(X)^\top S^\top S S^\top S\phi(X)}{m}$, 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 $n \ge p$ 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:*

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

From (13), we find that both the variance and bias terms are increasing for larger $\mathrm{df}_2(\kappa)$ and the excess risk explodes when $\mathrm{df}_2(\kappa) \to \mathrm{m}$.

Remark 3. *The value of $\mathrm{df}_2(\kappa)$ is influenced by three factors: 1) The explicit regularization parameter λ: According to the self-consistency equation in* (9)*, κ is positively correlated with λ. In (Bach 2023), it is suggested that choosing an appropriate λ can ensure $\mathrm{df}_2(\kappa) \ll \mathrm{m}$ 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 $\mathrm{df}_2(\kappa)$ *still depends on the choice of the feature mapping. Using a suitable feature mapping with hyperparameters or trainable feature mapping can further decrease* $\mathrm{df}_2(\kappa)$ *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* $\mathrm{df}_2(\kappa)$ *is also affected by the subsampling matrix. By employing an appropriate subsampling matrix, even through downsampling, it is possible to reduce* $\mathrm{df}_2(\kappa)$.

Since $\mathrm{df}_2(\kappa)$ defined on the covariance matrix of X for linear regression, $\mathrm{df}_2(\kappa)$ can only be reduced by selecting λ in (Bach 2023). We introduce another two ways to reduce $\mathrm{df}_2(\kappa)$, including feature mapping and subsampling. Especially, the trainable feature mapping is easier to implement using backpropagation. Beside $\mathrm{df}_2(\kappa)$, the excess risk is strictly decreasing in the sketching size m. Oversampling $m > n$ can reduce the risk but bring additional computational burdens. However, a favourable downsampling matrix where $m < n$ can still reduce $\mathrm{df}_2(\kappa)$ that compensates the increasing risk from smaller m.

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 $\hat{\theta} = \theta_* + \widehat{\Sigma}_{S\phi}^{-1} \frac{\phi(X)^\top S^\top S\varepsilon}{m}$.

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 $\mathrm{df}_2(\kappa) = \mathrm{rank}(\widehat{\Sigma}_{\mathrm{S}\phi}) = \mathrm{p}$. Substituting $\kappa = 0$ and $\mathrm{df}_2(\kappa) = \mathrm{p}$ to (13), 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:*

$$\begin{aligned} &\mathbb{E}_\varepsilon\left[\left\|\hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta})\right\|_{\Sigma_{S\phi}}^2\right] \sim \sigma^2 \frac{p}{m-p}, \\ &\left\|\mathbb{E}_\varepsilon(\hat{\theta}) - \theta_*\right\|_{\Sigma_{S\phi}}^2 = 0. \end{aligned} \tag{14}$$

In (14), we recover the classical results in Theorem 1 of (Hastie et al. 2022) for underparameterization. When $\gamma \to 1$, i.e. $p \to m$, 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 $\mathrm{df}_1(\kappa) = \mathrm{m}$ from Remark 2 and $\mathrm{df}_2(\kappa) \le \mathrm{df}_1(\kappa) = \mathrm{m}$.

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

$$\begin{aligned} &\mathbb{E}_\varepsilon\left[\left\|\hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta})\right\|_{\Sigma_{S\phi}}^2\right] \sim \sigma^2 \frac{\mathrm{df}_2(\kappa_0)}{m - \mathrm{df}_2(\kappa_0)}, \\ &\left\|\mathbb{E}_\varepsilon(\hat{\theta}) - \theta_*\right\|_{\Sigma_{S\phi}}^2 = \frac{m\kappa^2 \theta_*^\top (\Sigma_{S\phi} + \kappa I)^{-2} \Sigma_{S\phi} \theta_*}{m - \mathrm{df}_2(\kappa_0)}. \end{aligned} \tag{15}$$

The excess risk in over-parameterized regime depends on the differences between two effective dimensions $\mathrm{df}_1(\kappa_0) - \mathrm{df}_2(\kappa_0)$. If $\mathrm{df}_2(\kappa_0) \ll \mathrm{df}_1(\kappa_0)$, 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 $m \to \infty$. If $\mathrm{df}_2(\kappa_0) = \mathrm{rmdf}_1(\kappa_0)$, there is catastrophic overffitting 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 overffitting but the variance term is a constant.

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

Trainable Nonlinear Regression Model

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

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{n} \|S\phi(X)\theta - Sy\|_2^2 + \lambda \|\theta\|_2^2 \\ \text{s.t.} \quad & \{\lambda, \phi, S\} = \arg\min_{\lambda,\phi,S} \ \widehat{\mathrm{df}}_2(\lambda). \end{aligned} \tag{16}$$

The value of $\mathrm{df}_2(\kappa)$ depends on λ, the feature mapping ϕ, and the subsampling matrix S, as discussed in Remark 3. To solve (16), 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 $\widehat{\mathrm{df}}_2(\lambda) = \mathrm{tr}\left((\widetilde{\mathrm{X}}^\top \widetilde{\mathrm{X}})^2 (\widetilde{\mathrm{X}}^\top \widetilde{\mathrm{X}} + \lambda \mathrm{mI})^{-2}\right)$, where $\widetilde{X} = S\phi(X)$, 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

$$\begin{aligned} \widehat{\mathrm{df}}_2(\lambda) = &\mathrm{tr}\Big(\widetilde{\mathrm{X}}^\top (\widetilde{\mathrm{X}} \widetilde{\mathrm{X}}^\top + \lambda \mathrm{mI})^{-1} \widetilde{\mathrm{X}} \\ &\widetilde{X}^\top (\widetilde{X} \widetilde{X}^\top + \lambda m I)^{-1} \widetilde{X}\Big). \end{aligned} \tag{17}$$

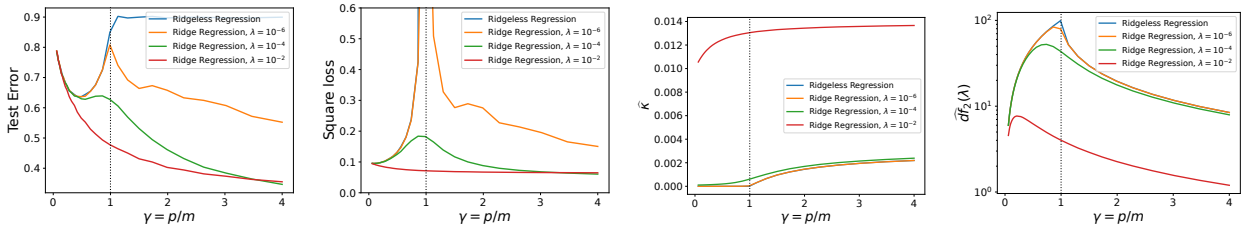

Figure 1: The testing error, testing loss, implicit regularization parameter $\widehat{\kappa}$, and effective dimension $\widehat{\mathrm{df}}_2(\widehat{\kappa})$ versus the increase of the feature dimension p on the MNIST dataset.

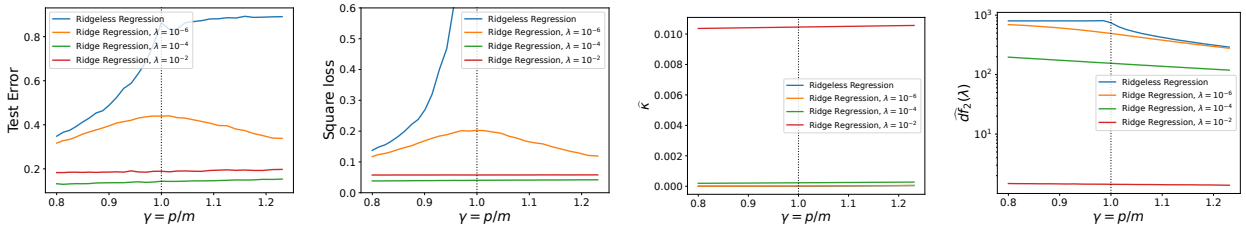

Figure 2: The testing error, testing loss, implicit regularization parameter $\widehat{\kappa}$, and effective dimension $\widehat{\mathrm{df}}_2(\widehat{\kappa})$ versus the increase of the subsampling size m on the MNIST dataset.

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

Example: Random Feature Regression with Effective Dimension (RFRed)

From (16), there are too many hyperparamters to optimize and the compute of (17) 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 (16) with the feature mapping

$$\phi(x) = \sqrt{\frac{2}{p}} \cos(W^\top x + b), \tag{18}$$

where the frequency matrix $W = [w_1, \cdots, w_p] \in \mathbb{R}^{d\times p}$ composed 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 $b = [b_1, \cdots, b_p] \in \mathbb{R}^p$ are drawn uniformly from $[0, 2\pi]$.

To improve the computational efficiency, we tune hyperparameters λ, S before the training and optimize the feature mapping ϕ during the training. To accelerate the solve of (16), we optimize θ and W jointly by minimize the the following objective

$$\mathcal{L}(\theta;\, W) = \frac{1}{n}\, \|S\phi(X)\theta - Sy\|_2^2 + \lambda\|\theta\|_2^2 + \beta\, \widehat{\mathrm{df}}_2(\lambda), \tag{19}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^p$ is defined in (18), $\widehat{\mathrm{df}}_2(\lambda)$ is defined in (17), $\lambda = 0$ 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 $\nabla_\theta \mathcal{L} = \frac{1}{n}\widetilde{X}_b^\top(\widetilde{X}_b\theta - \widetilde{y}_b)$ where $\{\widetilde{X}_b \in \mathbb{R}^{b\times p}, \widetilde{y}_b \in \mathbb{R}^b\}$ is a batch of $\{S\phi(X), Sy\}$ with the batch size b. We also use the batch data to approximate $\widehat{\mathrm{df}}_2(\lambda)$ where $\widetilde{X}$ in (17) is replaced by $\widetilde{X}_b$. 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 $\widehat{\mathrm{df}}_2(\lambda)$ requires $\mathcal{O}(\frac{p^2nT}{n\alpha})$.

Experiments

We utilize the random Fourier feature, as defined in equation (18), to provide an approximation of the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma^2\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2)$. It is important to note that the random Fourier features, specified in equation (18), 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 $\{0.01, \cdots, 1000\}$ and $\lambda \in \{0.1, \cdots, 10^{-5}\}$. 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 $\kappa \sim \widehat{\kappa} = 1/\mathrm{tr}[(\phi(\mathrm{X})\phi(\mathrm{X})^\top + \lambda \mathrm{nI})^{-1}]$ and use $\widehat{\mathrm{df}}_2(\widehat{\kappa})$ to approximate the key quantity $\mathrm{df}_2(\kappa)$. 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 $\mathrm{df}_2(\kappa)$ and

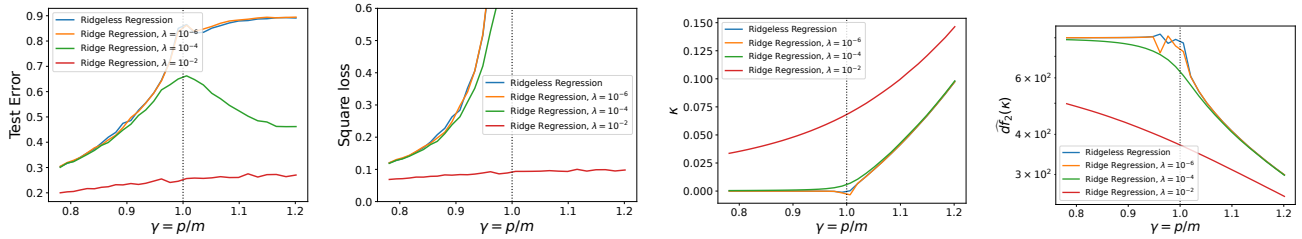

Figure 3: The testing error, testing loss, implicit regularization parameter κ, and effective dimension $\widehat{\mathrm{df}}_2(\kappa)$ versus the increase of the ROS size on the MNIST dataset.

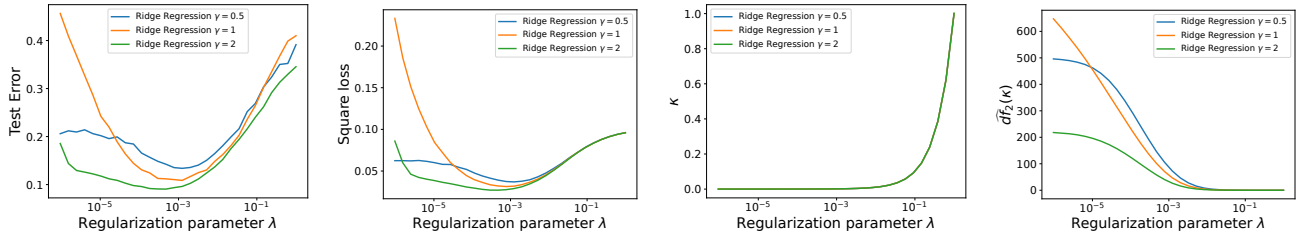

Figure 4: Testing error, testing loss, implicit regularization parameter κ, and effective dimension $\widehat{\mathrm{df}}_2(\kappa)$ w.r.t. different values of λ on the MNIST dataset.

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 $m - \mathrm{df}_2(\kappa)$ (Theorem 1). In the overparameterized setting ($\gamma > 1$), test errors decrease again as $\mathrm{df}_2(\kappa)$ 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 $\mathrm{rank}(\widehat{\Sigma}_{\mathrm{S}\phi}) \ll$ p. 3) Increasing p leads to a larger $\widehat{\kappa}$ approaching λ, and $\widehat{\kappa}_0$ is similar to $\widehat{\kappa}$ with smaller λ when $\lambda = 0$. Smaller $\mathrm{df}_2(\kappa)$ 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 overparameterized 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 $\widehat{\mathrm{df}}_2(\lambda)$ 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 $\widehat{\mathrm{df}}_2(\lambda)$.

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 $\widehat{\mathrm{df}}_2(\kappa)$ are positive correlated. 4) When the regularization parameter is small, the threshold $\gamma = 1$ have highest test error, while the underparameterized 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 Unicom 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 Downsampling. *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.

Richards, D.; Mourtada, J.; and Rosasco, L. 2021. Asymptotics of ridge (less) regression under general source condition. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3889–3897. PMLR.

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.

Proofs

Proof of Proposition 1. Let $L^2(X, P_x) = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_2^2 = \int |f(x)|^2 dP_x(x)\}$ be the square-integrable function space. For all $g \in L^2(X, \rho_X), \beta \in \mathbb{R}^p$, we have

$$\begin{aligned}
&\mathcal{S}_\phi : \mathbb{R}^p \to L^2(X, \rho_X), \qquad (\mathcal{S}_\phi \beta)(\cdot) = \langle \phi(\cdot), \beta \rangle, \\
&\mathcal{S}_\phi^* : L^2(X, \rho_X) \to \mathbb{R}^p, \qquad \mathcal{S}_\phi^* g = \int_X \phi(x) g(x) d\, P_x(x), \\
&\Sigma_\phi : \mathbb{R}^p \to \mathbb{R}^p, \qquad \Sigma_\phi = \mathcal{S}_\phi^* \mathcal{S}_\phi = \int_X \phi(x)\phi(x)^\top d\, P_x(x).
\end{aligned}$$

□

Proof of Proposition 2. Following the asymptotic equivalence provided in Proposition 1 (Bach 2023), we can prove our results.

Using Eq. (8) in Proposition 1 (Bach 2023) with $A = \Sigma_{S\phi}$, $z = -\lambda$ and $\varphi(z) = \frac{1}{-\kappa}$, we prove (10).

Using Eq. (9) in Proposition 1 (Bach 2023) with $A = \Sigma_{S\phi}, B = I$, $z = -\lambda$ and $\varphi(z) = \frac{1}{-\kappa}$, we prove (11).

Using Eq. (9) in Proposition 1 (Bach 2023) with $A = \theta_* \theta_*^\top, B = \Sigma_{S\phi}$, $z = -\lambda$ and $\varphi(z) = \frac{1}{-\kappa}$, we prove (12). □

Proof of Lemma 1. Using $\mathbb{E}_\varepsilon(\hat{\theta}) = \left(\widehat{\Sigma}_{S\phi} + \lambda\, I\right)^{-1} \widehat{\Sigma}_{S\phi} \theta_*$, we have

$$\begin{aligned}
\mathbb{E}_\varepsilon \left[\left\| \hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta}) \right\|_{\Sigma_{S\phi}}^2 \right] &= \mathbb{E}_\varepsilon \left(\left\| (\widehat{\Sigma}_{S\phi} + \lambda\, I)^{-1} \frac{\phi(X)^\top S^\top S \varepsilon}{m} \right\|_{\Sigma_{S\phi}}^2 \right) \\
&= \mathbb{E}_\varepsilon \left(\left\| (\widehat{\Sigma}_{S\phi} + \lambda\, I)^{-1} \frac{\phi(X)^\top S^\top S \varepsilon}{m} \right\|_{\Sigma_{S\phi}}^2 \right) \\
&= \frac{\varepsilon^\top S^\top S \phi(X)}{m} (\widehat{\Sigma}_{S\phi} + \lambda\, I)^{-1} \Sigma_{S\phi} (\widehat{\Sigma}_{S\phi} + \lambda\, I)^{-1} \frac{\phi(X)^\top S^\top S \varepsilon}{m} \\
&= \mathrm{tr} \left[\frac{\varepsilon^\top \mathrm{S}^\top \mathrm{S} \phi(\mathrm{X})}{\mathrm{m}} (\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\, \mathrm{I})^{-1} \Sigma_{\mathrm{S}\phi} (\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\, \mathrm{I})^{-1} \frac{\phi(\mathrm{X})^\top \mathrm{S}^\top \mathrm{S} \varepsilon}{\mathrm{m}} \right] \\
&= \frac{\sigma^2}{m} \mathrm{tr} \left[(\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\, \mathrm{I})^{-1} \frac{\phi(\mathrm{X})^\top (\mathrm{S}^\top \mathrm{S})^2 \phi(\mathrm{X})}{\mathrm{m}} (\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\, \mathrm{I})^{-1} \Sigma_{\mathrm{S}\phi} \right], \\
&= \frac{\sigma^2}{m} \mathrm{tr} \left[(\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\, \mathrm{I})^{-1} \widehat{\Sigma}_{\mathrm{S}\phi} (\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\, \mathrm{I})^{-1} \Sigma_{\mathrm{S}\phi} \right].
\end{aligned}$$

In the above proof, we use the cyclic property of matrix trace and $(S^\top S)^2 = S^\top S$ since $SS^\top = I_m$ from Assumption 4.

We then estimate the bias term

$$\begin{aligned}
\left\| \mathbb{E}_\varepsilon(\hat{\theta}) - \theta_* \right\|_{\Sigma_{S\phi}}^2 &= \mathbb{E}_\varepsilon \left[\left\| \left(\left(\widehat{\Sigma}_{S\phi} + \lambda\, I \right)^{-1} \widehat{\Sigma}_{S\phi} - I \right) \theta_* \right\|_{\Sigma_{S\phi}}^2 \right] \\
&= \mathbb{E}_\varepsilon \left[\left\| \left(\left(\widehat{\Sigma}_{S\phi} + \lambda\, I \right)^{-1} \widehat{\Sigma}_{S\phi} - \left(\widehat{\Sigma}_{S\phi} + \lambda\, I \right)^{-1} \left(\widehat{\Sigma}_{S\phi} + \lambda\, I \right) \right) \theta_* \right\|_{\Sigma_{S\phi}}^2 \right] \\
&= \mathbb{E}_\varepsilon \left[\left\| \lambda \left(\widehat{\Sigma}_{S\phi} + \lambda\, I \right)^{-1} \theta_* \right\|_{\Sigma_{S\phi}}^2 \right] \\
&= \lambda^2 \theta_*^\top (\widehat{\Sigma}_{S\phi} + \lambda\, I)^{-1} \Sigma_{S\phi} (\widehat{\Sigma}_{S\phi} + \lambda\, I)^{-1} \theta_*.
\end{aligned}$$

□

Proof of Theorem 1. From the bias-variance decomposition in Lemma 1, we use the asymptotic equivalents (10), (11) and

$\mathrm{df}_1(\kappa) - \mathrm{df}_2(\kappa) = \kappa \operatorname{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa \mathrm{I})^{-2}\right]$, we estimate the limiting variance

$$
\begin{aligned}
&\mathbb{E}_\varepsilon\left[\left\|\hat{\theta} - \mathbb{E}_\varepsilon(\hat{\theta})\right\|_{\Sigma_{S\phi}}^2\right] \\
=&\frac{\sigma^2}{m}\mathrm{tr}\left[(\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\,\mathrm{I})^{-1}\widehat{\Sigma}_{\mathrm{S}\phi}(\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\,\mathrm{I})^{-1}\Sigma_{\mathrm{S}\phi}\right] \\
=&\frac{\sigma^2}{m}\left[\mathrm{tr}\left((\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\,\mathrm{I})^{-1}\Sigma_{\mathrm{S}\phi}\right) - \lambda\mathrm{tr}\left((\widehat{\Sigma}_{\mathrm{S}\phi} + \lambda\,\mathrm{I})^{-2}\Sigma_{\mathrm{S}\phi}\right)\right] \\
\sim&\frac{\sigma^2\kappa}{m\lambda}\Bigg[\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-1}\right] - \kappa\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}\right] \\
&\qquad - \kappa\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}^2(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}\right]\cdot\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}\right]\cdot\frac{1}{\mathrm{m} - \mathrm{df}_2(\kappa)}\Bigg] \\
=&\frac{\sigma^2\kappa}{m\lambda}\Bigg[\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})\right] - \kappa\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}\right] \\
&\qquad - \kappa\,\mathrm{df}_2(\kappa)\cdot\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}\right]\cdot\frac{1}{\mathrm{m} - \mathrm{df}_2(\kappa)}\Bigg] \\
=&\frac{\sigma^2\kappa}{m\lambda}\left[\mathrm{df}_2(\kappa) - \kappa\,\mathrm{df}_2(\kappa)\cdot\mathrm{tr}\left[\Sigma_{\mathrm{S}\phi}(\Sigma_{\mathrm{S}\phi} + \kappa\,\mathrm{I})^{-2}\right]\cdot\frac{1}{\mathrm{m} - \mathrm{df}_2(\kappa)}\right] \\
=&\frac{\sigma^2\kappa}{m\lambda}\left[\mathrm{df}_2(\kappa) - (\mathrm{df}_1(\kappa) - \mathrm{df}_2(\kappa))\cdot\frac{\mathrm{df}_2(\kappa)}{\mathrm{m} - \mathrm{df}_2(\kappa)}\right] \\
=&\frac{\sigma^2\kappa}{m\lambda}\left[\frac{\mathrm{df}_2(\kappa)(\mathrm{m} - \mathrm{df}_1(\kappa))}{m - \mathrm{df}_2(\kappa)}\right] \\
=&\frac{\sigma^2}{\lambda}\left[\kappa\left(1 - \frac{1}{m}\mathrm{df}_1(\kappa)\right)\cdot\frac{\mathrm{df}_2(\kappa)}{m - \mathrm{df}_2(\kappa)}\right] \\
\sim&\,\sigma^2\frac{\mathrm{df}_2(\kappa)}{m - \mathrm{df}_2(\kappa)}.
\end{aligned}
$$

The last step is due to (9) where $\lambda \sim \kappa\left(1 - \frac{1}{m}\mathrm{df}_1(\kappa)\right)$.

Using the asymptotic equivalent (12), we have

$$
\begin{aligned}
&\left\|\mathbb{E}_\varepsilon(\hat{\theta}) - \theta_*\right\|_{\Sigma_{S\phi}}^2 \\
=&\,\lambda^2\theta_*^\top(\widehat{\Sigma}_{S\phi} + \lambda\,I)^{-1}\Sigma_{S\phi}(\widehat{\Sigma}_{S\phi} + \lambda\,I)^{-1}\theta_* \\
\sim&\,\kappa^2\theta_*^\top(\Sigma_{S\phi} + \kappa\,I)^{-2}\Sigma_{S\phi}\theta_* + \kappa^2\theta_*^\top(\Sigma_{S\phi} + \kappa\,I)^{-2}\Sigma_{S\phi}\theta_*\cdot\frac{\mathrm{df}_2(\kappa)}{m - \mathrm{df}_2(\kappa)} \\
=&\,\kappa^2\theta_*^\top(\Sigma_{S\phi} + \kappa\,I)^{-2}\Sigma_{S\phi}\theta_*\cdot\frac{m}{m - \mathrm{df}_2(\kappa)}.
\end{aligned}
$$

□

More Experiments

Impact of the Trainable Feature Mapping ϕ

Under different settings $\gamma = 0.5, 1, 2$, we compare random features estimators with fixed weights W in (18) with trainable random features methods with stochastic gradient descent (SGD) to update the weights W. We denote RF-SGD as the trainable feature mapping. We fix $n = 1000, \sigma^2 = 0.01, \lambda = 0.1$ and use different ratios γ. We record the test error, training loss, implicit regularization parameter κ and the effective dimension $\widehat{\mathrm{df}}_2(\kappa)$ w.r.t. the training epochs in Figure 5. That implies: 1) In terms of the training epochs, the test error, training loss, κ and $\widehat{\mathrm{df}}_2(\kappa)$ are positive correlated, where they decrease in the number of the features size. 2) The trainable feature mapping estimators achieve much lower testing errors and training losses

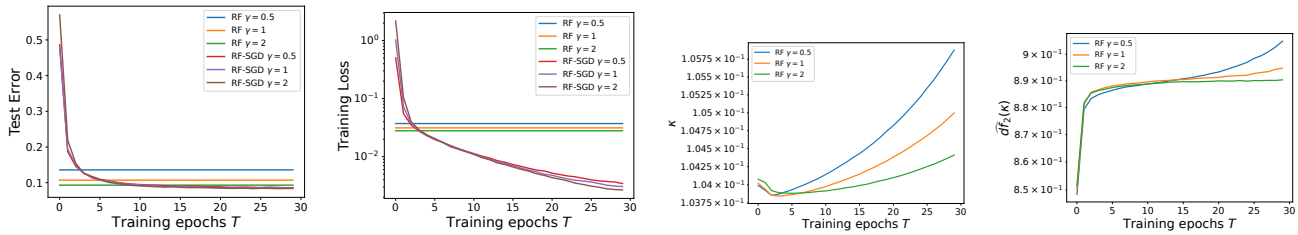

Figure 5: Testing error, testing loss, implicit regularization parameter κ, and effective dimension $\widehat{\mathrm{df}}_2(\kappa)$ w.r.t. the number of training epochs on the MNIST dataset.

Dataset	σ^2	λ	β
MNIST	0.01	0.0001	0.001
usps	0.1	1	0.00001

Table 1: Summary of datasets and hyperparameters.

than fixed feature mapping methods, which validates the effectiveness of the proposed training framework 19. 3) Under suitable regularization, the estimator with larger feature size p leads to lower test error.

Comparison Experiments

We conduct comparison experiments for RFRed, Ridge Regression with random features (Ridge RF) and Ridgeless Regression with random features and SGD (Ridgeless RF-SGD) that update the feature mapping and model parameter jointly. For all datasets from Libsvm datasets [2], we tune the hyperparameters λ, β, σ^2 via NNI [3]. For the under-parameterized regime, we set $n = 1000, p = 500$ where $\gamma = 0.5$; for the over-parameterized regime, we set $n = 1000, p = 2000$ where $\gamma = 2$.

We report results on the MNIST dataset in Figure 6 and results on the usps dataset in Figure 7. We find that 1) the proposed RFRed outperforms both Ridge RF and Ridgeles RF-SGD, while Ridge RF usually performs much worse than RFRed and Ridgeles RF-SGD; 2) Although the effective dimension of Ridge RF is much smaller than the others, RFRed and Ridgeless RF-SGD achieve better performance because we compute the effective dimension on mini-batch for RFRed and Ridgeless RF-SGD while we compute it on the entire dataset for Ridge RF.

Comparison with Related Work

We compare the results in this paper with recent asymptotic results on linear regression, feature mapping models, and subsampling models.

Linear regression. Using random matrix theory tools, (Hastie et al. 2022; Richards, Mourtada, and Rosasco 2021) provided asymptotic limit framework for high dimensional analysis of linear regression. However, these results still keep Stieltjes transforms that are hard to estimate, while in this paper we interpret the asymptotic results with effective dimensions. Meanwhile, we extend the analysis for linear models to nonlinear models with subsampling. Key results in (Hastie et al. 2022; Richards, Mourtada, and Rosasco 2021) can be special cases of this paper with $S = I_n$ and $\phi(X) = X$.

Feature mapping. (Mei and Montanari 2022) provided asymptotics for random features regression but the results independent on the model capacity terms and hard to follow. Considering random projections, (Bach 2023) derived the asymptotic equivalents that are related to effective dimensions, but it assumed the optimal estimator in the input space $\theta_* \in \mathbb{R}^d$ while this paper assumes $\theta_* \in \mathbb{R}^p$ give by feature mapping. Our results apply to random feature models when $\phi(X) = \cos(W^\top x + b)$ as defined in (Rahimi and Recht 2007) and random projections when $\phi(X) = XW$ where W is a random projection matrix.

Subsampling. Based on random matrix theory, (Dobriban and Liu 2019) analyzed asymptotics for various sketching matrices, including both i.i.d. sketching matrix and orthogonal sketching matrix. However, these results only apply to the under-parameterized regime when $n > d$. More recently, (Chen et al. 2023) studied asymptotic limits for sketching ridgeless regression in both under-parameterized and over-parameterized regimes, but the results remain self-consistency equations that are not intuitive enough. In this paper, we provide asymptotic limits for sketched nonlinear regression and the results are interpreted with traditional effective dimensions. Our results can be applied to orthogonal sketching regression when $\phi(X) = X$ and S is an orthogonal matrix.

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[3] https://nni.readthedocs.io/

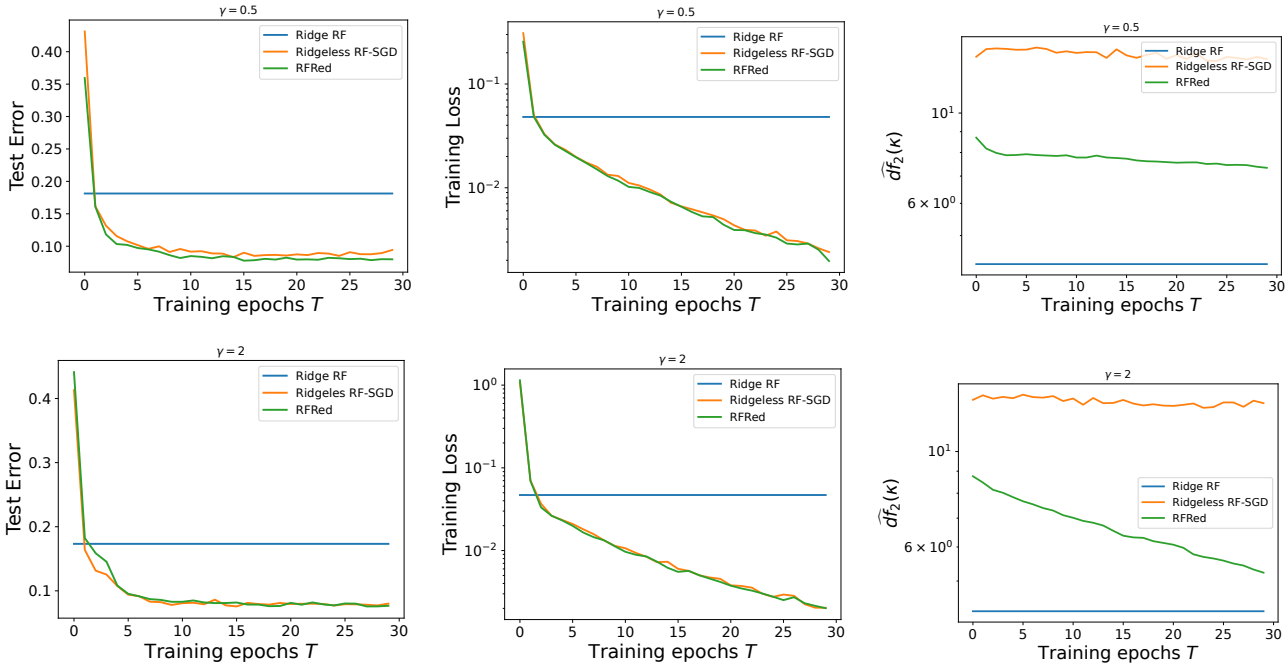

Figure 6: Testing error, testing loss, and effective dimension $\widehat{\mathrm{df}}_2(\kappa)$ w.r.t. the number of training epochs on the MNIST dataset for the under-parameterized setting ($\gamma = 0.5$) and the over-parameterized setting ($\gamma = 2$), respectively.

Conclusion

We explore the asymptotic properties of generalized nonlinear regression models and discover that the excess risk decreases as the effective dimension decreases. To leverage this finding, we propose an efficient nonlinear algorithm that dynamically adjusts the feature mapping to minimize the effective dimension. The techniques presented in this paper offer both theoretical and algorithmic insights for high-dimensional analysis of modern models and the design of new algorithms. There are several promising directions for further exploration. These include extending our results to i.i.d. sketching matrices (Dobriban and Liu 2019; Chen et al. 2023), relaxing the covariance conditions on nonlinear feature mapping (Mei and Montanari 2022), investigating them to classification tasks (Frei, Chatterji, and Bartlett 2022), and exploring connections with multilayer neural networks (Nakkiran et al. 2021).

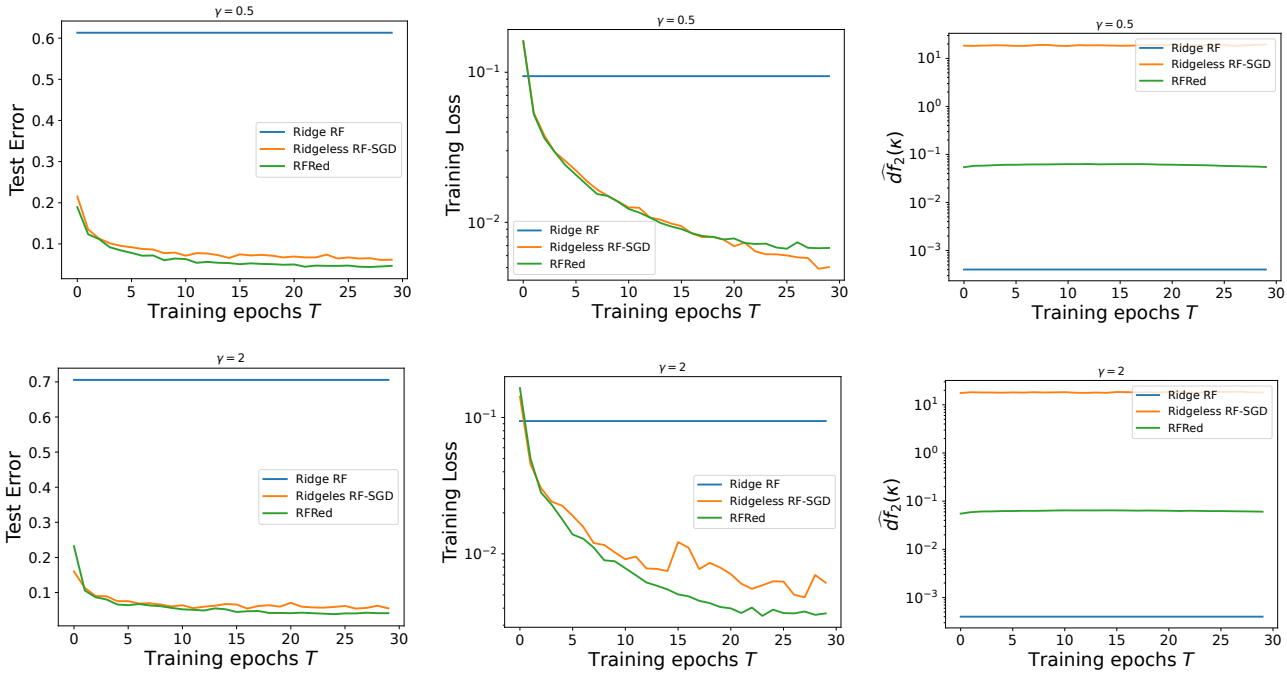

Figure 7: Testing error, testing loss, and effective dimension $\widehat{\mathrm{df}}_2(\kappa)$ w.r.t. the number of training epochs on the usps dataset for the under-parameterized setting ($\gamma = 0.5$) and the over-parameterized setting ($\gamma = 2$), respectively.