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

Jian Li\textsuperscript{1}, Yong Liu\textsuperscript{2*}, Weiping Wang\textsuperscript{1}

\textsuperscript{1}Institute of Information Engineering, Chinese Academy of Sciences
\textsuperscript{2}Gaoling 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 $\lambda$ 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). 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\textsuperscript{1}. Our contributions can be summarized as follows:

\begin{itemize}
  \item We provide a generalized asymptotic analysis framework for general nonlinear regression models, where the limit-
\end{itemize}
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 $x_1, \ldots, x_n \in \mathbb{R}^d$ 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 $y_i = \eta_i^\top x_i + \epsilon_i$, with $x_i$ and $\epsilon_i$ are independent, $\eta_i$ is the underlying parameter vector, $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. The optimal parameter estimate $\hat{\eta} \in \mathbb{R}^d$ satisfies $\mathbb{E}(y_i - \hat{\eta}_i^\top x_i) = \min_{\eta} \mathbb{E}(y_i - \eta_i^\top x_i)$.

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

**Linear Ridge Regression**

The linear ridge regression aims to solve the minimization problem:

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

which admits the closed-form solution

$$
\hat{\eta} = (\Sigma + \lambda I)^{-1} \Sigma \eta + (\Sigma + \lambda I)^{-1} X^\top \mathbb{E}(\Sigma) = \Sigma.
$$

**Generalized Nonlinear Regression Model**

Although the ridge linear regression has been well-studied in the high-dimensional setting (Dobriban and Wagner 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 $\theta_*$ 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 $y_i = f(x_i; \theta_*) + \epsilon_i$ where $\theta_* \in \mathbb{R}^p$ is the ideal estimator in the feature space and $\epsilon \in \mathbb{R}^n$ is the noise vector. The label noise $\epsilon$ is independent of $\phi(x_i)$ and follows a distribution on $\mathbb{R}$ such that $\mathbb{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$. We also assume the norm of $\theta_*$ 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

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

with the closed-form solution

$$
\hat{\theta} = (\Sigma_\phi + \lambda I)^{-1} \Sigma_\phi \theta_* + (\Sigma_\phi + \lambda n I)^{-1} \frac{\phi(X)^\top \epsilon}{n},
$$

where $\phi(X) = [\phi(x_1), \ldots, \phi(x_n)]^\top \in \mathbb{R}^{n \times p}$ is the feature matrix and $\Sigma_\phi = \frac{1}{n} \phi(X)^\top \phi(X) \in \mathbb{R}^{p \times p}$ 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 $W = [w_1, \ldots, w_p]^\top \in \mathbb{R}^{p \times d}$ 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:** $\phi(x) = \max\{Wx, 0\}$, 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

$$
\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{m} \|S \phi(X)\theta - S y\|_2^2 + \lambda \|\theta\|_2^2 \right\},
$$

where the closed form solution is

$$
\hat{\theta} = (\Sigma_\phi S + \lambda I)^{-1} \Sigma_\phi \theta_* + (\Sigma_\phi S + \lambda S^\top S)^{-1} \frac{\phi(X)^\top S^\top S \epsilon}{m},
$$

where $\Sigma_\phi S = \frac{1}{m} \phi(X)^\top S^\top S \phi(X) \in \mathbb{R}^{p \times p}$. 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 $x_2$. 

13501
• 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) $\phi(x) = \sqrt{\frac{2}{p}} \cos(Wx + b)$ provide continuous and bounded feature mapping.

**Proposition 1.** Under Assumption 2, the inputs in $X$ are sampled i.i.d. from $P$, and then the empirical covariance matrix of the feature mapping $\Sigma = \frac{1}{m} \phi(X)^T \phi(X)$ converges to a deterministic covariance matrix $\Sigma = E[\phi(x)\phi(x)^T]$ when $n \to \infty$.

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

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).** Suppose $\Sigma$ is invertible and bounded, and the eigenvalues of $\Sigma$ are positive and bounded, $\phi(X) = Z \Sigma^{1/2}$ 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 $SS^T = I_m$. Meanwhile, $S^T S$ converges to a deterministic matrix $\Sigma_S$.

**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 $\Sigma_{S \phi} = \frac{1}{m} \phi(X)^T S \phi(X)$ converges to a deterministic covariance matrix $\Sigma_{\phi} = \Sigma^{1/2} Z^T \Sigma S Z \Sigma^{1/2}$. The spectral distribution $\nu_{\Sigma_{S \phi}}$ of $\Sigma_{S \phi}$ converges to a limit probability distribution $\mu$ supported on $[0, +\infty)$ and $\Sigma$ is invertible and bounded in operator norm.

The above assumption implies there is no vanishing eigenvalues in the limiting $\mu$.

### Implicit Regularization and Asymptotic Equivalents

For any measure $G$ on $[0, \infty)$, we define the Stieltjes transform by $m_G(z) = \int \frac{\sigma d\mu}{1 + \sigma z}$, where $z \in \mathbb{C} \setminus \mathbb{R}^+$ and $\mu$ (which are mentions 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):

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

where $\mu$ 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 $v(-\lambda) = \frac{1}{m} \text{tr}([\tilde{K}_{S \phi} - z I])^{-1}$, 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 $\lambda > 0$, $\tilde{d}_1(\lambda) = \text{tr} \left( \Sigma_{S \phi} (\Sigma_{S \phi} + \lambda I)^{-1} \right) \sim d_1(\lambda) = \text{tr} \left( \Sigma_{S \phi} (\Sigma_{S \phi} + n I)^{-1} \right)$, where $\kappa = \frac{m}{m + \gamma}$ 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 $\lambda = 0$, $\kappa$ may not be zero and leads to implicit regularization in the over-parameterized regime $\gamma > 1$.

Using the definition of $\mu$ in Assumption 5, we have $\frac{1}{p} \tilde{d}_1(\kappa) = \frac{1}{p} \sum_{i=1}^p \frac{\sigma_i}{\sigma_i + \kappa} \to \int_0^{+\infty} \frac{\sigma d\mu}{1 + \sigma},$ which is decreasing in $\kappa$ and converges to one when $\kappa = 0$, while
the vector norm spectral functions of the empirical covariance operator, for the ridgeless regression
\[ \hat{\theta} = \arg \min_{\theta} \| \mathbf{S}(\mathbf{X}) \mathbf{X} - \mathbf{y} \| \mathbf{Q}^{-1} \|_2^2 \]
for the implicit regularization parameter \( \lambda \to \gamma > 0 \). We recover the bias-variance decomposition for the non-linear 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
\[
\text{Var}_{\varepsilon} \left[ \left\| \hat{\theta} - \theta_* \right\|^2_{\Sigma_{\phi}} \right] = \text{Var}_{\varepsilon} \left[ \left\| \hat{\theta} - \mathbb{E}_{\varepsilon} (\hat{\theta}) \right\|^2_{\Sigma_{\phi}} \right] + \| \mathbb{E}_{\varepsilon} (\hat{\theta}) - \theta_* \|^2_{\Sigma_{\phi}},
\]

where
\[
\text{Var} = \frac{\sigma^2}{m} \left[ (\hat{\Sigma}_{\phi} + \lambda I)^{-1} \hat{\Sigma}_{\phi} (\hat{\Sigma}_{\phi} + \lambda I)^{-1} \right],
\]
\[
(\text{Bias})^2 = \lambda^2 \theta_*^T (\hat{\Sigma}_{\phi} + \lambda I)^{-1} \hat{\Sigma}_{\phi} (\hat{\Sigma}_{\phi} + \lambda I)^{-1} \theta_*.
\]

Note that, in the proof of the variance term, there is a sketched covariance \( \hat{\phi}(\mathbf{X})^{T} \mathbf{S}^{T} \mathbf{S}^{T} \mathbf{S}_{\phi}(\mathbf{X}) \), 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-parametrized 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:
\[
\text{Var}_{\varepsilon} \left[ \left\| \hat{\theta} - \mathbb{E}_{\varepsilon} (\hat{\theta}) \right\|^2_{\Sigma_{\phi}} \right] \sim \frac{\sigma^2}{m} \frac{\text{df}_2(\kappa)}{m - \text{df}_2(\kappa)},
\]
\[
\left\| \mathbb{E}_{\varepsilon} (\hat{\theta}) - \theta_* \right\|^2_{\Sigma_{\phi}} \sim \frac{m \kappa^2 \theta_*^T (\hat{\Sigma}_{\phi} + \lambda I)^{-1} \hat{\Sigma}_{\phi} \theta_*}{m - \text{df}_2(\kappa)}.
\]
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 \( \text{df}_2(\kappa) \to m \).

**Remark 3.** The value of \( \text{df}_2(\kappa) \) is influenced by three factors: 1) The explicit regularization parameter \( \lambda \). According to the self-consistency equation in (8), \( \kappa \) is positively correlated with \( \lambda \). In (Bach 2023), it is suggested that choosing an appropriate \( \lambda \) can ensure \( \text{df}_2(\kappa) < m \) to prevent risk explosion. 2) The feature mapping: By considering the covariance on the feature mapping and subsampling of the inputs \( \mathbf{S}(\mathbf{X}) \) instead of the primal inputs \( \mathbf{X} \), 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 $d\hat{f}_2(\kappa)$ is also affected by the subsampling matrix. By employing an appropriate subsampling matrix, even through downsampling, it is possible to reduce $d\hat{f}_2(\kappa)$.

**Asymptotic Analysis of Ridgeless Regression**

We consider the limit when $\lambda = 0$ where the ridge regression estimator (6) becomes a minimum $\ell^2$-norm (ridgeless) estimator $\hat{\theta} = \theta + \frac{1}{m} \Sigma^{-1}_b \phi(X)^T S^T S \theta$.

We first consider the under-parameterized regime $\gamma < 1$ where $\kappa = \gamma$ as discussed in Remark 2. Since $\Sigma_b^{-1} \phi(X)$ is invertible, we have $d\hat{f}_2(\kappa) = \text{rank}(\Sigma_b^{-1}) = p$. Substituting $\kappa = 0$ and $d\hat{f}_2(\kappa) = p$ 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:

$$
E_{\epsilon} \left[ \left\| \hat{\theta} - E_{\epsilon}(\hat{\theta}) \right\|_{\Sigma_b}^2 \right] \sim \frac{\sigma^2}{m} \frac{p}{p - m}, \quad \left\| E_{\epsilon}(\hat{\theta}) - \theta \right\|_{\Sigma_b}^2 = 0.
$$

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 $\theta_*^T \phi(x)$ 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 $\phi(x) = W^T x$ 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 $\kappa$ is defined by $d\hat{f}_1(\kappa) = m$ from Remark 2 and $d\hat{f}_2(\kappa) \leq d\hat{f}_1(\kappa) = m$.

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

$$
E_{\epsilon} \left[ \left\| \hat{\theta} - E_{\epsilon}(\hat{\theta}) \right\|_{\Sigma_b}^2 \right] \sim \frac{\sigma^2}{{\frac{m}{m}}} \frac{d\hat{f}_2(\kappa_0)}{m - d\hat{f}_2(\kappa_0)},
$$

$$
E_{\epsilon}(\hat{\theta}) - \theta \parallel_{\Sigma_b}^2 = \frac{m\kappa_0^2 \theta_*^T (\Sigma_b + \kappa I)^{-2} \Sigma_b \theta_*}{m - d\hat{f}_2(\kappa_0)}.
$$

The excess risk in over-parameterized regime depends on the differences between two effective dimensions $d\hat{f}_1(\kappa_0) - d\hat{f}_2(\kappa_0)$. If $d\hat{f}_2(\kappa_0) \ll d\hat{f}_1(\kappa_0)$, the variance term tends to zero and the bias term tends to $\kappa^2 \theta_*^T (\Sigma_b + \kappa I)^{-2} \Sigma_b \theta_*$ when $m \rightarrow \infty$. If $d\hat{f}_2(\kappa_0) = m d\hat{f}_1(\kappa_0)$, 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 $\kappa_0 \sim \frac{1}{m} \Sigma_b(\Sigma_b + \kappa I)^{-1} \Sigma_b^{-1}$ and $d\hat{f}_2(\kappa_0)$ is decreasing in $\kappa_0$, we can optimize $S \phi(X)$ to increase $\kappa_0$ and reduce $d\hat{f}_2(\kappa_0)$ at the same time.

**Trainable Nonlinear Regression Model**

From Theorem 1, we notice that both the limiting variance and bias are increasing in $d\hat{f}_2(\kappa)$, and thus we lower the excess risk by reducing the effective dimension $d\hat{f}_2(\kappa)$. However, $d\hat{f}_2(\kappa)$ is hard to compute and thus we use the empirical effective dimension $\hat{d}f_2(\lambda)$ instead. That coincides with reducing $d\hat{f}_2(\lambda)$ can improve the fixed design risk $E_{\epsilon}[\|\hat{\theta} - \theta_*\|_{\Sigma_b}^2]$ = $\lambda^2 \text{tr}[\theta_* \theta^T (\Sigma_b + \kappa I)^{-1} \Sigma_b^{-1}] + \frac{\sigma^2}{m} d\hat{f}_2(\lambda)$. Therefore, smaller $d\hat{f}_2(\lambda)$ can lead to better performance. To obtain the smallest $d\hat{f}_2(\lambda)$, we devise a generalized nonlinear regression model with a bi-level problem

$$
\min_{\theta} \frac{1}{n} \|S \phi(X) \theta - S y\|_2^2 + \lambda \|\theta\|^2
$$

s.t. $\{\lambda, \phi, S\} = \arg \min_{\lambda, \phi, S} \hat{d}f_2(\lambda)$. (14)

The value of $d\hat{f}_2(\lambda)$ depends on $\lambda$, the feature mapping $\phi$, and the subsampling matrix $S$, as discussed in Remark 3. To solve (14), we alternate between optimizing the model parameter $\theta$ and the hyperparameters $\lambda, \phi, S$: 1) With fixed hyperparameters $\lambda, \phi, S$, the algorithm trains the nonlinear model $\theta$. 2) With a fixed nonlinear estimator $\theta$, the algorithm optimizes the hyperparameters $\lambda, \phi, S$.

To compute $\hat{d}f_2(\lambda)$, the computational complexity is typically impractical for the over-parameterized regime ($p > m$) at $O(p^3)$ time. For the under-parameterized regime, an alternative form is provided as

$$
\hat{d}f_2(\lambda) = \text{tr}(\tilde{X}^T \tilde{X} \lambda + \lambda m I)^{-1} \tilde{X} \tilde{X}^T (\tilde{X} \lambda + \lambda m I)^{-1} \tilde{X},
$$

where $\tilde{X} = S \phi(X)$. The time complexity for the above form is $O(p^2 n)$ since $p > n$. To accelerate the computational efficiency, we only compute $\hat{d}f_2(\lambda)$ for every $\alpha$-iterations rather 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 compute 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

$$
\phi(x) = \sqrt{\frac{\pi}{p}} \cos(W^T x + b),
$$

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

To improve the computational efficiency, we tune hyperparameters $\lambda, S$ before the training and optimize the feature
mapping $\phi$ during the training. To accelerate the solve of (14), we optimize $\theta$ and $W$ jointly by minimize the following objective

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

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

**Complexity.** Using batch stochastic gradient method, we have $\nabla_{\theta}L = \frac{1}{n} \overset{\sim}{X}_b^T (\hat{X}_b \theta - \hat{y}_b)$ where $\{\hat{X}_b \in \mathbb{R}^{b \times p}, \hat{y}_b \in \mathbb{R}^b\}$ is a batch of $\{S\phi(X), SY\}$ with the batch size $b$. We also use the batch data to approximate $\hat{d}_2(\lambda)$ where $\hat{X}$ in (15) is replaced by $\hat{X}_b$. The compute of $S\phi(X)$ consumes $O(mnp + ndp)$. With $T$ iterations, the update of $\theta$ takes $O(pbt)$ time, the update of $W$ consumes $O(p^2/2T)$, and the compute of $\hat{d}_2(\lambda)$ requires $O(nT/b)$. 

**Experiments**

We utilize the random Fourier feature, as defined in equation (16), to provide an approximation of the Gaussian kernel $K(x, x') = \exp(-\sigma^2\|x - x'\|^2/2)$. 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 $\sigma^2$ 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 $\kappa \sim \hat{\kappa} = 1/\text{tr}[(\phi(X)\phi(X)^T + \lambda n)^{-1}]$ and use $\hat{d}_2(\hat{\kappa})$ to approximate the key quantity $d_2(\kappa)$. The results are reported in Figure 1, which illustrated: 1) Mild regularization ($\gamma < 1$), 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 $m - d_2(\kappa)$ (Theorem 1). In the overparameterized setting ($\gamma > 1$), test errors decrease again as $d_2(\kappa)$ 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 $\text{rank}(\Sigma_{\phi\phi}) \ll p$. Increase $p$ leads to a larger $\hat{\kappa}$ approaching $\lambda$, and $\hat{\kappa}_0$ is

---

Figure 1: Testing error, testing loss, and effective dimension $\hat{d}_2(\hat{\kappa})$ versus the increase of the feature dimension $p$.

Figure 2: Testing error, testing loss, and effective dimension $\hat{d}_2(\hat{\kappa})$ versus the increase of the subsampling size $m$. 

13505
similar to $\hat{\mathcal{R}}$ with smaller $\lambda$ when $\lambda = 0$. Smaller $\hat{\mathcal{d}}_2(\kappa)$ 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 $\hat{\mathcal{d}}_2(\lambda)$ 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 $\hat{\mathcal{d}}_2(\lambda)$.

**Impact of the Different Regularization Parameter**

Under same settings as above experiments, we fix $n = 1000$ and vary the regularization parameter $\lambda \in \{10^{-5}, 1\}$ 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 $\lambda$ are similar even in different settings $\gamma = 0.5, 1, 2$, i.e. near $\lambda = 10^{-3}$, 2) The implicit regularization parameter $\kappa$ mainly depends on $\lambda$ rather than different $\gamma$. 3) Test error, squared loss, and the effective dimension $\hat{\mathcal{d}}_2(\kappa)$ 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.BJWYZJH012019100020098), the Unicom Innovation Ecological Cooperation Plan, and the CCF-Huawei Populus Grove Fund.

References


