A Survey of Double Descent in High-Dimensional Linear Regression[1]

Yilun Kuang **Linkai Ma**

Abstract

The recent observation of the double descent phenomenon in over-parametrized neural networks and kernel methods sparks renewed interest in the classical generalization error beyond the interpolation threshold $p = n$, where p is the feature dimension and n is the number of data points. This report reviews the risk of high-dimensional linear regression estimators in terms of the bias and variance decomposition in Hastie et al. (2022) with varying distributions over the features. Surprisingly, double descent is observed in linear regression, and we provided a detailed proof and a numerical simulation of the double descent phenomenon where the feature covariance matrix is isotropic. The linear isotropic feature also has deep connections to the linearization of neural networks in the infinite-width limit. **Code** for our numerical experiments: https://github.com/LinkaiMa/Math_Stats_Project

1. Introduction

Recent empirical success of over-parametrized neural networks challenged classical statistical wisdom of bias-variance tradeoff (Belkin et al., 2019; Hastie et al., 2009). Specifically, it's observed that after the interpolation threshold where the number of parameters p equals the number of data points n, the test risk begins a second-phase of decreasing, also known as "double descent" (Belkin et al., 2019).

Nakkiran et al. (2021) further shows that double descent is a robust phenomenon invariant to the choices of neural network architectures, datasets, and training methods. Moreover, double descents are results of either increased training time or growing model complexities (Nakkiran et al., 2021). Beyond neural networks, Canatar et al. (2021) also demonstrates that double descent exists in the settings of kernel regression by developing a closed-form generalization error derived using replica tricks from statistical mechanics.

In this report, we would like to go back to a more fundamental setting of linear regression in high dimensions. Specifically, we will review the work by Hastie et al. (2022), which shows the double descent phenomenon for minimum ℓ_2-norm least square regression with the over-parametrization ratio $\gamma := p/n$ for n data points and p features.

For the linear regression problem $y_i = x_i^\top \beta + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (see Section 2.1 for details), we consider the feature $x_i \in \mathbb{R}^p$ to be generated by an underlying linear model $x_i = \Sigma^{1/2} z_i$, where Σ is the feature covariance matrix and $z_i \sim \mathcal{N}(0, I_p)$, and a non-linear model $x_i = \varphi(W z_i)$, i.e. features of the random one-layer neural network. Depending on the geometry of Σ, we further divides the linear model into isotropic feature, correlated feature (equidistribute coefficients), and correlated feature (latent space model). Details on these settings can be found in Section 2 and Section 4.

Following Hastie et al. (2022), we provide a review of asymptotic and non-asymptotic risks of linear regression in each of the linear and non-linear model settings in terms of their bias and variance decomposition from Section 2 to Section 5.

Specifically, for linear generating models with isotropic features, the asymptotic risk either decreases as the over-parametrization ratio $\gamma \to \infty$ or exhibits a local minimum in the over-parametrization regime $\gamma \in (1, \infty)$ depending on the signal-to-noise ratio SNR (see Definition 2.7). Non-linear models with the generating process $x_i = \varphi(W z_i)$ also has the same asymptotic risk in the $\gamma \to \infty$ limit, connecting to the model linearization in the lazy training regime and neural tangent kernel observations in neural networks (Chizat et al., 2018; Jacot et al., 2018).

In linear models with correlated features, we observed either qualitatively similar results with isotropic features in the equidistributed coefficients case or the benefit of over-parametrization in the large $\gamma \to \infty$ limit for latent space model. All of these observations show that the double descent phenomenon passing the interpolation threshold $p = n$ is more general and universal with subtle changes based on different conditions.

Additionally, in Section 6 and Section 7 we provide our own proof of the bias and variance decomposition of min-norm regression estimator Lemma 3.1 and our own numerical experiments of asymptotic risk in the istropic feature case in Figure 2. We conclude the paper in Section 8.

Section 2 to Section 5 follows entirely from Hastie et al.

(2022) and thus citations are omitted.

2. Preliminaries

2.1. Model Assumption and Notations

Consider a total of n i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, where for all i the following holds

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon$$

$$y_i = x_i^\top \beta + \epsilon_i$$

with $x_i \sim P_x$, $x_i \in \mathbb{R}^p$. $\epsilon_i \sim P_\epsilon, \epsilon_i \in \mathbb{R}$, and $\beta \in \mathbb{R}^p$. For the distribution P_x, P_ϵ, we have $\mathbb{E}[x_i] = 0, \text{Cov}[x_i] = \Sigma$ and $\mathbb{E}[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2$. In matrix notation, we have

$$y = X\beta + \epsilon$$

where $y \in \mathbb{R}^{n\times 1}, X \in \mathbb{R}^{n\times p}, \beta \in \mathbb{R}^{p\times 1}$ and $\epsilon \in \mathbb{R}^{n\times 1}$. We consider each x_i as a row vector and β as a column vector. We denote:

$$\widehat{\Sigma_X} = \frac{1}{n}\sum_{i=1}^n x_i^T x_i, \widehat{\text{Cov}}(X, y) = \frac{1}{n}\sum_{i=1}^n x_i^T y_i$$

2.2. Linear Model

In the linear model, the feature $x_i \in \mathbb{R}^p$ are obtained according to the generating model $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ and z_i has i.i.d entries.

Definition 2.1. The feature $x_i \in \mathbb{R}^p$ is said to be isotropic if $\Sigma = I$, i.e. x_i is a vector of i.i.d entries in $\mathbb{R}^p$

Definition 2.2. The feature $x_i \in \mathbb{R}^p$ is said to be correlated if $\Sigma \neq I$.

For correlated feature x_i, the risk for the min-norm least-square regression depends on the geometry of (Σ, β). Details on the distribution over (Σ, β) are deferred to Section 4.

2.3. Non-Linear Model

Definition 2.3. The feature $x_i \in \mathbb{R}^p$ is said to be nonlinear if $x_i = \varphi(Wz_i)$ where $z_i \in \mathbb{R}^d, W \in \mathbb{R}^{p\times d}$ is a matrix of i.i.d entries and $\varphi(\cdot)$ is an element-wise nonlinear activation function.

2.4. Estimator

Definition 2.4. For a test sample $x_0 \sim P_x$ and an estimator $\hat{\beta}$ of β, the out-of-sample prediction risk (also risk) is given by $R_X(\hat{\beta}, \beta) = \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2 | X]$

Definition 2.5. The minimum ℓ_2 norm (min-norm) least square regression estimator is given by $\hat{\beta} = \arg\min_{b\in\mathbb{R}^p}\{||b||_2 : b \text{ minimizes } ||y - Xb||_2^2\}$

Definition 2.6. The ridge regression estimator is given by $\hat{\beta}_\lambda = \arg\min_{b\in\mathbb{R}^p}\{\frac{1}{n}||y - Xb||_2^2 + \lambda||b||_2^2\}$

Definition 2.7. The signal-to-noise ratio (SNR) is $\text{SNR} = \frac{r^2}{\sigma^2}$, where $r = ||\beta||_2$.

3. Linear Model - Isotropic Features

In this section, we will cover the asymptotic risk for linear models with isotropic features. To obtain the convergence statement, we first consider the decomposition of risk into bias and variance. Lemma 3.1 provides the closed-form expression.

Lemma 3.1. *Given the model assumption in Section 2.1, the risk $R_X(\hat{\beta}; \beta)$ of the min-norm regression estimator $\hat{\beta}$ satisfies the bias and variance decomposition*

$$B_X(\hat{\beta}; \beta) = \beta^\top \Pi\Sigma\Pi\beta$$

$$V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{n} Tr(\hat{\Sigma}^+\Sigma)$$

with $\hat{\Sigma} := X^\top X/n$ as the empirical covariance matrix of X and $\Pi := I - \hat{\Sigma}^+\Sigma$ as the projection onto the null space of X.

Proof. We present our proof of Lemma 3.1. See Section 6. □

Lemma 3.1 allows us to obtain the asymptotic risk for the min-norm least square regression estimator $\hat{\beta}$ under the isotropic feature and linear model assumption in Theorem 3.2.

Theorem 3.2. *Under the model assumption in Section 2.1 and Section 2.2, we assume that x_i's have independent entries with zero mean and unit variance.*

For $\gamma < 1$, assume that x_i have i.i.d. entries with finite 4th moment, and Σ is positive definite and deterministic with the smallest eigenvalue $\lambda_{\min}(\Sigma) \geq c > 0$.

For $\gamma > 1$, assume that either of the following holds

(i): $\forall i, j, k, \mathbb{E}[|(x_i)_j|^k] < C$ for some constant C

(ii): the entries for x_i's are i.i.d. and $\forall i, j, \mathbb{E}[|(x_i)_j|^{4+\delta}] < C$ for some $C \in \mathbb{R}, \delta > 0$

Then for the estimator we get from linear regression $\hat{\beta}$, as $n, p \to \infty$ with $\frac{p}{n} \to \gamma$, we have:

$$B_X(\hat{\beta}; \beta) \to \begin{cases} 0, & \text{for } \gamma < 1 \\ r^2(1 - \frac{1}{\gamma}), & \text{for } \gamma > 1 \end{cases}$$

$$V_X(\hat{\beta}; \beta) \to \begin{cases} \sigma^2\frac{\gamma}{1-\gamma}, & \text{for } \gamma < 1 \\ \sigma^2\frac{1}{\gamma-1}, & \text{for } \gamma > 1 \end{cases}$$

and

$$R_X(\hat{\beta}, \beta) \to \begin{cases} \sigma^2\frac{\gamma}{1-\gamma}, & \text{for } \gamma < 1 \\ r^2(1 - \frac{1}{\gamma}) + \sigma^2\frac{1}{\gamma-1}, & \text{for } \gamma > 1 \end{cases}$$

where $\gamma < 1$ corresponds to the under-parametrized regime and $\gamma > 1$ corresponds to the over-parametrized regime.

To interpret this theorem, notice that for the under-parametrized regime $\gamma < 1$, $\hat{\beta}$ is an unbiased estimator of β, and the risk increases as we closer to the interpolation threshold $p = n$ due to the rise in the variance. In other words, as $\gamma \to 1$, $R_X(\hat{\beta}; \beta) = V_X(\hat{\beta}; \beta) \to \infty$. The risk goes unbounded as we approach the interpolation threshold $p = n$ from the left.

For the over-parametrized regime $\gamma > 1$, $\hat{\beta}$ is not an unbiased estimator anymore. The overall risk is dependent on both the bias and the variance. This dependency can be captured by the SNR ratio r^2/σ^2. Figure 2 shows the asymptotic risk curve for different values of SNR. In the over-parametrized regime $\gamma > 1$, any SNR ≥ 1 will leads to a local minimum in the risk for $\gamma \in (1, \infty)$, and SNR ≤ 1 will result in the minimum of the risk for $\gamma \to \infty$.

4. Linear Model - Correlated Feature

4.1. Risk Approximation

Consider the same linear model setting $x_i \in \mathbb{R}^p$ in Section 2.2 but with $\Sigma \neq I$, Σ is deterministic and positive definite. Thus the features are not isotropic but correlated. The resulting risk will be shown to be dependent on the geometry of (Σ, β).

Let $\sum_{i=1}^p s_i v_i v_i^\top$ be the eigendecomposition of Σ. We can define the following empirical probability distribution over the eigenvalues s_i and the coefficient of β in the eigenbasis $\{v_i\}$ as a proxy for distribution over the pair (Σ, β):

$$\hat{H}_n(s) := \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}_{\{s \geq s_i\}}$$

$$\hat{G}_n(s) := \frac{1}{||\beta||_2^2} \sum_{i=1}^{p} \langle \beta, v_i \rangle^2 \mathbb{1}_{\{s \geq s_i\}}$$

We adopt the following assumptions:

Assumption 4.1. z_i has independent entries and $\mathbb{E}\{z_i\} = 0, \mathbb{E}\{z_i^2\} = 1, \mathbb{E}\{|z_i|^k\} \leq C_k < \infty, \forall i \leq p, k \geq 2$.

Assumption 4.2. $s_1 = ||\Sigma||_{\text{op}} \leq M, \int s^{-1} d\hat{H}_n(s) < M$

Assumption 4.3. $|1 - (p/n)| \geq 1/M, 1/M \leq p/n \leq M$

For the correlated feature, Lemma 3.1 is numerically hard to compute so we need an approximation of the bias and variance. We can define the predicted bias and variance of the min-norm regression as follows:

Definition 4.4. The predictive bias is

$$\mathscr{B}(\hat{H}_n, \hat{G}_n, \gamma) := ||\beta||_2^2 \left\{ 1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\hat{H}_n(s)} \right\} \cdot \int \frac{s}{(1+c_0\gamma s)^2} d\hat{G}_n(s)$$

Definition 4.5. The predictive variance is

$$\mathscr{V}(\hat{H}_n, \gamma) := \sigma^2 \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}$$

where c_0 is defined as follows:

Definition 4.6. Let $\gamma \in \mathbb{R}_{>0}$. $c_0 := c_0(\gamma, \hat{H}_n) \in \mathbb{R}_{>0}$ is the unique solution to

$$1 - \frac{1}{\gamma} = \int \frac{1}{1 + c_0 \gamma s} d\hat{H}_n(s)$$

Notice that $\mathscr{B}$ and $\mathscr{V}$ can be computed numerically. The following theorem provides a non-asymptotic bound on the approximation error of $\mathscr{B}$ and $\mathscr{V}$ to the true bias and variance and thus the true risk $R_X(\hat{\beta}; \beta)$.

Theorem 4.7. *Under the model assumption in Section 2.1 and Section 2.2 and Assumption 4.1, Assumption 4.2, and Assumption 4.3, we further assume that $s_p = \lambda_{\min}(\Sigma) > 1/M$. Then $\forall D > 0, \exists C = C(M, D)$ s.t. with probability at least $1 - Cn^{-D}$ we have*

$$|B_X(\hat{\beta}; \beta) - \mathscr{B}(\hat{H}_n, \hat{G}_n, \gamma)| \leq \frac{C||\beta||_2^2}{n^{1/7}}$$

$$|V_X(\hat{\beta}; \beta) - \mathscr{V}(\hat{H}_n, \gamma)| \leq \frac{C}{n^{1/7}}$$

where

$$R_X(\hat{\beta}; \beta) = B_X(\hat{\beta}; \beta) + V_X(\hat{\beta}; \beta)$$

is the bias-variance decomposition.

Theorem 4.7 can also be extended to the asymptotic regime by considering the weak convergence of $\hat{H}, \hat{G}$ to H, G.

Theorem 4.8. *Under the setting in Theorem 4.7, let $n, p \to \infty, p/n \to \gamma \in (0, \infty), \hat{H} \to H, \hat{G} \to G$. Then it holds almost surely*

$$\frac{1}{||\beta||_2^2} B_X(\hat{\beta}; \beta) \to \mathscr{B}_1(H, G, \gamma)$$

$$V_X(\hat{\beta}; \beta) \to \mathscr{V}(H, \gamma)$$

where

$$\mathscr{B}_1(H, G, \gamma) := \left\{ 1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} dH_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} dH_n(s)} \right\} \cdot \int \frac{s}{(1+c_0\gamma s)^2} dG_n(s)$$

Theorem 4.7 and Theorem 4.8 thus establish an approximation of the risk in both non-asymptotic and asymptotic regimes for non-isotropic features under the linear model $x_i = \Sigma^{1/2} z_i$ assumption. With these tools, we can analyze the risk with two different geometries of (Σ, β) encoded by $(\hat{H}, \hat{G})$.

4.2. Equidistributed Coefficients

Consider $G = H$. This is the scenario such that the parameter vector β is equidistributed in each eigenvector of Σ. It's shown in Hastie et al. (2022) that the resulting risk is qualitatively the same to the one in the case of the isotropic feature.

4.3. Latent Space Model

Now let β be aligned with the top eigenvectors of Σ. Also, let Σ be generated by $\Sigma = WW^\top + I$, where $W \in \mathbb{R}^{p \times d}$ with $d \ll p$. Hastie et al. (2022) show that in this regime larger over-parametrization generally decreases the risk further, resembling the double descent phenomenon observed in neural networks. Figure 1 copied from Hastie et al. (2022) provides an illustration. It can be seen that $\gamma \to \infty$ actually helps decrease the risk further, coinciding with observations made in over-parametrized neural networks.

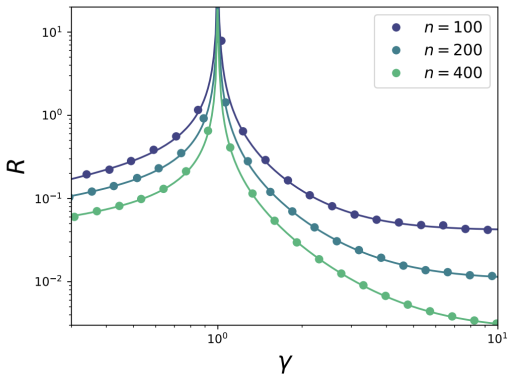

Figure 1. Risk for min-norm regression estimator $\hat{\beta}$ under the latent space model with the linear model assumption $x_i = \Sigma^{1/2} z_i$. Figure from Hastie et al. (2022)

5. Non-Linear Model

Now consider the setting of Definition 2.3, where $x_i = \varphi(W z_i)$ is the feature generated by a random one-layer neural network. The following theorem characterizes the asymptotic risk of the min-norm regression estimator $\hat{\beta}$ under the non-linear feature setting. Surprisingly, the asymptotic risk based on non-linear features is the same as the asymptotic risk of linear isotropic features in Theorem 3.2. Thus the linear isotropic feature assumption is potentially more general than we think.

Theorem 5.1. *Under the model assumption in Section 2.1 and the non-linear feature assumption $x_i = \varphi(W z_i)$, where $z_i \sim \mathcal{N}(0, I_d)$ and $W_{ij} \sim \mathcal{N}(0, 1/d)$, we further assume $|\varphi(x)| \le c_0(1 + |x|)^{c_0}$ for $c_0 > 0$ The activation function φ is also constrained to have $\mathbb{E}[\varphi(G)] = 0, \mathbb{E}[\varphi(G)^2] = 1, \mathbb{E}[G\varphi(G)] = 0$ for $G\mathcal{N}(0, 1)$.*

Then for $n, p, d \to \infty, p/n \to \gamma, d/p \to \phi \in (0, \infty), \gamma > 1$, it holds almost surely that

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} V_X(\hat{\beta}; \beta) = \frac{\sigma^2}{\gamma - 1}$$

which is the same as the variance in Theorem 3.2 for $\gamma > 1$ and linear isotropic feature. Additionally, if we assume $\mathbb{E}[\beta] = 0, Cov(\beta) = r^2 I_p/p$, the Bayes bias $B_X(\hat{\beta}_\lambda) = \mathbb{E}_\beta[B_X(\hat{\beta}_\lambda; \beta)]$ converges almost surely to

$$\lim_{\lambda \to 0^+} \lim_{n,p,d \to \infty} B_X(\hat{\beta}_\lambda) = \begin{cases} 0, & \text{for } \gamma < 1 \\ r^2(1 - \frac{1}{\gamma}), & \text{for } \gamma > 1 \end{cases}$$

which corresponds to the limiting bias in Theorem 3.2 in the over-parametrization regime.

Thus the random feature map produced by a one-layer neural network has the same asymptotic risk compared to the linear isotropic feature. This phenomenon potentially has deep connections to the neural tangent kernel regime of neural network function dynamics and the first-order Taylor approximation of infinite-wide neural network in the lazy training setting (Jacot et al., 2018; Chizat et al., 2018).

In the infinite-width limit, neural networks are asymptotically over-parametrized and can be approximated by the first-order Taylor term. Thus the risk for the linear isotropic features could be a good proxy for the risk of random features from a one-layer over-parametrized neural network.

6. Proof of Lemma 3.1

In this section, we will present of own proof of Lemma 3.1, which characterizes the bias and variances for the min-norm regression estimator $\hat{\beta}$.

Proof. The main idea behind this result is the bias-variance decomposition as we have seen in Section 7.2 of the class. This time, we are considering different definition of risk.

Notice that in class, the risk is defined to be $R(\hat{\beta}, \beta) = \mathbb{E}[||\hat{\beta} - \beta||^2 | X]$. But in this paper, $R_X(\hat{\beta}, \beta)$ is in fact equal

to $\mathbb{E}[||\hat{\beta}-\beta||_\Sigma^2|X]$, where $\forall z, ||z||_\Sigma^2 := z^T\Sigma z$

$$\begin{aligned}R_X(\hat{\beta},\beta) &= \mathbb{E}[(x_0^T\hat{\beta}-x_0^T\beta)^2|X]\\ &= \mathbb{E}[(x_0^T(\hat{\beta}-\beta))^2|X]\\ &= \mathbb{E}[(\hat{\beta}-\beta)^T x_0 x_0^T(\hat{\beta}-\beta)|X]\end{aligned}$$

Notice that this expectation is taken over x_0, the additional data sample and ϵ, the noises, by Fubini:

$$\begin{aligned}R_X(\hat{\beta},\beta) &= \mathbb{E}_{x_0,\epsilon}[(\hat{\beta}-\beta)^T x_0 x_0^T(\hat{\beta}-\beta)|X]\\ &= \mathbb{E}_\epsilon[\mathbb{E}_{x_0}[(\hat{\beta}-\beta)^T x_0 x_0^T(\hat{\beta}-\beta)|X]]\\ &= \mathbb{E}_\epsilon[(\hat{\beta}-\beta)^T\mathbb{E}_{x_0}[x_0x_0^T](\hat{\beta}-\beta)|X]\\ &= \mathbb{E}_\epsilon[(\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)|X]\\ &= \mathbb{E}[(\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)|X]\\ &= \mathbb{E}[||\hat{\beta}-\beta||_\Sigma^2|X]\end{aligned}$$

We then follow the same procedure for bias-variance decomposition as we did in class, but this time we are not assuming $\widehat{\Sigma_X}$ is invertible.

Our regression equation is:

$$X\beta = y$$

multiplying X^T on both the left and right, we obtain the normal equation:

$$X^TX\beta = X^Ty$$

This is not in general solvable, so we use the two norm minimizer for the difference of the left and the right hand side of the equation, which gives as the estimator:

$$\hat{\beta} = (X^TX)^\dagger X^Ty = \widehat{\Sigma_X}^\dagger\widehat{\text{Cov}}(X,y)$$

where † is the pseudoinverse.

We also notice that :

$$\begin{aligned}||x+y||_\Sigma^2 &= (x^T+y^T)\Sigma(x+y)\\ &= x^T\Sigma x + x^T\Sigma y + y^T\Sigma x + y^T\Sigma y\\ &= ||x||_\Sigma^2 + 2x^T\Sigma y + ||y||_\Sigma^2\end{aligned}$$

where the last step is justified because Σ is a Covariance matrix, hence symmetric. By taking the transpose, $x^T\Sigma y = y^T\Sigma x$.

We then try to introduce the expectation of our estimator $\hat{\beta}$ into the risk and decompose the risk into bias and variance.

$$\begin{aligned}R_X(\hat{\beta},\beta) &= \mathbb{E}[||\hat{\beta}-\beta||_\Sigma^2|X]\\ &= \mathbb{E}[||\beta-\mathbb{E}[\hat{\beta}|X]||_\Sigma^2|X] + \mathbb{E}[||\hat{\beta}-\mathbb{E}[\hat{\beta}|X]||_\Sigma^2|X]\\ &\quad - 2\mathbb{E}[(\beta-\mathbb{E}[\hat{\beta}|X])^T\Sigma(\hat{\beta}-\mathbb{E}[\hat{\beta}|X])|X]\end{aligned}$$

We denote the first term as $B_X(\hat{\beta},\beta)$, the second term as $V_X(\hat{\beta},\beta)$. The third term vanishes since both β and $\mathbb{E}[\hat{\beta}|X]$ are constants.

We also notice that:

$$\begin{aligned}&B_X(\hat{\beta},\beta)\\ &= \mathbb{E}[||\beta-\mathbb{E}[\hat{\beta}|X]||_\Sigma^2|X]\\ &= ||\beta-\mathbb{E}[\hat{\beta}|X]||_\Sigma^2 \qquad (1)\end{aligned}$$

$$\begin{aligned}&V_X(\hat{\beta},\beta)\\ &= \mathbb{E}[(\hat{\beta}^T-\mathbb{E}[\hat{\beta}|X]^T)\Sigma(\hat{\beta}-\mathbb{E}[\hat{\beta}|X])|X]\\ &= \mathbb{E}[(\hat{\beta}^T-\mathbb{E}[\hat{\beta}|X]^T)\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}(\hat{\beta}-\mathbb{E}[\hat{\beta}|X])|X]\\ &= \mathbb{E}[||\Sigma^{\frac{1}{2}}(\hat{\beta}-\mathbb{E}[\hat{\beta}|X])||_2^2|X]\\ &= \mathbb{E}[\text{Tr}((\Sigma^{\frac{1}{2}}(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))(\Sigma^{\frac{1}{2}}(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))^T)|X] \qquad (2)\\ &= \mathbb{E}[\text{Tr}(\Sigma^{\frac{1}{2}}(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))^T\Sigma^{\frac{1}{2}})|X]\\ &= \mathbb{E}[\text{Tr}((\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))^T\Sigma)|X] \qquad (3)\\ &= \text{Tr}(\mathbb{E}[(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))^T\Sigma|X]) \qquad (4)\\ &= \text{Tr}(\mathbb{E}[(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))(\hat{\beta}-\mathbb{E}[\hat{\beta}|X]))^T|X]\Sigma)\\ &= \text{Tr}(\text{Cov}[\hat{\beta}|X]\Sigma) \qquad (5)\end{aligned}$$

In (2), we used the same fact in the lecture notes: $||v||_2^2 = \text{Tr}(vv^T)$. In (3), we used the cyclic property of trace. In (4), we used the linearity of trace.

We try to further simplify our expressions for the bias and variance terms.

$$\begin{aligned}&\mathbb{E}[\hat{\beta}|X]\\ &= \mathbb{E}[\widehat{\Sigma_X}^\dagger\widehat{\text{Cov}}(X,y)|X]\\ &= \mathbb{E}[\widehat{\Sigma_X}^\dagger\frac{1}{n}\sum_{i=1}^n x_i^Ty_i|X]\\ &= \mathbb{E}[\widehat{\Sigma_X}^\dagger\frac{1}{n}\sum_{i=1}^n x_i^T(x_i\beta+\epsilon_i)|X]\\ &= \mathbb{E}[\widehat{\Sigma_X}^\dagger\frac{1}{n}\sum_{i=1}^n x_i^Tx_i\beta|X] + \mathbb{E}[\widehat{\Sigma_X}^\dagger\frac{1}{n}\sum_{i=1}^n x_i^T\epsilon_i|X]\\ &= \widehat{\Sigma_X}^\dagger\widehat{\Sigma_X}\beta + \widehat{\Sigma_X}^\dagger\mathbb{E}[\frac{1}{n}\sum_{i=1}^n x_i^T\epsilon_i]\\ &= \widehat{\Sigma_X}^\dagger\widehat{\Sigma_X}\beta \qquad (6)\end{aligned}$$

Let's denote $\Pi = I - \widehat{\Sigma_X}^\dagger\widehat{\Sigma_X}$, then the bias is:

$$\mathbb{E}[\hat{\beta}|X] - \beta = -\Pi\beta$$

Notice that Π is the projection onto the kernel of $\widehat{\Sigma_X}$, which is also the kernel of X. This is justified by the following:

$$\forall v \in \text{Ker}(X), \widehat{\Sigma_X} v = \frac{1}{n} X^T (Xv) = 0$$

$$\forall v \in \text{Ker}(\widehat{\Sigma_X}), ||Xv||_2^2 = v^T X^T X v = v^T \frac{1}{n} \widehat{\Sigma_X} v = 0$$

Therefore, when the row space of X is large, the kernel of X is small and our expected bias should be small.

By plugging into (1): $B_X(\hat{\beta}, \beta) = \beta^T \Pi^T \Sigma \Pi \beta$

Since Π is a projection matrix, it is symmetric. Then

$$B_X(\hat{\beta}, \beta) = \beta^T \Pi \Sigma \Pi \beta \tag{7}$$

For the variance part, notice that:

$$\begin{aligned}
&\text{Cov}[\hat{\beta}|X] \\
&= \text{Cov}[(X^T X)^\dagger X^T y|X] \\
&= \text{Cov}[(X^T X)^\dagger X^T (X\beta + \epsilon)|X] \\
&= \text{Cov}[(X^T X)^\dagger X^T \epsilon] \\
&= (X^T X)^\dagger X^T \text{Cov}[\epsilon] X ((X^T X)^\dagger)^T && (8) \\
&= (X^T X)^\dagger X^T \text{Cov}[\epsilon] X ((X^T X)^T)^\dagger && (9) \\
&= (X^T X)^\dagger X^T \text{Cov}[\epsilon] X (X^T X)^\dagger \\
&= \sigma^2 (X^T X)^\dagger X^T X (X^T X)^\dagger \\
&= \sigma^2 (X^T X)^\dagger && (10) \\
&= \sigma^2 (n\widehat{\Sigma_X})^\dagger \\
&= \frac{\sigma^2}{n} \widehat{\Sigma_X}^\dagger
\end{aligned}$$

In (8) we are using the stability of multivariate Gaussian after affine transformation. In (9,10) we are using the properties of pseudoinverse: $(A^\dagger)^T = (A^T)^\dagger$ and $A^\dagger A A^\dagger = A^\dagger$

Plugging the above into (5), we get:

$$V_X(\hat{\beta}, \beta) = \frac{\sigma^2}{n} \text{Tr}(\widehat{\Sigma_X}^\dagger \Sigma) \tag{11}$$

Notice that in the isotropic case, the above agrees with our lecture notes. □

7. Numerical Experiments

We aim to recover our previous analysis of the asymptotic from numerical experiments on the isotropic case. We first tried to recover the simulation results from the paper, i.e. Figure 2. We used the same parameters as the authors of this paper. $\sigma = 1, n = 200, p = \lfloor \gamma n \rfloor$. The X in our experiment also has i.i.d. entries generated from $\mathcal{N}(0, 1)$. We test different values of SNR and for each SNR, we generated β from multinormal distribution and scaled properly according to the SNR. Then for different values of γ, we calculated the numerical risk for 50 samples and took the average as our risk. We also plotted the theoretical risk curve, the results of Theorem 1. The null risk, i.e. the risk when $\hat{\beta} = 0$ is also plotted as a dotted horizontal line.

We have successfully recovered Figure 2 in the paper:

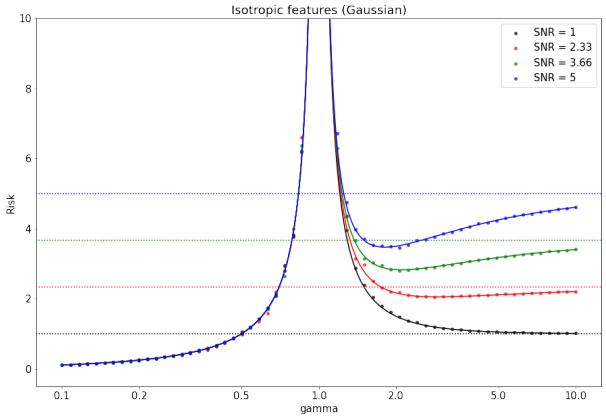

Figure 2. Numerical simulation of the asymptotic risk curve for the min-norm least square estimator $\hat{\beta}$ under the linear model with isotropic features. The x-axis represents the over-parameterization ratio $\gamma := p/n$. Here SNR:= r^2/σ^2.

In addition, we repeated the simulation for X with i.i.d. entries generated by the Rademacher distribution, i.e. $+1$ or -1 with equal probability. The Rademacher distribution also has zero mean but has variance $= \frac{1}{4}$, which doesn't completely satisfy the requirement for Theorem 1. However, our numerical results still converged to the same curve. We then tried to loosen the i.i.d. assumption. We first tried to use some repeated rows for X. We then tried to add a small perturbation, $h\mathcal{N}(0, I)$ on the repeated rows, where $h = 0.1$.

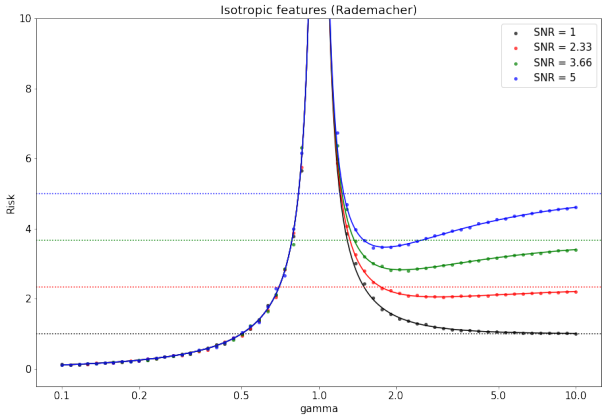

Figure 3. Numerical simulation for X with i.i.d. entries generated from the Rademacher distribution.

It turns out that using dependent rows on our simulations has a rather small effect on the results when the portion of the dependent rows is small. However, when we add a small perturbation to these rows, destroying the "identically distributed" assumption, our results have completely deviated the asymptotic results from Theorem 1.

Interestingly, the deviation of the results were mostly in the region where $\gamma > 1$. In the $\gamma < 1$ region, the deviation is relatively minor. This is because in the paper, the asymptotic results used different assumptions for $\gamma < 1$ and for $\gamma > 1$. The resutls for $\gamma < 1$ comes from random matrix theory and seems to be more universal.

See below for the simulation results:

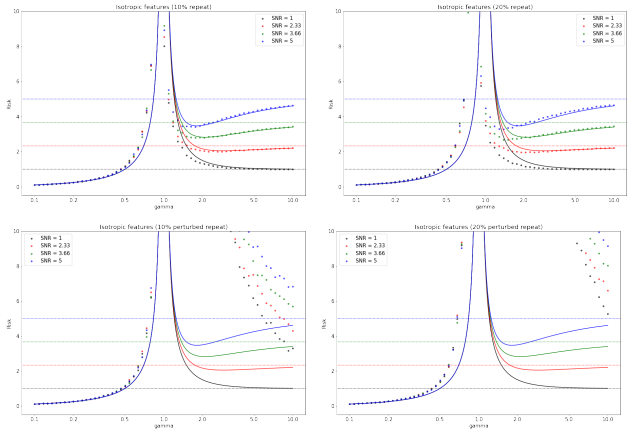

Figure 4. X with entries from standard Gaussian. Using 10%/ 20% repeated rows. Adding/without a samll Gaussian perturbation to X

8. Discussion and Conclusion

In this report, we review key results from Hastie et al. (2022) and show that double descent exists for linear regression in high-dimensions. Specifically, we discuss the asymptotic and non-asymptotic risks of the min-norm regression estimators in terms of the bias and variance decompositions with linear isotropic features, linear correlated features, and non-linear features. We are able to provide a detailed proof of the bias and variance terms of the min-norm regression estimator and reproduce the experiments on linear isotropic features from the paper.

Interestingly, the asymptotic risk for linear models with isotropic features is the same compared to non-linear features. In neural network literature, it's known that in the asymptotic infinite-width limit, the neural network can be linearized by a first-order Taylor expansion and the training dynamics under gradient descent can be captured by the neural tangent kernel (Chizat et al., 2018; Jacot et al., 2018). Such a coincidence in the asymptotic risk thus should not be an accident, and there are deep theoretical connections to be explored.

Another revelation from Hastie et al. (2022) not included in this report is that under ridge regularization, the double descent phenomenon dissolves as the λ parameter in the norm regularization term increase. The existence of double descent in neural networks thus implies that we haven't found the optimal regularization for neural networks yet (Grosse, 2021). A more interesting question will be that is regularization even necessary. Hastie et al. (2022) further shows that under suitable conditions the risk of the min-norm regression estimator is lower than that of the ridge regression estimator.

Overall, the study of double descent in linear regression opens up new avenues to be pursued and there are more interesting questions to be asked in the ensuing decades.

Acknowledgements

Yilun Kuang (yk2516@nyu.edu) and Linkai Ma (lm4307@nyu.edu) have equal contributions to this project. Yilun Kuang provides the review and Linkai Ma presents the proof derivations and numerical experiments. Yilun Kuang and Linkai Ma thank Prof. Jonathan Niles-Weed and TA Aram-Alexandre Pooladian for the NYU MATH-GA 2830 Mathematical Statistics class and helpful suggestions.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

Canatar, A., Bordelon, B., and Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. 2018. doi: 10.48550/ARXIV.1812.07956. URL https://arxiv.org/abs/1812.07956.

Grosse, R. A toy model: Linear regression. 2021. URL https://www.cs.toronto.edu/~rgrosse/courses/csc2541_2021/readings/L01_intro.pdf.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, dec 2021. doi: 10.1088/1742-5468/ac3a74. URL https://dx.doi.org/10.1088/1742-5468/ac3a74.