universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Beyond the Bias-Variance Trade-off: Exploring Double Descent in Statistical Modeling

verfasst von | submitted by

Abdelbast Nassiri

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von | Supervisor:	Assoz. Prof. Dr. Philipp Christian Petersen M.Sc.

Abstract

Das Double-Descent-Phänomen stellt den klassischen Bias-Variance-Trade-off infrage, indem es zeigt, dass eine Erhöhung der Modellkomplexität über den Interpolationspunkt hinaus in einigen Fällen zu einer verbesserten Generalisierungsleistung führen kann. Obwohl dieses Phänomen in den letzten Jahren intensiv untersucht wurde, bleibt die Literatur fragmentiert, wobei Erkenntnisse über verschiedene Modelle und Settings verstreut sind. Diese Arbeit bietet eine umfassende Synthese der bestehenden Forschung zum Double-Descent-Phänomen und präsentiert eine einheitliche Perspektive auf sein Verhalten in verschiedenen statistischen Modellen.

Wir beginnen mit einer systematischen Übersicht über Double Descent in linearen Regressionsmodellen, wobei wir die Rolle von Überparametrisierung, Regularisierung und das Zusammenspiel zwischen Modellkomplexität und Stichprobengröße hervorheben. Anschließend untersuchen wir Double Descent in Random-Feature-Modellen, wobei wir Erkenntnisse über den Einfluss der Feature-Dimensionalität und den Übergang von unterparametrisierten zu überparametrisierten Regimes zusammenfassen. Danach wenden wir uns neuronalen Netzen (NNs) zu und fassen empirische und theoretische Ergebnisse zusammen, die Double Descent mit Optimierungsdynamiken, impliziter Regularisierung und den einzigartigen Eigenschaften von Deep-Learning-Modellen in Verbindung bringen.

Schließlich geben wir einen Überblick über Techniken zur Minderung des Double-Descent-Effekts, darunter Ridge-Regression, Early Stopping und Ensemble-Methoden. Durch die Strukturierung und Verfeinerung bestehender Ergebnisse bieten wir einen klaren Leitfaden für Praktiker, die sich den Herausforderungen überparametrisierter Modelle stellen müssen.

Diese Arbeit zielt nicht darauf ab, neue theoretische oder empirische Beiträge zu liefern, sondern dient als strukturierte und zugängliche Synthese des aktuellen Wissensstands zum Double-Descent-Phänomen. Durch die Konsolidierung und Klärung bestehender Arbeiten wollen wir Lücken in der Literatur schließen und eine Grundlage für zukünftige Forschungen in diesem sich schnell entwickelnden Bereich schaffen.

Acknowledgment

I would like to express my heartfelt gratitude to everyone who has supported me throughout the journey of completing this master's thesis.

First and foremost, I am deeply indebted to my supervisor, **Dr. Philipp Christian Petersen**, for his invaluable guidance, constructive feedback, and continuous encouragement. His expertise and patience have been instrumental in shaping the direction and completion of this work.

My sincere appreciation goes to my colleagues and peers at the university of Vienna, whose collaboration and support made this academic journey more enriching and enjoyable.

I am grateful to my family and friends for their unwavering support and understanding during the challenging times of this endeavor. Their belief in me has been my anchor throughout.

Lastly, I acknowledge the resources and infrastructure provided by the university of Vienna, which made this thesis possible.

This work is dedicated to all those who have inspired and motivated me during this process.

Contents

1 Introduction

Machine learning has become an indispensable tool in solving complex problems across various domains, from medical diagnostics to autonomous systems. Central to its success is the ability to build models that generalize well—that is, to perform accurately on unseen data. Achieving this requires balancing the trade-offs inherent in model complexity, a challenge traditionally framed by the bias-variance trade-off.

The bias-variance trade-off provides a foundational understanding of generalization. It posits that models with low complexity tend to underfit, leading to high bias but low variance, while overly complex models tend to overfit, resulting in low bias but high variance. This framework suggests an optimal level of complexity where these two sources of error are minimized, yielding the best generalization performance. However, modern machine learning practices, particularly with highly overparameterized models like deep neural networks, often defy this classical intuition.

A striking phenomenon that emerges in this context is Double Descent (DD). Unlike the traditional U-shaped curve of the bias-variance trade-off, the test error in double descent exhibits a non-monotonic behavior. Beyond the interpolation threshold — where the model perfectly fits the training data— the test error decreases again, revealing a second descent. This unexpected behavior has challenged conventional wisdom and has significant implications for both theory and practice in machine learning.

While observations of similar behavior existed previously (Vallet and Refregier 1989), the work by (Belkin, Hsu, Ma, et al. 2019) popularized the term "Double Descent". This influential paper offered a new perspective, suggesting that DD may bridge the gap between classical statistical learning theory and the practical approaches used in modern machine learning.

The discovery of DD sparked a surge in research activity. Two main areas of exploration emerged:

- Expanding the Scope of DD: Studies like those by (Nakkiran, Kaplun, et al. 2019) have confirmed

the presence of DD behavior in common deep learning networks. They've also shown that DD can occur along other axes, such as the amount of training data or the number of training epochs.

- Unveiling the Mechanisms: Researchers like (Hastie et al. 2020), (Belkin, Hsu, and Xu 2020) have focused on theoretical explanations for DD. Their work primarily explores the phenomenon in linear or kernel regression settings, aiming to understand the fundamental mechanisms behind it.

This thesis provides a comprehensive exploration of the DD phenomenon in machine learning, examining its behavior across various models and settings to deepen our understanding of its implications for generalization. We begin by analyzing DD in linear regression 3, where we investigate how risk evolves with increasing model complexity, particularly in the over-parameterized regime, and explore the role of regularization in shaping the risk curve. Next, we study DD in random features models 4, a simplified framework that bridges classical and modern machine learning, to understand how feature dimensionality and interpolation thresholds influence the risk behavior. We then extend our analysis to neural networks 5, empirically and theoretically examining how does the width of neural networks contribute to DD, offering insights into the success of over-parameterized models in practice. Finally, we discuss techniques to mitigate double descent 7, exploring strategies such as regularization, early stopping, and ensembling to control risk and improve generalization.
By systematically studying DD across these models and proposing mitigation strategies, this thesis aims to provide a holistic understanding of this phenomenon and its implications for the design and training of modern machine learning systems.

2 Statistical learning theory

2.1 Foundations of statistical learning theory

In statistical learning theory, the supervised learning task focuses on deriving an effective predictor $h_n : \mathcal{X} \longrightarrow \mathcal{Y}$ using a training dataset $\mathcal{D}_n$. This dataset $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n))\}$ comprises $n \in \mathbb{N}$ independent and identically distributed (i.i.d) samples of random variables, which are defined over the space $\mathcal{X} \times \mathcal{Y}$ and are generated according to the distribution $\mathcal{D}$. The objective is to select a predictor h from a predefined set of possible predictors, i.e. Hypothesis set $\mathcal{H}$. This selection process aims to ensure that h accurately predicts the outputs Y for new inputs X by effectively capturing the underlying distribution $\mathcal{D}$.

Definition 2.1 *(Loss function) A loss function is a function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ that is used to measure the distance between the predicted and true label.*

Examples of loss functions.
Zero-one loss: $L_{0,1}(y, y') = 1_{y' \neq y}$ for $\mathcal{Y} = \{0, 1\}$.
Squared loss: $\mathcal{Y} = \mathbb{R}^k$ $L(y, y') = \|y - y'\|^2$.
Binary cross entropy: $\mathcal{Y} = [0, 1]$ $L(y, y') = -(y \log(y') + (1 - y) \log(1 - y'))$.

In this thesis, we will focus on using the squared loss function.

Definition 2.2 *(Generalization Error) Let L be a loss function on $\mathcal{Y} \times \mathcal{Y}$. Let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ and let $h \in \mathcal{H}$. The risk of h is defined by*

$$\mathcal{R}_L(h) = \mathbb{E}_{\mathcal{D}}(L(h(x), y))$$

In practice, we cannot compute the generalization error $\mathcal{R}_L(h)$ since we do not know the distribution $\mathcal{D}$, instead we can compute the error on a sample $\mathcal{D}_n$.

Definition 2.3 *(Empirical Risk) Let L be a loss function on $\mathcal{Y} \times \mathcal{Y}$. Let $h \in \mathcal{H}$, and $\mathcal{D}_n := (x_i, y_i)_{i=1}^n$ be a training sample. The empirical risk is defined as*

$$\hat{\mathcal{R}}_{\mathcal{D}_n, L}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)$$

Empirical risk minimization (ERM) is an algorithm that chooses a hypothesis with smallest empirical risk. Given a hypothesis set $\mathcal{H}$ and a sample $\mathcal{D}_n$, ERM returns the solution to the optimization problem $h_{\mathcal{D}_n}^{ERM} = \arg\min_{h \in \mathcal{H}} \hat{\mathcal{R}}_{\mathcal{D}_n, L}(h)$.

2.2 Bias - Variance trade-off

Assume that $y = h(x) + \epsilon$, with $\mathbb{E}[\epsilon] = 0, Var(\epsilon) = \sigma^2$. Then, for any $x \in \mathbb{R}^d$, the expected error of a predictor h_n obtained through ERM using the squared loss function on a random dataset $\mathcal{D}_n$ can be decomposed into

$$\begin{aligned}
\mathbb{E}_x[(y - h_n(x))^2] &= \mathbb{E}_x[(h(x) + \epsilon - h_n(x))^2] \\
&= \mathbb{E}\left[(h_n(x) - h(x))^2\right] + 2\mathbb{E}\left[(h_n(x) - h(x))\epsilon\right] + \mathbb{E}[\epsilon^2] \\
&= \mathbb{E}\left[(h_n(x) - h(x))^2\right] + \sigma^2 \\
&= \mathbb{E}\left[(h_n(x) - \mathbb{E}[h_n(x)] + \mathbb{E}[h_n(x)] - h(x))^2\right] + \sigma^2 \\
&= \mathbb{E}\left[(h_n(x) - \mathbb{E}[h_n(x)])^2 + 2(h_n(x) - \mathbb{E}[h_n(x)])(\mathbb{E}[h_n(x)] - h(x)) + (\mathbb{E}[h_n(x)] - h(x))^2\right] + \sigma^2 \\
&= \underbrace{(\mathbb{E}[h_n(x)] - h(x))^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(h_n(x) - \mathbb{E}[h_n(x)])^2\right]}_{\text{Variance}} + \sigma^2.
\end{aligned}$$

Where we used for the last equality that $\mathbb{E}[h_n(x) - \mathbb{E}[h_n(x)]] = 0$, and that $h(x)$ is not random, but fixed deterministic function of x.

The "Bias-Variance" trade-off: we want to choose h_n to balance between reducing the bias term and the variance term in order to make the lowest squared loss. If we choose to increase the complexity of the model h_n (i.e., one that can closely approximate the true function h) it becomes more flexible and sensitive to the data, which allows us to reduce the bias term. However, this increased flexibility means the h_n can easily capture the random noise in the training data which causes an increase in the variance

term. The bottom line is to strike a balance between bias and variance.

Classical statistic learning focuces on finding the sweet spot between bias and variance to avoid under-fitting and over-fitting, with model selection is often guided by the U-shaped curve of the generalization error.
In modern statistical learning, however, very large models with enough parameters to achieve near-zero training error are commonly used. These models can perfectly fit (or "interpolate") the training data and, intriguingly, still generalize well. This behavior challenges the traditional U-shaped view and is part of what is known as the "Double Descent" phenomenon. In DD, error initially follows the classical pattern but then decreases again as model complexity further increases, allowing over-parameterized models to generalize effectively even when they reach the interpolation threshold.

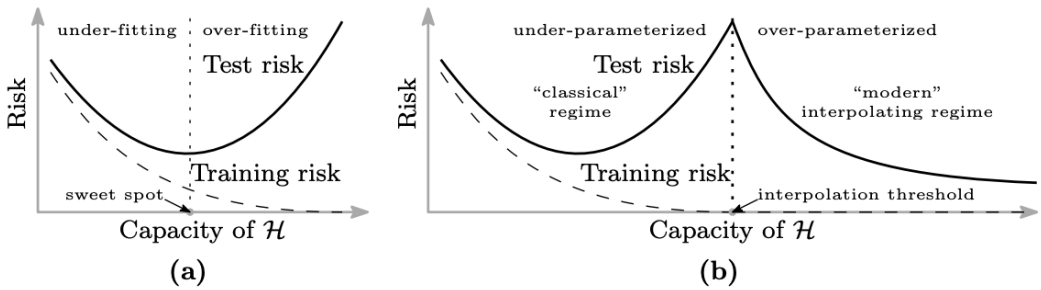

Figure 1: (a) The classical U-shaped risk curve arising from the bias-variance trade-off. (b) The DD risk curve.
Taken from (Belkin, Hsu, Ma, et al. 2019).

3 Double descent in linear regression

3.1 Setting

In this section we consider the hypothesis set $\mathcal{H}_d$ formed by linear functions. $\mathcal{H}_d$ is the set of functions $h : \mathbb{R}^d \longrightarrow \mathbb{R}$ of the form:

$$h(x) = x^\top \theta, \quad \text{with } \theta \in \mathbb{R}^d$$

Let $x \in \mathbb{R}^d$, $\epsilon \in \mathbb{R}$ be independent random variables with x drawn from some distribution P_x and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We define the random variable $y = h^*(x) + \epsilon$, with $h^* \in \mathcal{H}_d$. Given n samples $(x_i, y_i)_{i=1}^n$ from an (unknown) distribution $\mathcal{D}$, we want to learn a linear model $h(x) = x^\top \hat{\theta}$ from $\mathcal{H}_d$ for estimating y given x. That is, we want to find h that solves the problem

$$\min_{h \in \mathcal{H}_d} \mathbb{E}_{\mathcal{D}}((h(x) - y)^2) \tag{1}$$

Using empirical risk minimization we can write this problem as

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \|X\theta - y\|^2 \tag{2}$$

with $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, for simplicity we assume that X has full rank.
The solution for this optimization problem is given by

$$\hat{\theta} = X^\dagger y = \begin{cases} (X^\top X)^{-1} X^\top y & n > d \\ X^\top (X X^\top)^{-1} y & n \leq d \end{cases},$$

with the symbol † denoting the Moore-Penrose pseudo-inverse.
When $n \geq d$, we are in the "under-parameterized" regime, and there is a unique minimizer. In this case, the design matrix X has full column rank (since X is supposed to be full rank), ensuring that the normal equations have a unique solution.
When $n < d$, we are in the "over-parameterized" regime, and there are infinitely many minimizers. Since X is full row rank, these minimizers interpolate the data perfectly. In this regime, the inductive bias is used toward finding the solution with the smallest ℓ_2 norm, i.e., the minimum norm solution. This means that the solution can be written as

$$\hat{\theta} = X^\dagger y = \begin{cases} \arg\min_\theta \frac{1}{2} \|X\theta - y\|^2 & n \geq d \\ \arg\min_{\theta : X\theta = y} \|\theta\|^2 & n < d \end{cases} \tag{3}$$

For the hypothesis set $\mathcal{H}_d$ of linear models, the number of parameters is tied to the number of features and there is no natural way to adjust the capacity of the model without simultaneously adjusting the data distribution. To analyze the DD in this context, we employ two distinct approaches. The first approach involves performing feature selection by choosing a number p from $\{0, 1, ..., d\}$ of features such that $p < n$ and then gradually increasing p. In the second approach, we keep the feature-to-sample ratio constant $\frac{d}{n} \to \gamma$ as $n, d \to \infty$.

3.2 Isotropic Gaussian features

We consider the linear regression problem from the last setting with isotropic Gaussian covariates, that is $x \sim \mathcal{N}(0, I_d)$.

3.2.1 Feature selection

Let $p \in \{0, ..., d\}$ for d arbitrary large ($d >> n$) and consider the sub-problem

$$\min_{\theta_P \in \mathbb{R}^p} \frac{1}{2} \|X_P \theta_P - y\| \tag{4}$$

with $P \subset \{1, ..., d\}$ with cardinality p and $X_P \in \mathbb{R}^{n \times p}$ being the design matrix using only indices in P. Then $\hat{\theta}$ is given by

$$\hat{\theta}_p = X_P^{\dagger} y, \quad \hat{\theta}_{P^c} = 0. \tag{5}$$

The following theorem is the main theorem in (Belkin, Hsu, and Xu 2020) and gives an expression for the generalization of the minimum norm solution $\hat{\theta}$ restricted to the set of indices P.

Theorem 3.1 *(Belkin, Hsu, and Xu 2020) Let $(x, \epsilon) \in \mathbb{R}^d \times \mathbb{R}$ independent random variables with $x \sim \mathcal{N}(0, I)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\theta \in \mathbb{R}^d$. We assume that the response variable y is defined as $y = x^{\top}\theta + \epsilon$. Pick any $p \in \{0, ..., d\}$ and $P \subset \{1, ..., d\}$ of cardinality p. Let X_P be the selected P columns sub-matrix of X, and define $\hat{\theta}$ with $\hat{\theta}_P = X_P^{\dagger} y$ and $\hat{\theta}_{P^c} = 0$. The generalization error or risk of the predictor associated to $\hat{\theta}$ is:*

$$\mathbb{E}_{x,y}((y - x^{\top}\hat{\theta})^2) = \begin{cases} (\|\theta_{P^c}\|^2 + \sigma^2)(1 + \frac{p}{n-p-1}) & if\ p \leq n-2 \\ \infty & if\ n-1 \leq p \leq n+1 \\ \|\theta_P\|^2(1 - \frac{n}{p}) + (\|\theta_{P^c}\|^2 + \sigma^2)(1 + \frac{n}{p-n-1}) & if\ p \geq n+2 \end{cases} \tag{6}$$

Proof: See the proof in section 8.1.

Corollary 3.2 *Let P be a uniformly random subset of $\{1, ..., d\}$ of cardinality p. Under the setting of theorem 3.1 and taking the expectation with respect to P. The risk of the predictor associated to $\hat{\theta}$ is:*

$$\mathbb{E}_{x,y}((y - x^{\top}\hat{\theta})^2) = \begin{cases} ((1 - \frac{p}{d})\|\theta\|^2 + \sigma^2)(1 + \frac{p}{n-p-1}) & p \leq n-2 \\ \|\theta\|^2(1 - \frac{n}{d}(2 - \frac{d-n-1}{p-n-1})) + \sigma^2(1 + \frac{n}{p-n-1}) & p \geq n+2 \end{cases} \tag{7}$$

Proof: Since P is a uniformly random subset of $[d]$ of cardinality p, we have

$$\mathbb{E}(\|\theta_P\|^2) = \frac{p}{d}\|\theta\|^2, \quad \mathbb{E}(\|\theta_{P^c}\|^2) = (1 - \frac{p}{d})\|\theta\|^2.$$

Plugging into 3.1 completes the proof.

Assuming that $d > n + 1$ we observe the following:
In the under-parameterized regime ($p \leq n - 2$) the risk is given by $\mathbb{E}_{x,y}((y - x^{\top}\hat{\theta})^2) = ((1 - \frac{p}{d})\|\theta\|^2 + \sigma^2)(1 + \frac{p}{n-p-1})$, where we notice an increase in the risk up to the interpolation point ($p = n$) with the minimum at the point $p = 0$.
In the over-parameterized regime ($p \geq n + 2$), the risk decreases with p and reaches the minimum at the point $p = d$. We also notice that if

$$\|\theta\|^2 + \sigma^2 < \|\theta\|^2(1 - \frac{n}{d}) + \sigma^2(1 + \frac{d}{n-d-1})$$

then

$$\frac{\|\theta\|^2}{\sigma^2} > \frac{d}{d-n-1}$$

which means that if the signal-to-noise ratio $\frac{\|\theta\|^2}{\sigma^2}$ is big enough, the global minimum is achieved in the over-parameterized regime at $p = d$.

Let us address why the risk peaks at the interpolation point $n = p$. When $n < d$, there are many interpolating estimators, i.e., many solutions $\{\hat{\theta} : X\hat{\theta} = y\}$, including ones with small norms. This implies that we can find estimators that fit the data well without overfitting the noise, due to the abundance of degrees of freedom. In contrast, when $n = d$ (the interpolation point), there is exactly one interpolating estimator. However, this estimator typically has a high norm because it must fit not only the underlying signal but also the noise present in the data. The unique solution at this point resulting in a model of higher complexity that captures the noise, leading to a high risk.

3.2.2 Constant parameterization rate

Another interesting case, common practice in literature (Hastie et al. 2020), (Mei and Montanari 2019) considers the asymptotic setting where $n, d \to \infty$ while maintaining a constant ratio $\frac{d}{n} \to \gamma \in]0, \infty[$. The generalization error for finite d, n and deterministic P satisfies (6). However in the infinite case there is no feature selection, resulting in $p = d$. Consequently, $P^c = \emptyset$, leading to $\|\theta_{P^c}\| = 0$, passing to the limit we get.

$$(1 + \frac{p}{n-p-1}) \to (1 + \frac{\gamma}{1-\gamma}) = \frac{1}{1-\gamma}$$

Hence

$$\mathbb{E}_{x,y}((y - x^\top \hat{\theta})^2) = \begin{cases} \sigma^2 \frac{1}{1-\gamma} & if\ \gamma < 1 \\ \|\theta\|^2(1 - \frac{1}{\gamma}) + \sigma^2 \frac{\gamma}{\gamma-1} & if\ \gamma > 1 \end{cases} \tag{8}$$

On $]0, 1[$ the generalization error is dictated only by the variance and it increases with γ, and it is minimum when $\gamma = 0$. When $\gamma > 1$, the generalization error is given by

$$\|\theta\|^2(1 - \frac{1}{\gamma}) + \sigma^2 \frac{\gamma}{\gamma - 1} > \sigma^2$$

This means that the minimal generalization error is always attained in the under-parametrized regime.

3.2.3 Empirical Results

In this section, we verify the theoretical results presented earlier. To do so, we perform a Monte Carlo simulation with $M = 100$ repetitions. Our setup uses $n = 50$ samples and a feature dimension $d = 100$, with the number of active features from P with $|P| = p$ progressively increased from 1 to 100. The

experiment follows these steps:

- Generate the Design Matrix: Create the design matrix $X \in \mathbb{R}^{n \times d}$ with rows drawn independently from a standard normal distribution, $\mathcal{N}(0, I_d)$.
- Generate Noise: Draw noise values ϵ from a normal distribution $N(0, \sigma^2)$. Compute True Labels: Define the true labels using the model $Y = X\theta^* + \epsilon$, where θ^* represents the true parameter vector of our choice.
- Calculate the parameter estimator $\hat{\theta}$ using 5 with P deterministic or uniformly random .
- For each p, we compute both the generalization error and the test error. Following the set up used in (Belkin, Hsu, and Xu 2020), we take the true parameter θ^* such that

$$\theta_i^* \propto \frac{1}{j}, \qquad \|\theta^*\|^2 = 1$$

We repeat these steps for $M = 100$ and take the average over the two errors.

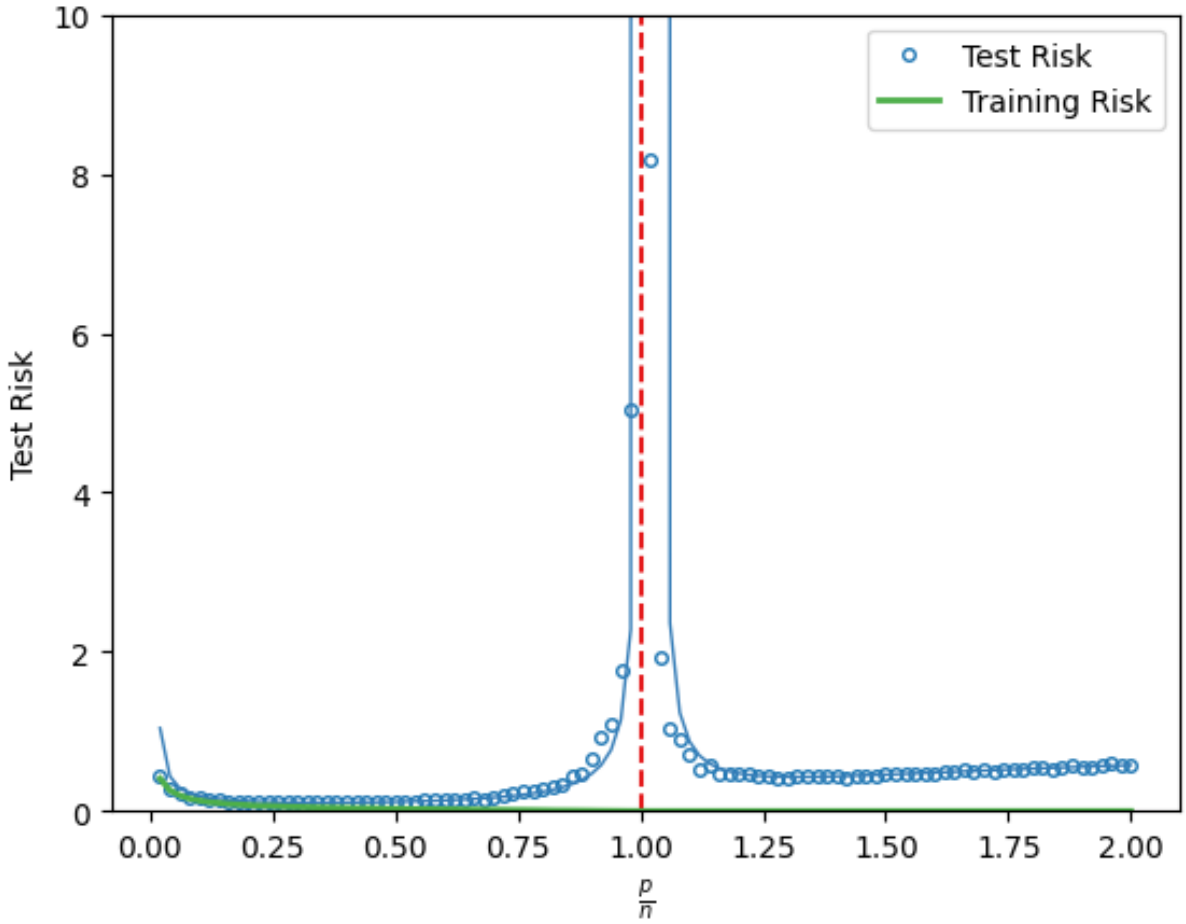

Figure 2: Plot of the generalization risk as a function of the over-parameterization ratio $\frac{p}{n}$ in the deterministic choice of the set P with $\sigma^2 = \frac{1}{25}$. Green line is the training loss, the blue line is the theoretical generalization loss from theorem 3.1 and blue circles are the average over 100 MC repetition of the generalization loss.

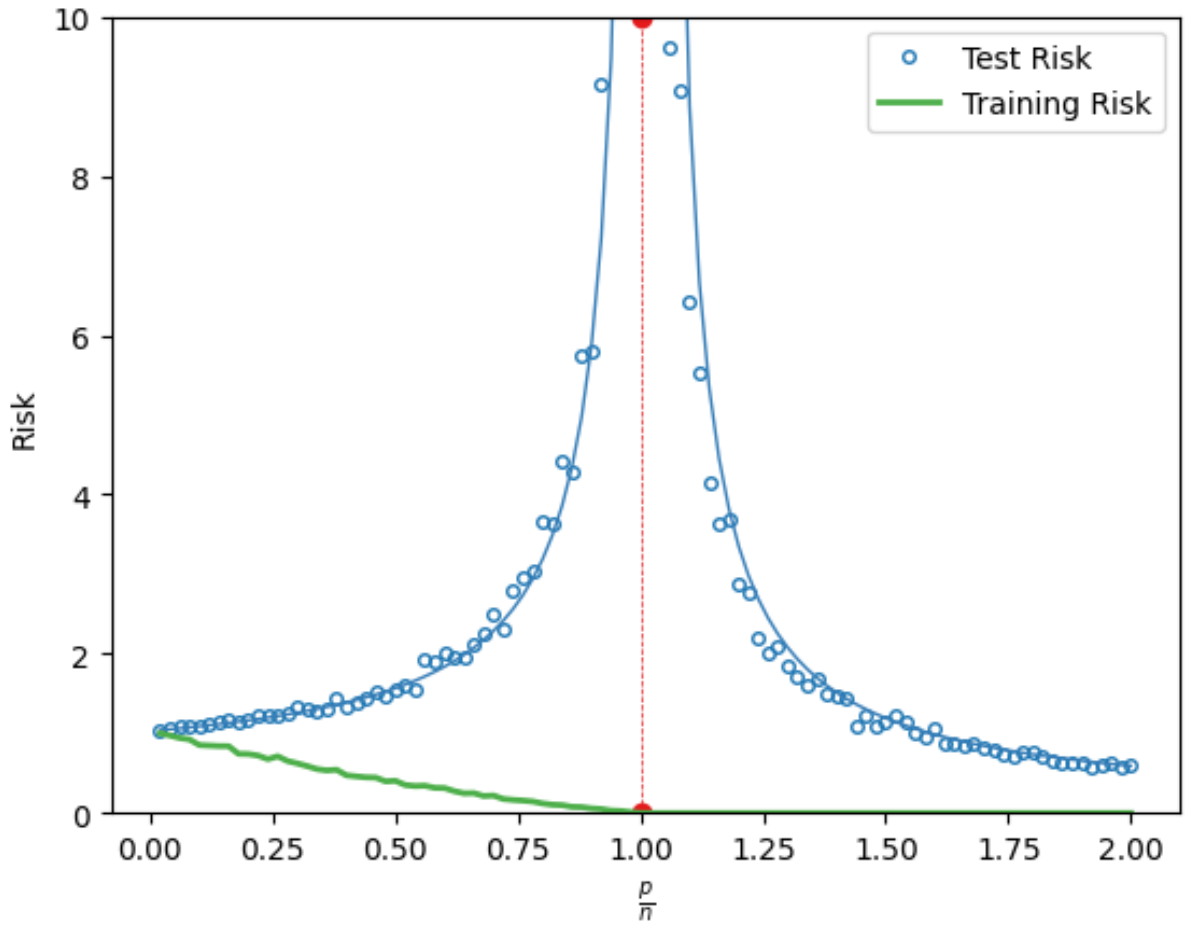

Figure 3: Plot of the generalization risk as a function of the over-parameterization ratio $\frac{p}{n}$ in the uniformly random choice of the set P with $\sigma^2 = \frac{1}{25}$. Green line is the training loss, the blue line is the theoretical generalization loss from corollary 3.2 and blue circles are the average over 100 MC repetition of the generalization loss.

We now support the claim that if the signal-to-noise $\frac{\|\theta^*\|^2}{\sigma^2}$ exceeds $\frac{d}{d-n-1}$ the global minimum is achieved in the over-parameterized regime.

Previously, we followed the setup from (Belkin, Hsu, and Xu 2020) with $\|\theta^*\|^2 = 1$ and $\sigma^2 = \frac{1}{25}$. We now fix $\|\theta^*\|^2 = 1$, and because σ must be less than $\sqrt{\frac{d-n-1}{d}}$ for the global minimum to be in the over-parametrized regime, we take three values for σ

$$\sigma_1 = \frac{1}{2}\sqrt{\frac{d-n-1}{d}}, \quad \sigma_2 = \sqrt{\frac{d-n-1}{d}}, \quad \sigma_3 = 2\sqrt{\frac{d-n-1}{d}}$$

and expect that for σ_1 the global minimum for the generalization error to be in over-parameterized regime.

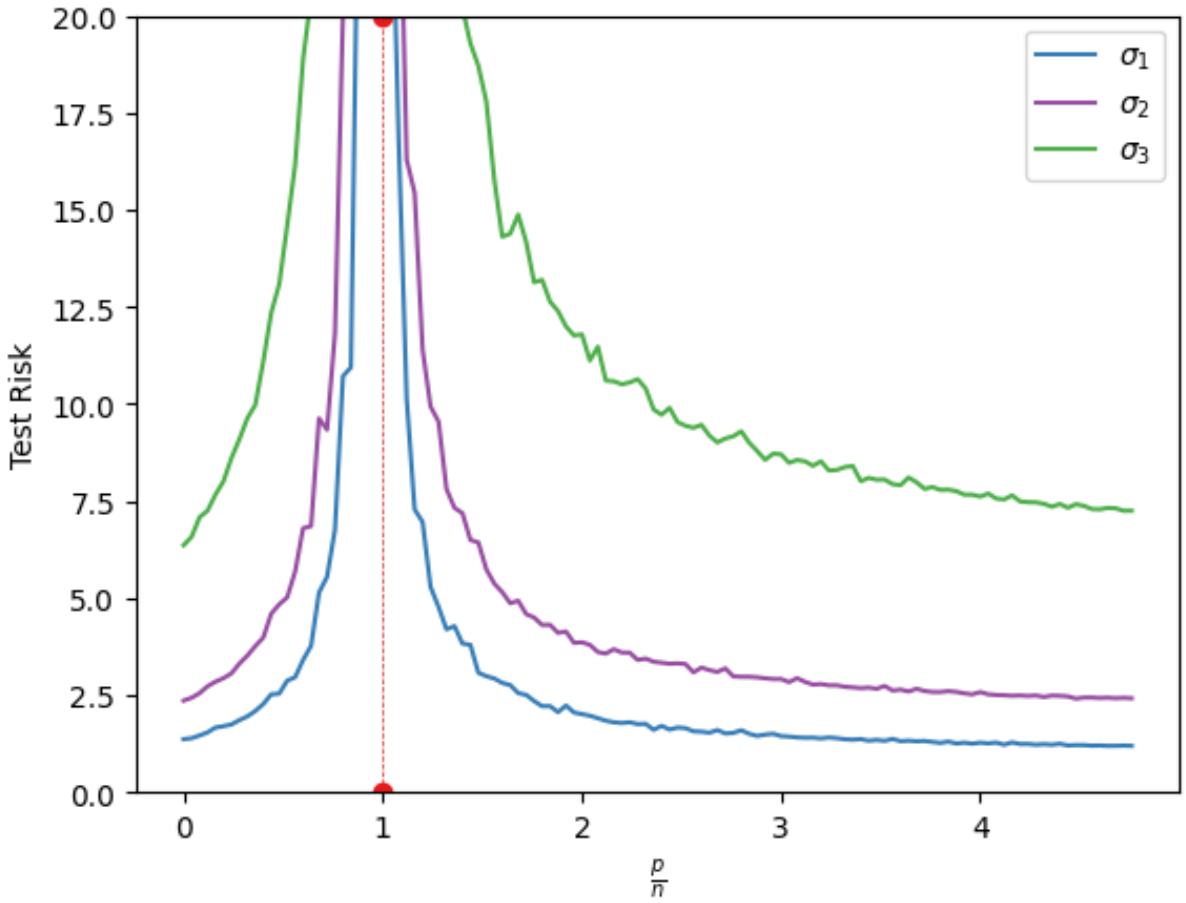

Figure 4: Plot of the generalization risk as a function of the over-parameterization ratio $\frac{p}{n}$ for different values of σ, with $d = 120$ and $n = 40$.

This figure shows that indeed for σ_1, the global minimum of the generalization error lies in the over-parameterized regime.

3.3 Anisotropic Gaussian features

Unlike in the previous setting, we now allow correlation between the features. We take $x \sim \mathcal{N}(0, \Sigma)$ to be of the form $x = \Sigma^{\frac{1}{2}} z$ for $\Sigma \neq I_p$ symmetric positive definite covariance matrix, and $z \sim \mathcal{N}(0, I)$. This is the setting described by (Hastie et al. 2020) where they consider more general setting for x being drawn form some distribution P_x. The results are more complicated than the ones from the previous subsection, as the min risk regression depends on the empirical distribution $\hat{H}_n$ of the descending ordered eigenvalues $(\lambda_1, ..., \lambda_p)$ of Σ, and $\hat{G}_n$ the coefficients of θ in the base of the eigenvectors $(v_1, ..., v_p)$ of Σ, which are defined as

$$\hat{H}_p(\lambda) = \frac{1}{p} \sum_{i=1}^{p} 1_{\{\lambda \geq \lambda_i\}}, \quad \hat{G}_p(\lambda) = \frac{1}{\|\theta\|^2} \sum_{i=1}^{p} \langle \theta, v_i \rangle 1_{\{\lambda \geq \lambda_i\}}.$$

Assumption 3.3 *Let us assume that:*

a) The covariates satisfy $x = \Sigma^{\frac{1}{2}} z$*, with* z *having independent entries with mean 0, variance 1 and finite moments.*

b) There exists a constant $M > 0$ *such that*

$$\frac{1}{M} < \lambda_p \leq \lambda_1 = \|\Sigma\|_{op} \leq M, \ \int \frac{1}{\lambda} d\hat{H}_p(\lambda) < M.$$

c) There exist $N > 0$ *such that*

$$|1 - \frac{p}{n}| \geq \frac{1}{N}, \ \frac{1}{N} \leq \frac{p}{d} \leq N.$$

Note on the assumptions. Assumption *b)* requires the eigenvalues of the Σ to be bounded and not accumulate near 0, and that the minimum eigenvalue is bounded away from 0. Assumption *c)* requires the ratio $\frac{p}{d}$ to be bounded away from the interpolation threshold $\frac{p}{d} = 1$.

Theorem 3.4 *(Hastie et al. 2020) Assume that assumptions 3.3 are satisfied and let* $\frac{p}{n} \to \gamma \in]0, \infty[$ *as* $n, p \to \infty$*. Define* $c_0 = c_0(\gamma, \hat{H}_p)$ *to be the unique non-negative solution of*

$$1 - \frac{1}{\gamma} = \int \frac{1}{1 + c_0 \gamma \lambda} d\hat{H}_p(\lambda) \tag{9}$$

Further assume that $\hat{H}_p$ *and* $\hat{G}_p$ *converge weakly in probability to probability measures on* $[0, \infty[$ H *rep.* G *as* $p, n \to \infty$*. Then almost surely*

$$R(\theta, \hat{\theta}) = \mathbb{E}_X(\|\theta - \hat{\theta}\|_\Sigma^2) = \underbrace{\|\mathbb{E}_X(\hat{\theta}) - \theta\|_\Sigma^2}_{B(\theta,\hat{\theta})} + \underbrace{Tr(Cov_X(\hat{\theta})\Sigma)}_{V(\theta,\hat{\theta})} \longrightarrow \mathcal{B}(H, G, \gamma) + \mathcal{V}(H, \gamma)$$

with

$$\mathcal{B}(\hat{H}_p, \hat{G}_p, \gamma) = \|\theta\|^2 \left(1 + \gamma c_0 \frac{\int \frac{\lambda^2}{(1+c_0\gamma\lambda)^2} d\hat{H}_p(\lambda)}{\int \frac{\lambda}{(1+c_0\gamma\lambda)^2} d\hat{H}_p(\lambda)}\right) \int \frac{\lambda}{(1 + c_0\gamma\lambda)^2} d\hat{G}_p(\lambda) \tag{10}$$

and

$$\mathcal{V}(\hat{H}_p, \gamma) = \sigma^2 \gamma c_0 \frac{\int \frac{\lambda^2}{(1+c_0\gamma\lambda)^2} d\hat{H}_p(\lambda)}{\int \frac{\lambda}{(1+c_0\gamma\lambda)^2} d\hat{H}_p(\lambda)}. \tag{11}$$

Proof: See the proof in (Hastie et al. 2020).

Note: Numerical evaluation of $\mathcal{B}(\hat{H}_p, \hat{G}_p, \gamma)$ and $\mathcal{V}(\hat{H}_p, \gamma)$ can generally be approached efficiently, as both involve integral computations and matrix-based operations. However, the most complex part is solving the equation (9).

We consider the special case where $\Sigma = I$, the distribution function then are given by:

$$\hat{H}_p(\lambda) = \frac{1}{p}\sum_{i=1}^{p} 1_{\{\lambda \geq \lambda_i\}} = 1_{\{\lambda \geq 1\}}$$

and

$$\hat{G}_p(\lambda) = \frac{1}{\|\theta\|^2}\sum_{i=1}^{p}\langle\theta, v_i\rangle 1_{\{\lambda\geq\lambda_i\}} = \frac{1}{\|\theta\|^2}\sum_{i=1}^{p}\langle\theta, e_i\rangle 1_{\{\lambda\geq 1\}} = \frac{1}{\|\theta\|^2}1_{\{\lambda\geq 1\}}\|\theta\|^2 = 1_{\{\lambda\geq 1\}}$$

By the definition of the Riemann–Stieltjes integral,we have

$$\int \frac{\lambda}{(1+c_0\gamma\lambda)^2} d\hat{H}_p(\lambda) = \sum_i \frac{\lambda_i}{(1+c_0\gamma\lambda_i)^2}\Delta\hat{H}_p(\lambda_i) = \frac{1}{(1+c_0\gamma)^2}.$$

with $\Delta\hat{H}_p(\lambda_i) = \hat{H}_p(\lambda_{i+1}) - \hat{H}_p(\lambda_i)$. We do the same for

$$\int \frac{\lambda^2}{(1+c_0\gamma\lambda)^2} d\hat{H}_p(\lambda) = \sum_i \frac{\lambda_i^2}{(1+c_0\gamma\lambda_i)^2}\Delta\hat{H}_p(\lambda_i) = \frac{1}{(1+c_0\gamma)^2}.$$

Replacing the integrals in the equations (10) and (11), and taking $c_0 = \frac{1}{\gamma(\gamma-1)}$ being the solution for (9), gives that

$$B(\theta,\hat{\theta}) = \|\mathbb{E}_X(\hat{\theta}) - \theta\|^2 \longrightarrow \|\theta\|^2(1-\frac{1}{\gamma})$$

and

$$V(\theta,\hat{\theta}) = Tr(Cov_X(\hat{\theta})) = \mathbb{E}\left[\|\mathbb{E}\left[\hat{\theta}\right] - \hat{\theta}\|^2\right] \longrightarrow \sigma^2\frac{1}{\gamma-1}$$

Note that this expression is the same found in the previous section for $\gamma > 1$, with $\mathbb{E}((y - x^\top\hat{\theta})) = B(\theta,\hat{\theta}) + V(\theta,\hat{\theta}) + \sigma^2$.

3.3.1 Empirical Results

In this section, we present an illustrative experiment to show case the DD in the Anisotropic Gaussian features case.

We consider the setup in the previous experiments this time with a covariance matrix with eigenvalues proportional to $\frac{1}{m}$, for $m \in \{1, ..., d\}$.

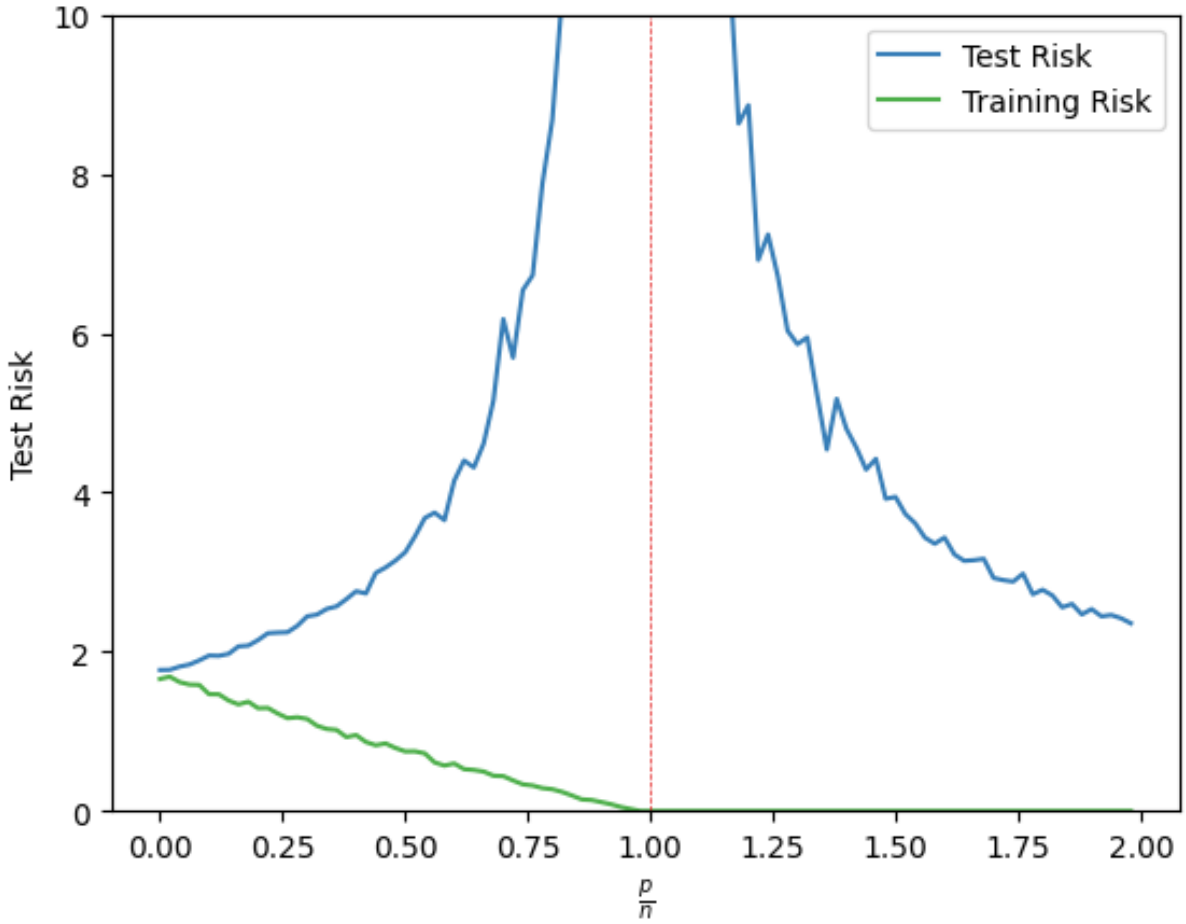

Figure 5: Plot of the generalization risk as a function of the over-parametrization ratio $\frac{p}{n}$ in the case of Anisotropic features.

4 Double descent in random features

In this section, we explore the DD in the random features model, which can be viewed as a two layer neural network with a fixed first layer represented by the random matrix W and trainable weights θ in the second layer. We consider the family of functions $\mathcal{H} = \{f : f(x) = \langle \theta, \phi(Wx) \rangle\}$, parameterized by the trainable θ and a fixed random feature map $\phi(Wx)$. We show that the DD occurs not only in the generalization error but also in the conditioning number of the Gram matrix ($\Phi\Phi^*$ or $\Phi^*\Phi$)of the feature matrix $\Phi := \phi(XW^\top)$.

4.1 Setting

We consider a student-teacher setup, in which the data is generated by a model $F : \mathbb{R}^d \to \mathbb{R}$ with additive noise, and the student aims to minimize the square loss:

$$(x_i, \epsilon_i) \overset{\text{i.i.d}}{\sim} P_x \times P_\epsilon \quad y_i = F(x_i) + \epsilon_i, \quad L(X, f) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{12}$$

with $\mathbb{E}(x_i) = 0, Cov(x_i) = \Sigma, \mathbb{E}(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$. We are interested in the generalization error $\mathcal{R}(f) = \mathbb{E}_x((F(x) - f(x))^2)$, with f is in the Reproduced Kernel Hilbert Hypothesis set $\mathcal{H} = \{f : f(x) = \langle \theta, \phi(Wx) \rangle\}$, with $W = [w_1, ..., w_p]^\top \in \mathbb{R}^{p \times d}$ is the weight matrix and $\phi : \mathbb{R} \to \mathbb{R}$ is a lipschitz continuous non linear activation function acting component-wise on Wx and has bounded Gaussian moments, i.e. $\mathbb{E}(\phi(G)^k) < \infty, \forall k \in \mathbb{N}$ for $G \sim \mathcal{N}(0, 1)$.
The analysis is done assuming the following assumptions.

Assumption 4.1 *Gaussian features $x_i \sim \mathcal{N}(0, I_d)$, and $w_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, d^{-1} I_d)$.*

Assumption 4.2 *We work in the following asymptotic setting:*

$$n, p, d \to \infty, \quad \frac{d}{n} \to \gamma, \quad \frac{p}{n} \to \psi, \quad \gamma, \psi \in]0, \infty[.$$

We consider the following bias-variance decomposition of the test error decoupling the randomness in the data and the weight initialization.

$$\mathcal{R}(\hat{\theta}) = \mathbb{E}_x \left[\left(\phi(x^\top W^\top)\hat{\theta} - F(x) \right)^2 | W \right] \tag{13}$$

$$= \underbrace{\mathbb{E}_x \left[\left(\mathbb{E}\left[\phi(x^\top W^\top)\hat{\theta} | W \right] - F(x) \right)^2 \right]}_{B(\hat{\theta})} + \underbrace{\mathbb{E}_x \left[\left(\phi(x^\top W^\top)\hat{\theta} - \mathbb{E}\left[\phi(x^\top W^\top)\hat{\theta} | W \right] \right)^2 \right]}_{V(\hat{\theta})} \tag{14}$$

The following theorem by (Ba et al. 2020), which draws significant inspiration from (Hastie et al. 2020) gives an expression of the variance term in RF model.

Theorem 4.3 *Given assumptions 4.1 and 4.2, we have*

$$V(\hat{\theta}) = \begin{cases} \sigma^2 \frac{\psi}{1-\psi} & \psi < 1, \\ -\sigma^2 \lim_{\xi \to 0} \left[\psi \frac{\partial}{\partial x} m_1(\xi, c_1 x, c_2 x)|_{x=0} + \frac{\partial}{\partial x} m_2(\xi, c_1 x, c_2 x)|_{x=0} + \frac{\psi - 1}{\xi^2} \right] & \psi > 1 \end{cases}$$

with $m_1(\xi, \rho, \eta), m_2(\xi, \rho, \eta)$ are the unique solution of the following fourth degree equations in $\{|m_1(\xi, \rho, \eta)|, |m_2(\xi, \rho, \eta)| \leq \frac{1}{Im(\xi)}\}$:

$$m_1^{-1} = -\xi - \rho - \gamma^{-1}\psi\eta^2 m_1 - c_1 m_2 + \frac{\eta^2 \gamma \psi m_1^2 (c_2 m_2 - \eta) - 2\eta c_2 m_1 m_2 + c_2^2 m_2 m_1^2}{m_1(c_2 m_2 - \eta) - \gamma\psi^{-1}} \tag{15}$$

$$m_2^{-1} = -\xi - \psi m_1 + \frac{\psi c_2 m_1^2 (c_2 m_2 - \eta)}{m_1(c_2 m_2 - \eta) - \gamma\psi^{-1}} \tag{16}$$

Where variables ρ, ξ, η satisfy $Im(\xi) > 0$ or $\xi < 0$, $\rho \geq \eta \geq 0$ and the constants c_1, c_2 defined as,

$$c_1 = \mathbb{E}(\phi(G)^2) - \mathbb{E}(\phi(G))^2, c_2 = \mathbb{E}(\phi(G))^2 \tag{17}$$

for $G \sim \mathcal{N}(0, 1)$.

proof: See the proof in section 8.2.

This theorem demonstrates the presence of DD behavior in random feature (RF) models, as the variance reaches a peak when $p = n$, before declining as the over-parameterization ratio ψ increases.

Remark. The theorem above holds irrespective of the teacher model F.

Proposition 4.4 *Given assumptions 4.1, 4.2, then for a linear teacher F we have, $B(\hat{\theta}) \to \infty$ as $\psi \to 1$. Furthermore, $B(\hat{\theta})$ is finite when $\psi > 1$.*

4.2 Double descent in the condition number of the random feature matrix

It has been demonstrated in (Chen and Schaeffer 2021) that the Random Feature model not only exhibits DD behavior in its generalization error, but also in the condition number of the Gram matrix associated with the feature matrix Φ.

Theorem 4.5 *((Chen and Schaeffer 2021)) Assume that $x_i \sim \mathcal{N}(0, \gamma^2 I_d)$, and $w_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2 I_d)$hold, and take $\phi(x) = \exp(ix)$(Random Fourier map), then we have the following*

*(1) (**under-parameterized regime** $n > p$) if the following holds*

$$n \geq C\eta^{-2} p \log(\frac{2p}{\delta}) \tag{18}$$

$$p \leq \sqrt{\delta}\eta(4\gamma^2\sigma^2 + 1)^{\frac{d}{4}} \tag{19}$$

for some $\gamma, \eta > 0$, then with probability at least $1 - 2\gamma$ we have

$$\left|\lambda_k(\frac{1}{n}\Phi^*\Phi) - 1\right| \leq 3\eta$$

for all $k \in \{1, ..., p\}$

*(2) (**over-parameterized regime** $n < p$) if the following holds*

$$p \geq C\eta^{-2} n \log(\frac{2n}{\delta}) \tag{20}$$

$$n \leq \sqrt{\delta}\eta(4\gamma^2\sigma^2 + 1)^{\frac{d}{4}} \tag{21}$$

for some $\gamma, \eta > 0$, then with probability at least $1 - 2\gamma$ we have

$$\left|\lambda_k(\frac{1}{p}\Phi\Phi^*) - 1\right| \leq 3\eta$$

for all $k \in \{1, ..., n\}$

*(3) (**Interpolation threshold**) If $n = p \geq 2$, then the eigenvalues satisfy*

$$\mathbb{E}_{x,w}\left[\lambda_{\min}(\frac{1}{p}\Phi^*\Phi)\right] \leq (1-\frac{1}{p})^{\frac{1}{2}}\frac{1}{(4\gamma^2\sigma^2+1)^{\frac{d}{4}}} + \frac{1}{p}$$
$$\mathbb{E}_{x,w}\left[\lambda_{\max}(\frac{1}{p}\Phi^*\Phi)\right] \geq 2 - \frac{1}{p}. \tag{22}$$

The constant $C > 0$ is no greater than 4 if $\min\{n, p\} > 25$.

Proof: See the proof in section 8.3

This theorem examines the behavior of the condition number of the Gram matrix of the Fourier feature map as a function of the number of data points n and the number of features p, irrespective of the data dimension d, and we notice that in the extreme case ($n >> p$ or $p >> n$), the parameter η in 19 can be chosen to be very small, which means that the eigenvalues of the Gram matrix concentrate around 1. At the interpolation threshold, if we assume that $p = n \leq (4\gamma^2\sigma^2+1)^{\frac{d}{4}}$, then using the Markov inequality on the $\lambda_{\min}$ of the scaled Gram matrix $\frac{1}{n}\Phi^*\Phi$ we have

$$\mathbb{P}(\lambda_{\min} \geq \frac{1}{\sqrt{n}}) \leq \mathbb{E}(\lambda_{\min})\sqrt{n} \leq \sqrt{(1-\frac{1}{n})}\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} \leq \frac{2}{\sqrt{n}}$$

then, for σ, γ ($\sigma, \gamma = 2$) and $d(\geq 44)$ big enough, $\lambda_{\min}$ is zero with probability almost 1. Meanwhile, from 22 we know that $\lambda_{\max}$ must be bigger than 1. This means that the normalized Gram matrix becomes ill-conditioned at least at the interpolation point, which proves that the condition number of the normalized Gram matrix of the feature map undergoes DD.

Remark. (Chen and Schaeffer 2021) argue that the theorem can be extended to other features (activation functions) and probability distribution.

4.3 Numerical results

In this section we illustrate the DD phenomenon in both Random Feature (RF) model with varying activation functions and examine the condition number of the Gram matrix of the feature maps.

The experiment setup is as follows.

1. Data Generation:
   - We generate a design matrix $X \in \mathbb{R}^{n\times d}$ where each column is drawn from a $\mathcal{N}(0, I)$ for $n = 100, d = 10, 20, 30$.
   - A weight matrix $W \in \mathbb{R}^{p\times d}$ with columns from $\mathcal{N}(0, \frac{1}{d}I)$ for varying values of p ranging from 1 to 200.
2. Feature mapping :
   - We construct the feature matrix ϕ by applying the activation function component wise to $XW^\top$.
   - Two activation functions are used:

- ReLU: $ReLU(x) = \max\{0, x\}$
- Sigmoid : $\phi(x) = \frac{1}{1+\exp(-x)}$.

3. Compute the Variance, Bias and the condition number of the Gram matrix of Φ:

   - generate the true labels using $y = \Phi\theta^* + \epsilon$ with θ^* is such that $\theta_i^* \propto \frac{1}{j}$ and $\|\theta^\star\|^2 = 1$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
   - Compute the solution $\hat{\theta} = \Phi^\dagger y$.
   - Compute the variance and bias using the formulas for $V(\hat{\theta})$ and $B(\hat{\theta})$.
   - Compute the condition number for Gram matrix of Φ

We repeat these steps for $M = 100$ times and plot the average.

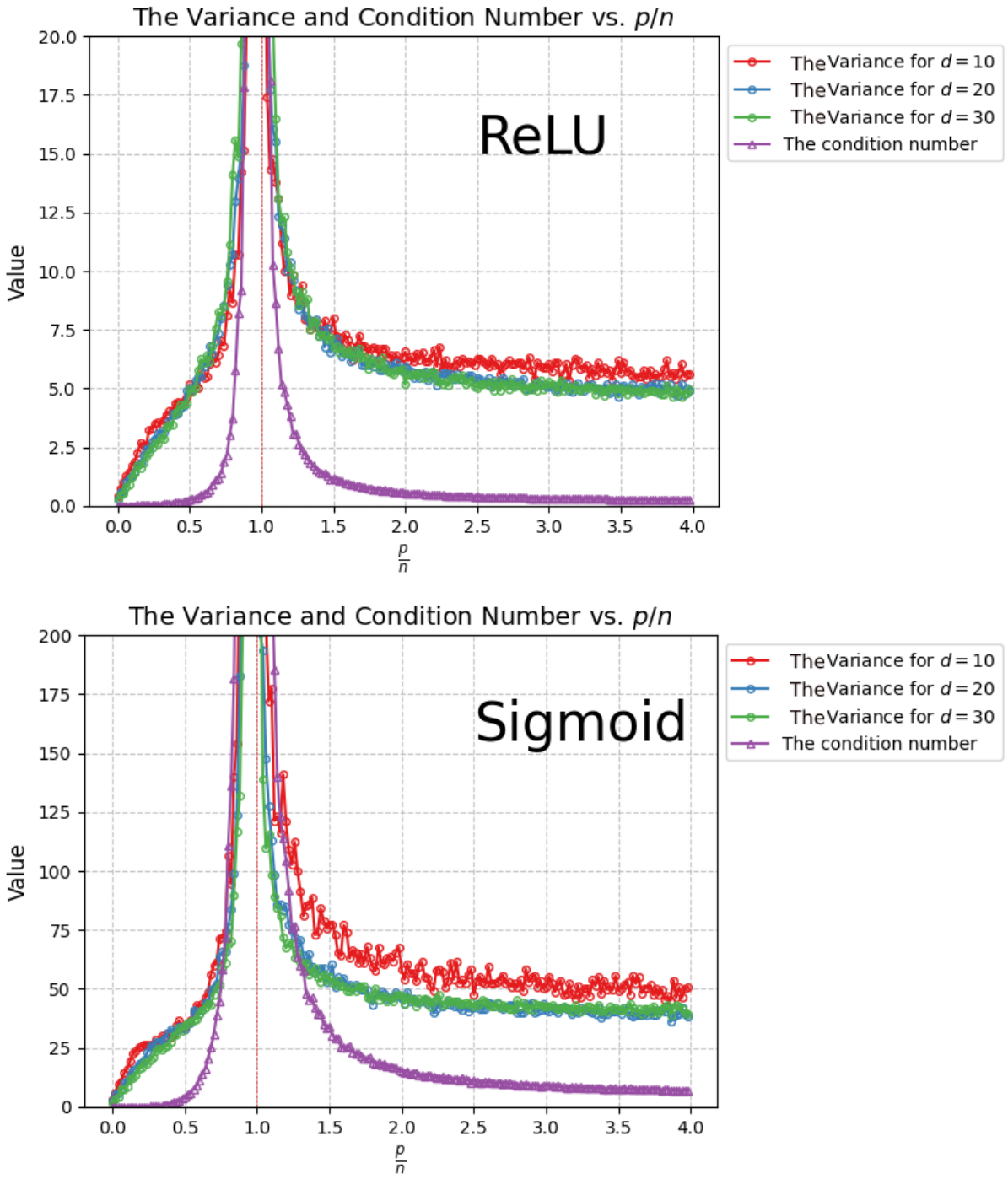

Figure 6: Plot of the condition number and the Variance for different data dimensions and activation functions. The curves are rescaled for visual purpose.

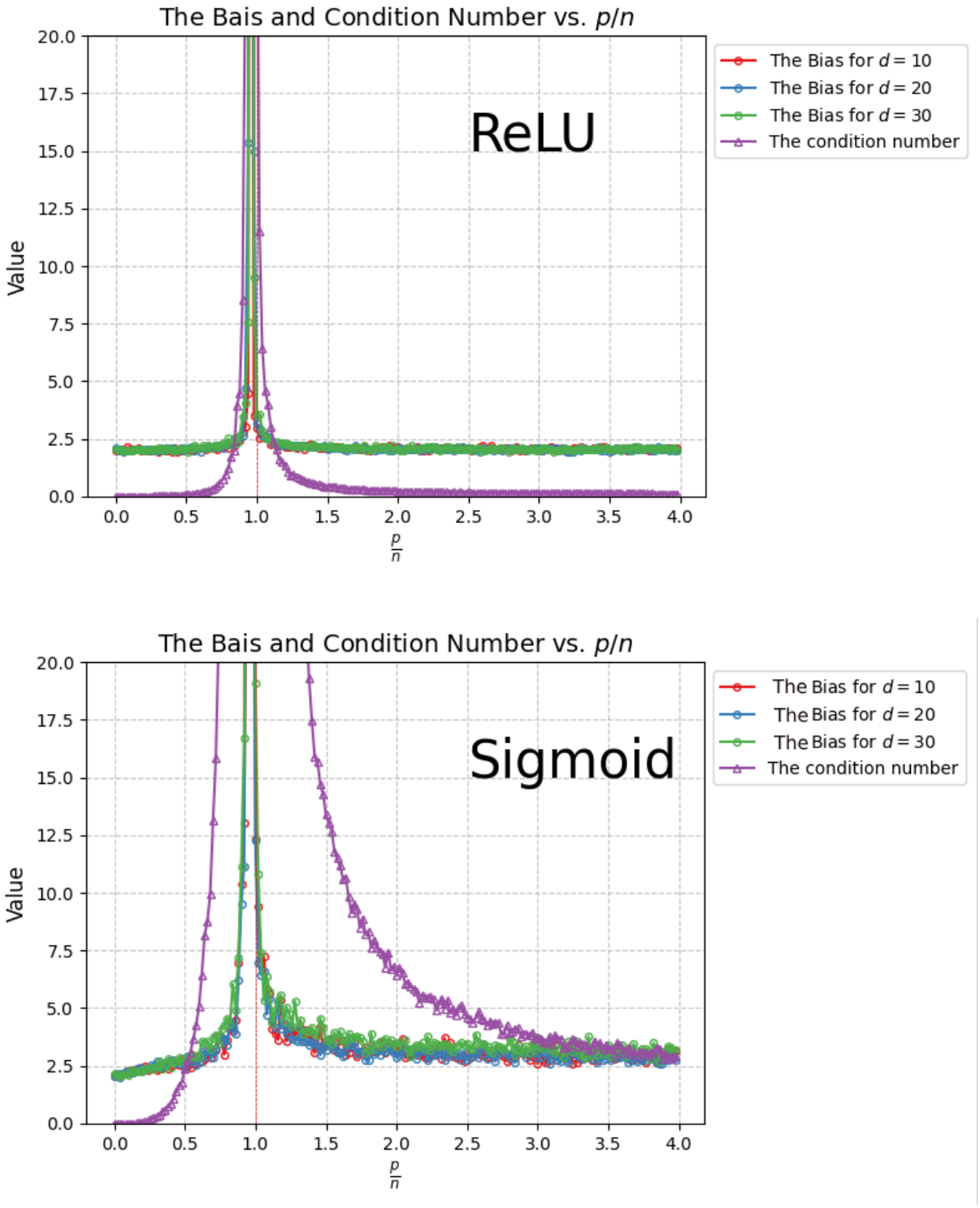

Figure 7: Plot of the condition number and the Bias for different data dimensions and activation functions. The curves are rescaled for visual purpose.

5 Double descent in neural networks

Model wise DD phenomenon in neural networks trained using different loss functions has been discussed in (Singh, Lucchi, et al. 2022). The analysis of the authors relies on the influence functions (which are

commonly used in robust statistics) for maximum likelihood-type estimators, i.e. $\hat{\theta} = \arg\max_\theta \mathcal{M}(x, \theta)$ for some likelihood function $\mathcal{M}$, along with the add-one procedure (a symmetric procedure to the more classical leave-one-out) to derive an expression for the generalization error in finite with neural networks. The idea is to express the generalization error using the add-one procedure which allows in a second stage to use the influence function to provide the lower bound for this loss.

5.1 Setting

Assume that the input data $z = (x, y)$ is drawn from a (unknown) distribution $\mathcal{D}$. The main task is to find an estimate parameter $\hat{\theta}$ such that $\mathcal{D}_{\hat{\theta}} \approx \mathcal{D}$. Given a finite sample $S = \{z_1 = (x_1, y_1), ..., z_n = (x_n, y_n)\}$, $\hat{\theta}$ can be computed by some statistical functional T on the empirical distribution $\mathcal{D}_n$ of S i.e. $\hat{\theta} = T(\mathcal{D}_n)$. We assume that our data is a tuple (x, y) with $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$, and consider the the neural network function $f : \Theta \times \mathbb{R}^d \to \mathbb{R}^k$, with $\Theta \subset \mathbb{R}^p$. We note by $\hat{\theta}_S \in \Theta$ the parameter estimator obtained by training to convergence the model on the data set S using a loss function $\ell : (\mathbb{R}^d \times \mathbb{R}^k) \times \Theta \to \mathbb{R}$ that is twice differentiable in the parameter θ, and compute its generalization loss using $\mathcal{L}(\hat{\theta}_S) := \mathbb{E}_z(\ell(z, \hat{\theta}_S))$.

Definition 5.1 *(Influence function) The influence function of a statistic functional T given a distribution $\mathcal{D}$ is defined as*

$$IF(z; T, \mathcal{D}) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\mathcal{D} + \epsilon\delta_z) - T(\mathcal{D})}{\epsilon} \tag{23}$$

evaluated at a point z where such limit exists, and δ_z is the Dirac distribution.

Influence functions are a useful tool in robust statistics and lately gained popularity in deep learning (Koh and Liang 2017) as a natural tool to analyse the change in distributions.
A classic result for the influence function at a point z on the parameter $\hat{\theta}$ is given by the following proposition from (Cook and Weisberg 1982).

Proposition 5.2 *(Cook and Weisberg 1982) The influence function of the maximum likelihood like estimator $\hat{\theta}_{\mathcal{D}}$ based on the distribution $\mathcal{D}$ evaluated at a point z takes the following form:*

$$IF(z : \hat{\theta}_{\mathcal{D}}, \mathcal{D}) = -[H_{\mathcal{L}}(\hat{\theta}_{\mathcal{D}})]^{-1} \nabla_\theta \ell(z, \hat{\theta}_{\mathcal{D}})$$

with the Hessian $H_{\mathcal{L}}(\hat{\theta}_{\mathcal{D}}) := \nabla^2_\theta \mathcal{L}(\hat{\theta}_{\mathcal{D}})$ contains the second derivative of the loss $\mathcal{L}(\theta) = \mathbb{E}_z(\ell(z, \theta))$ with respect to the parameter θ.

Add-one procedure studies the sensitivity of the parameter estimator $\hat{\theta}$ when an additional data point is added to the training set. Adding a point $z' = (x', y') \sim \mathcal{D}$ to the training set S can be interpreted as introducing a small perturbation in the original distribution $\mathcal{D}_n$ which leads to the new distribution $\tilde{\mathcal{D}}_{n+1} = (1-\epsilon)\mathcal{D}_n + \epsilon\delta_{z'}$ with the perturbation amount $\epsilon = \frac{1}{n+1}$. However retraining the model for each added point is rather slow. Fortunately, influence functions give us an efficient approximation. An

equivalent way to state the definition of the influence function is by defining $G = \delta_z - \mathcal{D}$, this gives:

$$IF(z; T, \mathcal{D}) = \lim_{\epsilon \to 0} \frac{T(\mathcal{D} + \epsilon G) - T(\mathcal{D})}{\epsilon} = \lim_{\epsilon \to 0} \frac{T(\mathcal{D}_\epsilon) - T(\mathcal{D})}{\epsilon} = \frac{d\hat{\theta}}{d\epsilon}|_{\epsilon=0}$$

where the last equality comes from the fact that the influence function measure how the model parameter $\hat{\theta}$ changes when the data distribution is perturbed by adding a small amount of weight at the point z. Next, we apply the chain rule to measure the change of functions of $\hat{\theta}$ at a point z', which gives

$$\begin{aligned} IF(\ell(z', \theta), \hat{\theta}_S, \mathcal{D}_S) &= \frac{d\ell(z', \hat{\theta}_S)}{d\epsilon}|_{\epsilon=0} = \nabla_\theta \ell(z', \hat{\theta}_S)^\top \frac{d\hat{\theta}_S}{d\epsilon}|_{\epsilon=0} \\ &= -\nabla_\theta \ell(z', \hat{\theta}_S)^\top [H_{\mathcal{L}}^S(\hat{\theta}_S)]^{-1} \nabla_\theta \ell(z', \hat{\theta}_S). \end{aligned}$$

Where $H_{\mathcal{L}}^S$ is computed on the set S.

Remark. The matrix $H_{\mathcal{L}}^S$ can in general be written for $k = 1$ as follows:

$$\begin{aligned} H_{\mathcal{L}}^S(\theta) = \nabla_\theta^2 \mathcal{L}_S(\theta) = H_o^S(\theta) + H_f^S(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_\theta(x_i) \left(\nabla_f^2 \ell_i\right) \nabla_\theta f_\theta(x_i)^\top \\ + \frac{1}{n} \sum_{i=1}^{n} \nabla_f \ell_i \nabla_\theta^2 f^\theta(x_i). \end{aligned} \tag{24}$$

Theorem 5.3 *((Singh, Lucchi, et al. 2022)) Consider the parameter estimator $\hat{\theta}_S$ based on the input samples S. Then the generalization risk $\mathcal{L}(\hat{\theta}_S) := \mathbb{E}_z(\ell(z, \hat{\theta}_S))$ takes the form:*

$$\mathcal{L}(\hat{\theta}_S) = \tilde{\mathcal{L}}(\hat{\theta}_S) + \frac{1}{n+1} Tr\left(\left[H_{\mathcal{L}}^S(\hat{\theta}_S) + \lambda I\right]^{-1} C_{\mathcal{L}}^{\mathcal{D}}(\hat{\theta}_S)\right) + \mathcal{O}(\frac{1}{n^2}).$$

Where $|S| = n$, $\tilde{\mathcal{L}}(\hat{\theta}_S) := \mathbb{E}_z(\ell(z, \hat{\theta}_{S \cup z}))$ is the expectation of the one sample training loss and $C_{\mathcal{L}}^{\mathcal{D}}(\hat{\theta}_S) := \mathbb{E}_z\left[\nabla_\theta \ell(z, \hat{\theta}_S) \nabla_\theta \ell(z, \hat{\theta}_S)^\top\right]$ is the (uncentered) covariance of loss gradients.

Proof: See the proof in (Singh, Lucchi, et al. 2022).

The expression of the generalization loss in t his theorem is used to understand how the model's performance changes when a new data point z' is added to the training set. Specifically, it helps in analyzing the sensitivity of the model to the inclusion of new data points and how this affects the generalization error.

Remark: When trained sufficiently long enough so that the training loss on a single sample is close to zero, the expectation of the one sample training loss $\mathcal{L}(\hat{\theta})$ becomes negligible.

The regularization term ($\lambda > 0$) is added to the Hessian since for neural networks the Hessian is in general rank deficient.

Let us define the covariance of network Jacobians when computed on the set of samples S to be

$$\mathcal{C}_f^S(\theta) = \frac{1}{|S|} Z_S(\theta) Z_S(\theta)^\top, \text{with} \quad Z_S := [\nabla_\theta f(\theta, x_1), ..., \nabla_\theta f(\theta, x_n)]^\top \in \mathbb{R}^{p \times kn}.$$

Having established this, we now introduce the assumptions required for the analysis and focus on the Mean-Squared-Error (MSE) loss function $\ell(z,\theta) = \ell(f(\theta,x),y) = \frac{1}{2}\|y - f(\theta,x)\|^2$.

Definition 5.4 *(Sub-Gaussian random vectors)A random variable X is sub-Gaussian if there exists $K > 0$ called the sub-Gaussian moment of X such that*

$$\mathbb{P}(|X| > t) \leq 2e^{-\frac{t^2}{K^2}}.$$

The sub-Gaussian norm of a random vector $x \in \mathbb{R}^d$ is defined as

$$\|x\|_{\psi_2} = \sup_{p \geq 1} \left\{ \frac{1}{\sqrt{p}} \mathbb{E}\left[|X|^p\right]^{\frac{1}{p}} \right\}.$$

Examples of sub-Gaussian random variables:

- **Gaussian**: A standard normal random variable X is sub-Gaussian with $\|x\|_{\psi_2} \leq C$, where C is a constant.

- **Bernoulli**: A Bernoulli random variable X is a sub-Gaussian random variable with $\|x\|_{\psi_2} = 1$

- **Bounded**: Any bounded random variable X ($|X| \leq M$ a.s) is a sub-Gaussian random variable with $\|x\|_{\psi_2} \leq M$.

Assumption 5.5 *There exists a sample $(x,y) \sim \mathcal{D}$ with non-zero probability α such that $\sigma^2(x,y) := \|\nabla_f \ell(f(\hat{\theta},x),y)\|^2 > 0$*

Assumption 5.6 *The functional Hessian at the optimum $\hat{\theta}_S$ is zero, i.e.,*

$$H_f^S(\hat{\theta}_S) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}[\nabla_f \ell_i]_j \nabla_\theta^2 f_j(\hat{\theta}_S, x_i) = 0.$$

Assumption 5.7 *The minimum non-zero eigenvalue $\lambda_{\min}$ of covariance of the network Jacobians at optimum $C_f^S(\hat{\theta}_S)$ is bounded by the corresponding one at initialization θ^0 over some common set S, i.e., $A\lambda_{\min}(C_f^S(\theta^0)) \leq \lambda_{\min}(C_f^S(\hat{\theta}_S)) \leq B\lambda_{\min}(C_f^S(\theta^0))$, with constants $0 < A, B < \infty$*

Assumption 5.8 *The columns of $Z_S(\theta^0)$ are sub-Gaussian independent random vectors at initialization θ^0.*

Assumption 5.9 *we work in the following asymptotic setting:*

$$n, p \to \infty \quad \textit{and} \quad \frac{p}{n} \in]0, \infty[$$

Note on the assumptions. Assumption 5.5 only requires that the generalization error is not zero in at least one point in the distribution $\mathcal{D}$. Otherwise it will not be of interest. Assumption 5.6, as shown

empirically by (Singh, Bachmann, et al. 2021) to hold, indicates that the rank of the functional Hessian converges to zero after sufficient training, especially near the interpolation threshold where losses and gradients approach zero. Assumption 5.7 supports a lower bound on the population risk by ensuring that the minimum non-zero eigenvalue of the covariance of the network Jacobians matrix remains stable from initialization to optimum. Assumption 5.8, only requires sub-Gaussianity of the covariance at initialization. This assumption, common in DD study (Kuzborskij et al. 2021), enables precise characterization of the minimum eigenvalue that appears in theorem 8.5, and supports predictions on DD behavior via Random Matrix Theory.

Theorem 5.10 *Under the assumptions 5.5, 5.6, 5.7 5.8, 5.9 and taking the limit of the regularization* $\lambda \to 0$*, the generalization error for the finite width neural network trained using the MSE for* $k = 1$ *takes the following form and diverges almost surely at* $p = \frac{1}{c^2}n$

$$\mathcal{L}(\hat{\theta}_S) \geq \tilde{\mathcal{L}}_S(\hat{\theta}_S) + \frac{\sigma_{\min}^2 \alpha A \lambda_{\min}\left(C_f^{\tilde{\mathcal{D}}}(\theta^0)\right)}{B\|C_f^{\mathcal{D}}(\theta^0)\|_2 \left(1 - c\sqrt{\frac{p}{n}}\right)^2}. \tag{25}$$

where the constant $c > 0$ *depends only in the sub-Gaussian norm of the columns of* $Z_S(\theta^0)$*, and* $\tilde{\mathcal{D}}$ *is as in 8.5*

Proof: See the proof in section 8.4.

Theorem 5.10 provides a lower bound for the generalization error of NNs trained using MSE loss. We can see that this error diverges when $p \approx n$ and that the second term on the RHS decreases as the denominator increases for $p >> n$.

Remark. It is worth noting that (Singh, Lucchi, et al. 2022) never claimed that this lower bound is tight bound for the generalization error but the aim was to isolate the source of the DD.

6 Epoch Wise Double Descent

Interestingly, the DD behavior manifests not only with increases in model size "model-wise DD", but also as training time "epoch-wise DD" as observed by (Nakkiran, Kaplun, et al. 2019), and the aim of this section is to show that the epoch-wise DD unlike the model-wise is not tied to over-parameterized models, but both under and over-parameterized models can have epoch-wise DD (Heckel and Yilmaz 2020).

6.1 Setting

We consider the following linear model $y = x^\top \theta^* + \epsilon$, with $x \in \mathbb{R}^d$, $\Sigma = diag(\sigma_1^2, ..., \sigma_d^2)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of x. Let $X = \frac{1}{\sqrt{n}}[x_1, ..., x_n] \in \mathbb{R}^{n \times d}$ is the design matrix and $y = \frac{1}{\sqrt{n}}[y_1, ..., y_n]$ drawn from a given data distribution $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, x_n)\}$ with n iid data points.

In this section, we consider the estimate based on early stopping gradient descent applied to the empirical

risk $\hat{R}(\theta) = \|X\theta - y\|^2$ initialized at $\theta^0 = 0$.
The iterates of the gradient descent are given by

$$\theta^{t+1} = \theta^t - \frac{1}{2}\text{diag}(\eta)\nabla_\theta \hat{R}(\theta^t) \quad \text{for} \quad t = 1, 2, 3 \cdots$$

with $\text{diag}(\eta)$ is a diagonal matrix containing the stepsizes $\eta_i > 0$ allowing different stepsized for all the features.
The difference between the parameter estimate returned by the gradient descent at time $t+1$ and the true parameter θ^* is given by

$$\begin{aligned}\theta^{t+1} - \theta^* &= \theta^t - \frac{1}{2}\text{diag}(\eta)\nabla_\theta \hat{R}(\theta^t) - \theta^* \\ &= \theta^t - \text{diag}(\eta)X^\top X\theta^t + \text{diag}(\eta)X^\top y - \theta^* \\ &= (I - \text{diag}(\eta)X^\top X)(\theta^t - \theta^*) + \text{diag}(\eta)X^\top \epsilon.\end{aligned}$$

with $\epsilon = y - X\theta^* = [\epsilon_1, \cdots, \epsilon_n]^\top$.
The idea used for this analysis by (Heckel and Yilmaz 2020) is to analyse an approximation $\tilde{\theta}^t$ of the parameter estimator at time t, which is given by

$$\tilde{\theta}^{t+1} - \theta^* = (I - \text{diag}(\eta)\Sigma)(\tilde{\theta}^t - \theta^*) + \text{diag}(\eta)X^\top \epsilon. \tag{26}$$

since Σ is an estimate of $X^\top X$.

Note: In the paper (Heckel and Yilmaz 2020), there appears to be a recurring typographical error where the relationship $X^\top X \approx \Sigma^2$ is stated. This approximation is not generally true unless the eigenvalues of Σ satisfy $\sigma_i^2 \in \{0, 1\}$ for all i.

Theorem 6.1 *((Heckel and Yilmaz 2020)) In the setting above, suppose that the stepsize satisfies $\eta_i \leq \frac{1}{\sigma_i^2}$ for all $i = 1, 2, ..., d$. Then, in the under-parameterized regime ($d << n$) it holds with probability at least $1 - 2d^{-5} + 2de^{\frac{-n}{8}} - e^{-d} - 2e^{-32}$ that*

$$\left|R(\theta) - \tilde{R}(\tilde{\theta}^t)\right| \leq c\left(\frac{\max_i \eta_i^2\sigma_i^4}{\min_i \eta_i\sigma_i^4}\frac{d}{n}\left(\|\Sigma\theta^*\|^2 + \frac{d}{n}\sigma^2\log(d)\right) + \frac{\sigma^2}{n}\sqrt{d}\right) \tag{27}$$

with $\tilde{R}(\tilde{\theta}^t) = \sigma^2 + \sum_{i=1}^d \underbrace{\sigma_i^2(\theta^*)^2(1-\eta_i\sigma_i^2)^{2t} + \frac{\sigma^2}{n}(1-(1-\eta_i\sigma_i^2)^t)^2}_{U_i(t)}$, and c is a constant.

Proof: See the proof in section 8.5.
This theorem provides, with high probability, an approximation of the risk of the gradient descent at time t in the under-parameterized regime. As a result, we notice that in the expression $\tilde{R}(\tilde{\theta})^t$ the $U_i(t)$ terms have contrasting monotonicities: $\sigma_i^2(\theta^*)^2(1-\eta_i\sigma_i^2)^{2t}$ decreases in time and the other one $\frac{\sigma^2}{n}(1-(1-\eta_i\sigma_i^2)^t)^2$ increases in time. The combined behavior of $U_i(t)$ can be described as follows:

- For small t: the decreasing term $\sigma_i^2(\theta^*)^2(1-\eta_i\sigma_i^2)^{2t}$ dominates over the increasing term $\frac{\sigma^2}{n}(1-(1-$

$\eta_i\sigma_i^2)^t)^2$ and therefore the function U_i decreases.

- Transition point: there exists a point $\tilde{t}$ such that U_i reaches its minimum value. This minimum depends on the parameters η_i and σ_i.
- For $t > \tilde{t}$: The increasing term starts to dominate, causing U_i to increase.

Thus, U_i has a U-shaped profile in relation to t. Consequently, the generalization error, as a superposition of U-shaped functions, can exhibit a DD over epochs if the minima occur at different times. This phenomenon gives rise to the observed epoch-wise DD effect.

Remark: In the next section we are going to show that this behavior can be eliminated by choosing a stepsize for each parameter entry, so that the minima overlap at the same point.

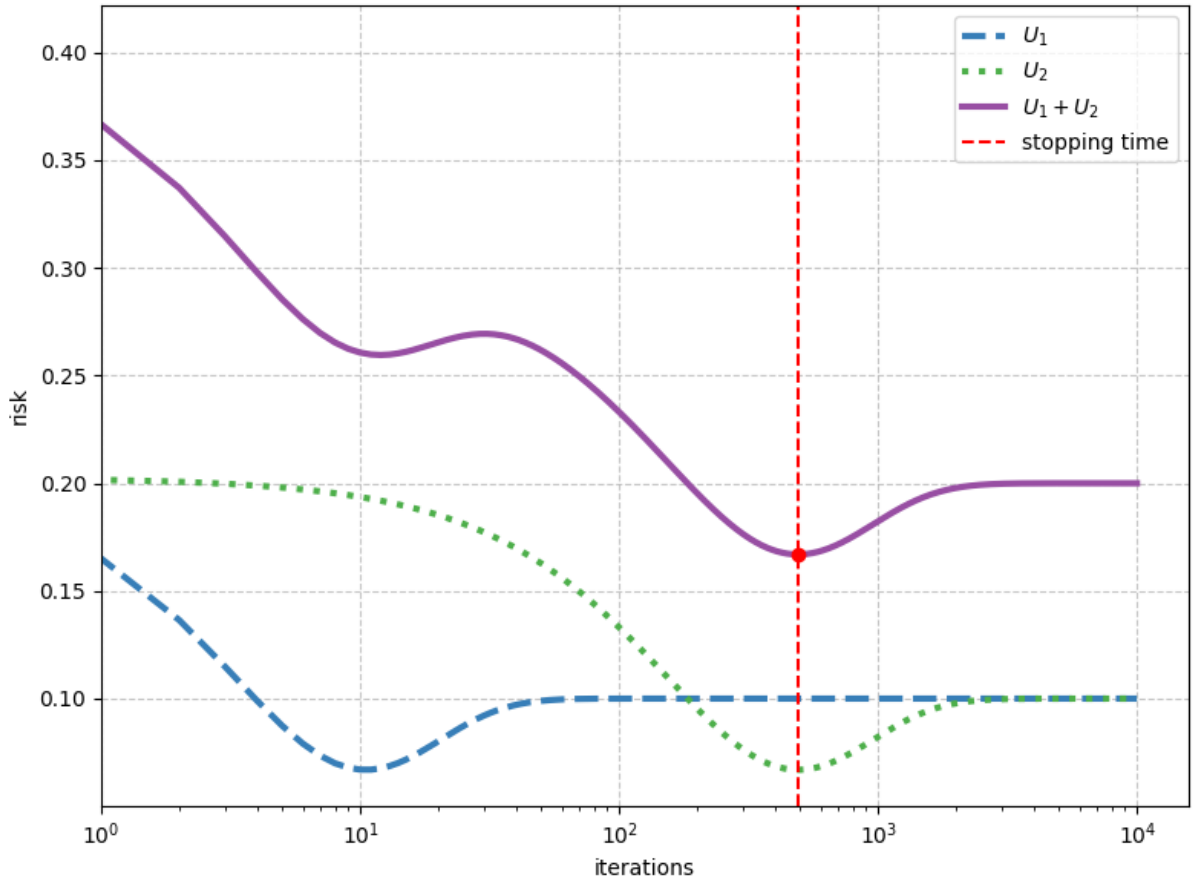

Figure 8: Plot of two function $U_1(t)$ and $U_2(t)$ and their sum, with $\theta_1^* = 0.45, \sigma = 1, \sigma_1 = 1$ and stepsize $\eta_1 = 0.1$ for $U_1(t)$, and $\theta_2^* = 3, \sigma_2 = 0.15, \eta_2 = 0.1$ for $U_2(t)$.

Figure 8 illustrates that the superposition of the two U-shaped functions U_1 and U_2, exhibits a DD.

7 Mitigating The Double Descent

In this section, we explore several techniques aimed at mitigating the DD phenomenon. These techniques include regularization through ridge regression (Mei and Montanari 2019), early stopping (Heckel and

Yilmaz 2020) and performance improvement via ensemble methods (Wilson and Izmailov 2020), (Loureiro et al. 2022).

7.1 Ridge regression

In this section, we demonstrate that for a specific class of linear models, applying optimal regularization effectively prevents model-wise DD. In other words, given a fixed sample size, larger models do not perform worse than smaller ones when appropriately regularized.

We consider the setting of 3.1 with the random choice of the set of indices P. The generalization error of a parameter estimator $\hat{\theta}_p(X_P)$ is given by $\mathcal{R}(\hat{\theta}_p) = \mathbb{E}_{x_p,y}(x_p^\top \hat{\theta}_p - y)$.

For any estimator $\hat{\theta}_P$ as a function of the observed samples, its expected risk is

$$\bar{\mathcal{R}}(\hat{\theta}) = \mathbb{E}_P \mathbb{E}_{X_P,y}\left[(X_p^\top \hat{\theta} - y)^2\right]. \tag{28}$$

Let $\hat{\theta}_{p,\lambda}$ be the solution for the regularized least squares problem

$$\hat{\theta}_{p,\lambda} = \arg\min_{\theta \in \mathbb{R}^p} \|X_p\theta - y\|^2 + \lambda\|\theta\|^2, \quad \lambda > 0 \tag{29}$$

and λ_p^{opt} be the optimal ridge regression (that achieves the minimum expected risk) for a model of size p and number of samples n

$$\lambda_p^{opt} = \arg\min \mathcal{R}(\hat{\theta}_{p,\lambda}). \tag{30}$$

Let $\hat{\theta}_p^{opt}$ be the estimator that corresponds to λ_p^{opt} that is

$$\hat{\theta}_p^{opt} = \arg\min_{\theta \in \mathbb{R}^p} \|X_p\theta - y\|^2 + \lambda_p^{opt}\|\theta\|^2. \tag{31}$$

We now state the following theorem that shows that the generalization error is monotone when using optimal ℓ_2 regularization.

Theorem 7.1 *((Nakkiran, Venkat, et al. 2021)) In the setting above, for all $d \in \mathbb{N}, p \leq d, n \in \mathbb{N}$ we have*

$$\mathcal{R}(\hat{\theta}_p^{opt}) \leq \mathcal{R}(\hat{\theta}_{p+1}^{opt}). \tag{32}$$

proof: See the proof in (Mei and Montanari 2019).

7.1.1 Empirical Results

We conduct experiments to support and validate the theoretical results presented above, extending the analyses for various settings and beyond linear regression.

Isotropic features. We consider the setting of theorem 7.1, with $n = 50, d = 100, \sigma = \frac{1}{3}, \|\theta^*\|^2 = 1$.

We take the optimal regularization parameter λ_{opt} as in lemma 3 by (Nakkiran, Venkat, et al. 2021) and average the generalization risk over $M = 50$ MC iterations.

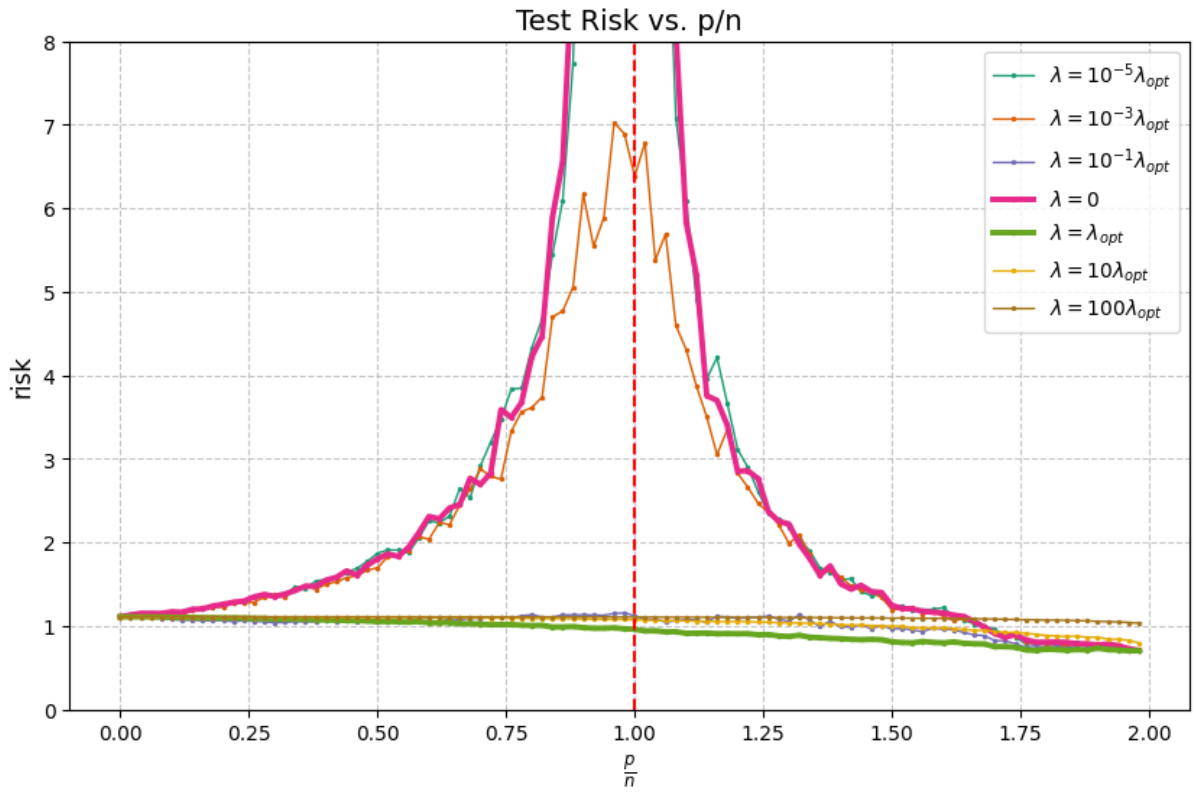

Figure 9: Plot of the average risk over 50 Monte Carlo iteration of ridge regression for $d = 100, n = 50, X \sim \mathcal{N}(0, I)$, noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\|\theta^\star\|^2 = 1$.

Anisotropic features. We follow the same setting as above but this time with $\Sigma = diag(\lambda_1, ..., \lambda_d)$ with exponential decay of the eigenvalues $\{\lambda_i\}_{i=1}^d$.

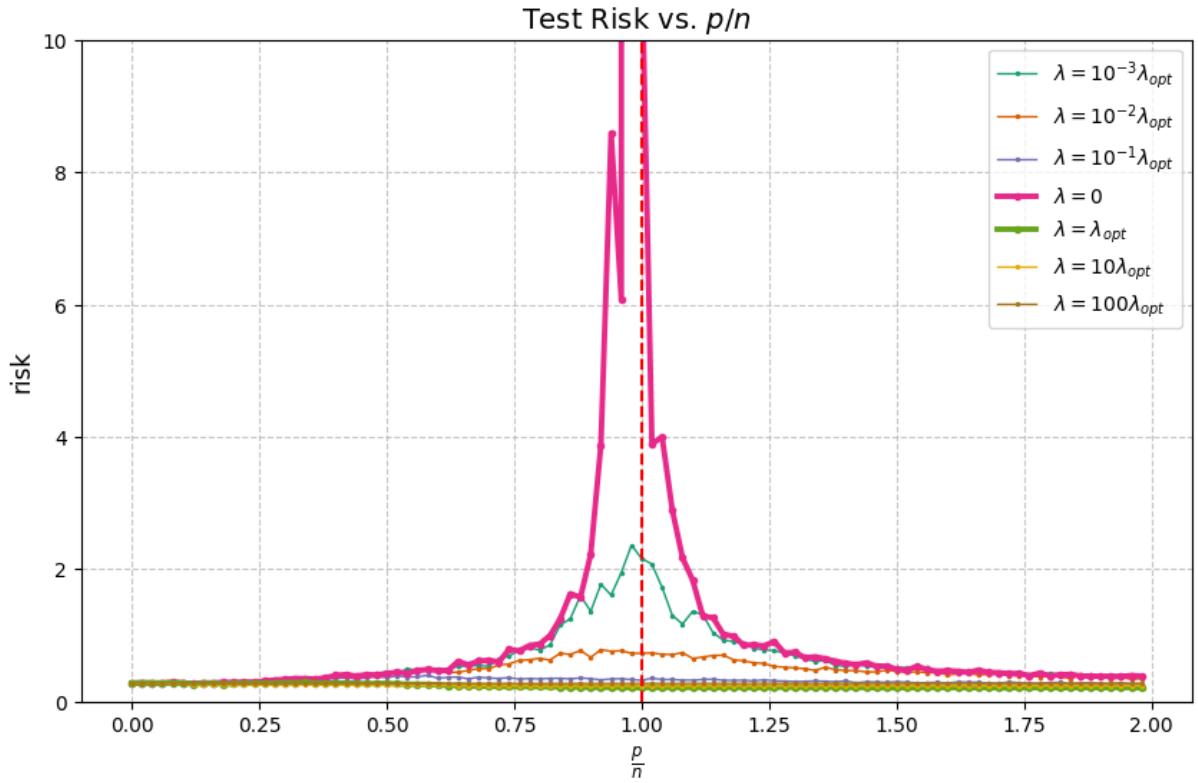

Figure 10: Plot of the average risk over 50 Monte Carlo iteration of ridge regression for $d = 100, n = 50, X \sim \mathcal{N}(0, \Sigma)$, noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\|\theta^\star\|^2 = 1$.

Random Fourier features model. We consider RFF applied to the MNIST dataset, an image classification problem with 10 classes. The input dimension is $d = 784$, corresponding to the 28×28 pixel images, and we vary the number of random Fourier features between 10 and 6000.

The random Fourier features are generated by sampling from a $\mathcal{N}(0, \sigma^2 I)$ distribution, with $\sigma = 1/5$. For optimization, we use *Stochastic Gradient Descent* (SGD) with a learning rate of 0.01, momentum of 0.95, and a batch size of 64. Training is performed for 600 epochs on $n = 1000$ using MSE loss.

We investigate the impact of ℓ_2 regularization by testing regularization parameters $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$.

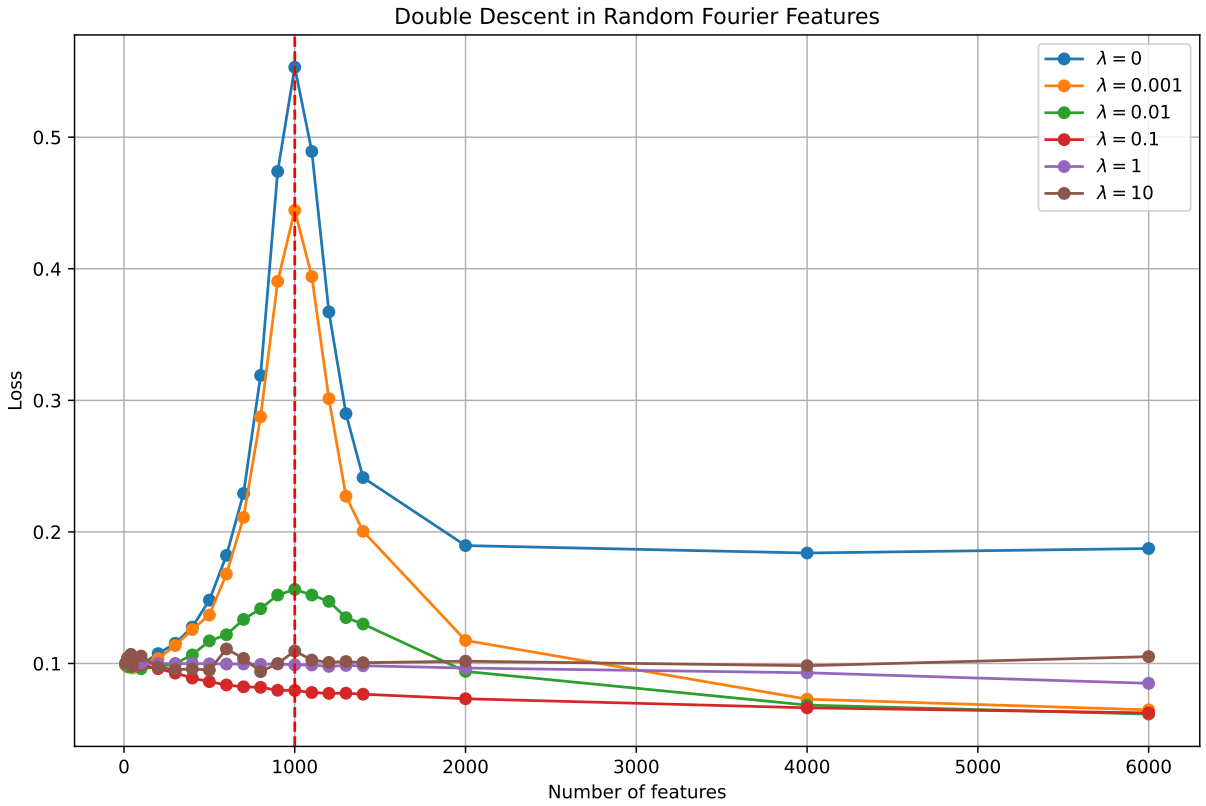

Figure 11: Double descent curve for RFF model on MNIST dataset. The interpolation threshold is achieved at $N = 1000$.

7.2 Early Stopping

Here, we introduce a method to mitigate DD described in 6 by selecting a step size that aligns the minima of each feature to occur simultaneously, as shown in figure 12. This approach helps ensure that features reach their lowest points together, which can smooth out fluctuations in generalization error and reduce the effects of DD.

Proposition 7.2 *In the context of 6. Pick an optimal early stopping time $\tilde{t} \geq 1$. The minimum of the risk expression $\min_{\eta_1,\ldots,\eta_d} \min_t \tilde{R}(\tilde{\theta}^t)$ is achieved at iteration $\tilde{t}$ by choosing the stepsizes to be*

$$\eta_i = \frac{1}{\sigma_i^2}\left(1 - \left(\frac{\frac{\sigma^2}{n}}{\sigma_i^2(\theta^*)^2 + \frac{\sigma^2}{n}}\right)^{\frac{1}{\tilde{t}}}\right). \tag{33}$$

Proof: Taking the derivative with respect to t of each component $U_i(t)$ of the risk $\tilde{R}(\tilde{\theta}^t)$ and setting it to zero gives the stepsize

$$\eta_i = \frac{1}{\sigma_i^2}\left(1 - \left(\frac{\frac{\sigma^2}{n}}{\sigma_i^2(\theta^*)^2 + \frac{\sigma^2}{n}}\right)^{\frac{1}{\tilde{t}}}\right).$$

replacing η_i in the expression gives

$$\min_t U_i(t) = \frac{\frac{\sigma^2}{n}\sigma_i^2\theta_i^{*2}}{\frac{\sigma^2}{n} + \sigma_i^2\theta_i^{*2}},$$

which is independent of time t.

This proposition provides a theoretical approach to eliminate the epoch-wise DD described in 6. However, in practice, we may not know the variance of the characteristics σ_i of the true parameter θ^*. A more practical approach involves treating the stepsizes as hyper-parameters and appropriately scaling them during the training process. This allows for an effective mitigation of DD, even in the absence of precise knowledge about the underlying feature statistics or model parameters.

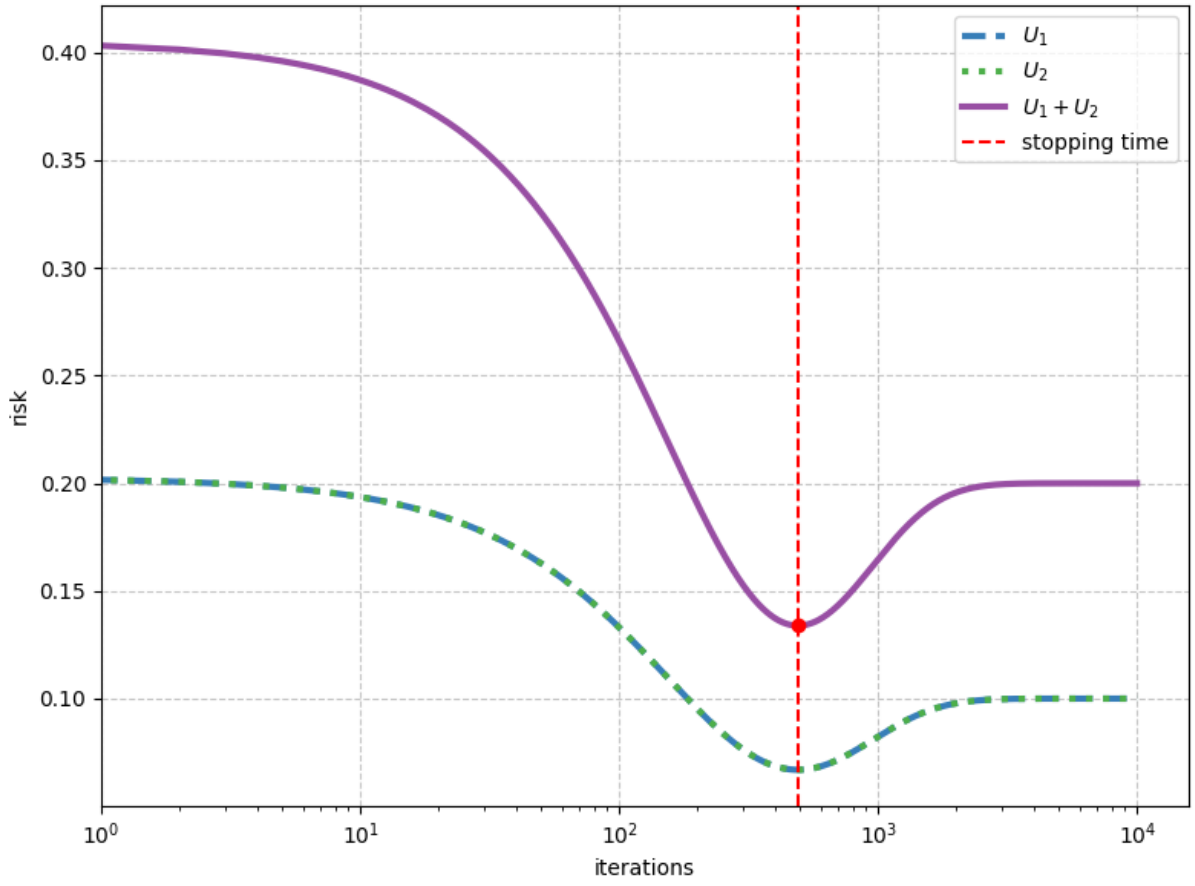

Figure 12: Plot of two functions $U_1(t)$ and $U_2(t)$ in 27, with the stepsizes η_1 and η_2 following 33, and their sum. This shows that the step size 33 mitigates the DD by forcing the minima of the U-shaped functions U_i to align at the same iteration.

7.3 Ensemble methods

Ensemble methods that train multiple learners and combine them by boosting or bagging are known to be significantly more accurate than individual learners. Recently, this method has gained attention as an effective strategy to eliminate the phenomenon of DD (Wilson and Izmailov 2020), (Loureiro et al. 2022).

7.3.1 Setting

We consider an ensemble of K predictors, each of them depending on the parameter $w_k \in \mathbb{R}^p$ for $k \in [K]$, and independently trained on the dataset $(x_i, y_i)_{i=1}^n$.
We assume that the data is generated from a target

$$y = \theta^{*T} x \quad \theta^* \sim \mathcal{N}(0, \rho I)$$

and consider the following family of random feature ensemble predictors

$$f(x) = \frac{1}{K} \sum_{k=1}^{K} \frac{w_k^\top \phi(F_k x)}{\sqrt{p}}$$

Where $\{F_k\}_{k=1}^K$ with $F_k \in \mathbb{R}^{p \times d}$ is a family of independently sampled random matrices and $\phi : \mathbb{R} \to \mathbb{R}$ is an activation function acting component-wise on $F_k x$.
The models are trained independently through the empirical risk minimization:

$$\hat{w}_k = \arg\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left\| y_i - \frac{w^\top \phi(F_k x_i)}{\sqrt{p}} \right\|^2$$

Theorem 7.3 *(Loureiro et al. 2022) Suppose that as $d, n, p \to \infty$ their ratio $\alpha = \frac{p}{n} \in]0, \infty[, \gamma = \frac{d}{p} \in]0, \infty[$, and the matrix $FF^\top$ has a well-defined asymptotic spectral distribution. Then in the limit the generalization error satisfies*

$$\mathbb{E}_{(x,y)}(\|y - \hat{f}(x)\|^2) = \rho + \frac{q_0 - q_1}{K} + q_1 - 2m \xrightarrow{K \to \infty} \rho + q_1 - 2m$$

with $\hat{f}(x) = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{w}_k^\top \phi(F_k x)}{\sqrt{p}}$, m, q_0 and q_1 are given by the following equations

$$\hat{v} = \frac{1}{\alpha(1+v)} \qquad \hat{q}_0 = \frac{\rho - 2m + q_0}{\alpha(1+v)^2},$$
$$\hat{m} = \frac{1}{\alpha(1+v)} \frac{1}{\sqrt{\gamma}} \qquad \hat{q}_1 = \frac{\rho - 2m + q_1}{\alpha(1+v)^2}.$$

Proof: See the proof in (Loureiro et al. 2022).

The theorem above provides a characterization of the random feature ensemble in the asymptotic setting. The authors claim that the turn q_0 is responsible for the fluctuation and hence DD, and is suppressed by increasing the number of predictors k. Notably, the authors did not derive explicit closed-form expressions for either the fluctuation term q_0 or the fluctuation-free term q_1. Furthermore, the auxiliary variables $\hat{v}, \hat{m}, \hat{q}_0$ and $\hat{q}_1$ introduced to capture the effective quantities in the high-dimensional ensemble of random feature models are also not provided in closed form. Solving the associated self-consistent equations requires numerical techniques, primarily due to their dependence on integrals over the spectral distribution of the covariance matrix.

References

Ba, Jimmy et al. (2020). "Generalization of Two-Layer Neural Networks: An Asymptotic Viewpoint". In: URL: https://openreview.net/pdf?id=H1gBsgBYwH.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, et al. (2019). "Reconciling Modern Machine Learning Practice and the Bias-Variance Trade-off". In: *arXiv:1812.11118*.

Belkin, Mikhail, Daniel Hsu, and Ji Xu (2020). "Two Models of Double Descent for Weak Features". In: *arXiv:1903.07571*.

Breiman, Leo and David Freedman (1983). "How Many Variables Should Be Entered in a Regression Equation?" In: *Journal of the American Statistical Association* 78.381, pp. 131–136.

Chen, Zhijun and Hayden Schaeffer (2021). "Conditioning of Random Feature Matrices: Double Descent and Generalization Error". In: *arXiv:2110.11477*.

Cheng, Xiuyuan and Amit Singer (2012). "The Spectrum of Random Inner-Product Kernel Matrices". In: *arXiv:1202.3155*.

Cook, R. Dennis and Sanford Weisberg (1982). "Residuals and Influence in Regression". In: *Journal of the American Statistical Association*.

Foucart, Simon and Holger Rauhut (2013). *A Mathematical Introduction to Compressive Sensing*. Springer.

Hastie, Trevor et al. (2020). "Surprises in High-Dimensional Ridgeless Least Squares Interpolation". In: *arXiv:1903.08560*.

Heckel, Reinhard and Fatih Furkan Yilmaz (2020). "Early Stopping in Deep Networks: Double Descent and How to Eliminate It". In: *arXiv:2007.10099*.

Koh, Pang Wei and Percy Liang (2017). "Understanding Black-box Predictions via Influence Functions". In: *arXiv:1703.04730*.

Kuzborskij, Ilja et al. (2021). "On the Role of Optimization in Double Descent: A Least Squares Study". In: *arXiv:2107.12685*.

Louart, Cosme, Zhenyu Liao, and Romain Couillet (2018). "A Random Matrix Approach to Neural Networks". In: *The Annals of Applied Probability* 28.2, pp. 1190–1248.

Loureiro, Bruno et al. (2022). "Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High Dimension". In: *arXiv:2201.13383*.

Mei, Song and Andrea Montanari (2019). "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve". In: *arXiv:1908.05355*.

Nakkiran, Preetum, Gal Kaplun, et al. (2019). "Deep Double Descent: Where Bigger Models and More Data Hurt". In: *arXiv:1912.02292*.

Nakkiran, Preetum, Prayaag Venkat, et al. (2021). "Optimal Regularization Can Mitigate Double Descent". In: *arXiv:2003.01897*.

Pennington, Jeffrey and Pratik Worah (2017). "Nonlinear Random Matrix Theory for Deep Learning". In: *Advances in Neural Information Processing Systems*, pp. 2637–2646.

Singh, Sidak Pal, Gerhard Bachmann, and Thomas Hofmann (2021). "Analytic Insights into Structure and Rank of Neural Network Hessian Maps". In: *arXiv:2106.16225*.

Singh, Sidak Pal, Aurelien Lucchi, et al. (2022). "Phenomenology of Double Descent in Finite-Width Neural Networks". In: *arXiv:2203.07337*.

Vallet, Cailton and Refregier (1989). "Linear and nonlinear extension of the pseudoinverse solution for learning boolean functions". In: *Europhysics Letters, 9(4):315, 1989*.

Vershynin, Roman (2011). "Introduction to the Non-Asymptotic Analysis of Random Matrices". In: *arXiv:1011.3027*.

Wilson, Andrew G. and Pavel Izmailov (2020). "Bayesian Deep Learning and a Probabilistic Perspective of Generalization". In: *arXiv:2002.08791*.

8 Omitted Proofs

8.1 Proof of Theorem 3.1

Under-parameterized regime ($p \leq n$): We have the following lemma by (Breiman and Freedman 1983).

Lemma 8.1 *((Breiman and Freedman 1983)) Define*

$$\sigma^2 = Var(\epsilon), \quad \sigma_p^2 = Var(\sum_{p+1}^{\infty} \theta_i x_i | x_1, ..., x_p), \quad U_{n,p} = \mathbb{E}(\mathbb{E}((y - \hat{y}_{n+1})^2 | (y_i, x_j)_{i \leq n, j \leq n})),$$

where $\hat{y}_{n+1}$ is the estimate for y after training on a set of size n. If $p \leq n-2$, then

$$U_{n,p} = (\sigma^2 + \sigma_p^2)\left(1 + \frac{p}{n-p-1}\right),$$

and for $n \geq p \geq n-1$, we have $U_{n,p} = \infty$.

Since in the setting of theorem 3.1 $x \sim \mathcal{N}(0, I_p)$, we have

$$\sigma_p^2 = Var(\sum_{p+1}^{\infty} \theta_i x_i | x_1, ..., x_p) = Var(\sum_{i \in P^c} \theta_i x_i) = \sum_{i \in P^c} \theta_i^2 Var(x_1) = \sum_{i \in P^c} \theta_i^2 = \|\theta_{P^c}\|^2.$$

Furthermore,

$$U_{n,p} = \mathbb{E}(\mathbb{E}((y - \hat{y}_{n+1})^2 | (y_i, x_j)_{i \leq n, j \leq n})) = \mathbb{E}((y - \hat{y})^2) = \mathbb{E}((y - x^\top \hat{\theta})^2).$$

Hence,

$$\mathbb{E}((y - x^\top \hat{\theta})^2) = \begin{cases} (\sigma^2 + \|\theta_{P^c}\|^2)\left(1 + \frac{p}{n-p-1}\right) & if\ p \leq n-2 \\ \infty & if\ n \geq p \geq n-1. \end{cases}$$

Over-parameterized regime ($p \geq n$) The mean-squared-error prediction can be written as

$$\mathbb{E}_x((y - x^\top \hat{\theta})^2) = \sigma^2 + \|\theta - \hat{\theta}\|^2 = \sigma^2 + \|\theta_P - \hat{\theta}_P\|^2 + \|\theta_{P^c} - \hat{\theta}_{P^c}\|^2.$$

since $\hat{\theta}_{P^c} = 0$, we have that the risk of $\hat{\theta}$ is

$$\mathbb{E}_x((y - x^\top \hat{\theta})^2) = \sigma^2 + \mathbb{E}(\|\theta_P - \hat{\theta}_P\|^2) + \|\theta_{P^c}\|^2.$$

In this regime we know that the pseudo-inverse $X_p^\dagger$ is given by $X_P^\dagger = X_P^\top (X_P X_P^\top)^{-1}$. Set $\eta = y - X_P \theta_P$. Then

$$\begin{aligned}\theta_P - \hat{\theta}_P &= \theta_P - X_P^\top (X_P X_P^\top)^{-1}(\eta + X_P \theta_P) \\ &= (I - X_P^\top (X_P X_P^\top)^{-1} X_P)\theta_P - X_P^\top (X_P X_P^\top)^{-1}\eta.\end{aligned}$$

We notice that $X_P(I - X_P^\top (X_P X_P^\top)^{-1})\theta_P = (X_P - X_P)\theta_P = 0$, which implies that $(I - X_P^\top (X_P X_P^\top)^{-1} X_P)$ is in the null space $N(X_P)$ of X_P. Furthermore, by taking $v = -(X_P X_P^\top)\eta$, we have that $X_P^\top v = -X_P^\top (X_P X_P^\top)\eta$ which means that the vector $-X_P^\top (X_P X_P^\top)\eta$ is in the row space of X_P. By the Pythagorean theorem we have that

$$\|\hat{\theta}_P - \theta_P\|^2 = \|(I - X_P^\top (X_P X_P^\top)^{-1} X_P)\theta_P\|^2 + \|X_P^\top (X_P X_P^\top)^{-1}\eta\|^2.$$

First term. Note that $\Pi_P := X_P^\top (X_P X_P^\top)^{-1} X_P$ is an orthogonal projection for the row space of X_P. Indeed,

$$\Pi_P^2 = X_P^\top (X_P X_P^\top)^{-1} X_P X_P^\top (X_P X_P^\top)^{-1} X_P = X_P^\top (X_P X_P^\top)^{-1} X_P = \Pi_P,$$

and

$$\Pi_P^\top = (X_P^\top (X_P X_P^\top)^{-1} X_P)^\top = X_P^\top (X_P X_P^\top)^{-1} X_P = \Pi_P.$$

Hence, Pi_P is an orthogonal projection. Therefore $\Pi_P \theta_P \perp \theta_P$, and by the Pythagorean theorem

$$\|\theta_P - \Pi_P \theta_P\|^2 = \|\theta_P\|^2 - \|\Pi_P \theta_P\|^2.$$

By rotational symmetry of the standard normal distribution, it follows that

$$\mathbb{E}(\|\Pi_p \theta_P\|^2) = \frac{n}{p}\|\theta_P\|^2.$$

Therefore,

$$\mathbb{E}(\|(I - X_P^\top (X_P X_P^\top)^{-1} X_P)\theta_P\|^2) = \|\theta_P\|^2 (1 - \frac{n}{p})$$

Second term. we use the trace trick $a^\top b = tr(a^\top b)$ to write

$$\begin{aligned}\|X_P^\top (X_P X_P^\top)^{-1}\eta\|^2 &= tr((X_P^\top (X_P X_P^\top)^{-1}\eta)^\top X_P^\top (X_P X_P^\top)^{-1}\eta) \\ &= tr(\eta^\top (X_P X_P^\top)^{-1} X_P X_P^\top (X_P X_P^\top)^{-1}\eta) \\ &= tr(\eta^\top (X_P X_P^\top)^{-1}\eta) \\ &= tr((X_P X_P^\top)^{-1}\eta^\top \eta).\end{aligned}$$

Using the fact that $\eta = y - X_P\theta_p = \epsilon + X\theta - X_P\theta_p = \epsilon + X_{P^c}\theta_{P^c}$, and the fact that $\epsilon + X_{P^c}\theta_{P^c}$ is independent from $X_P\theta_P$, we get

$$\mathbb{E}(\|X_P^\top (X_P X_P^\top)^{-1}\eta\|^2) = tr(\mathbb{E}(X_P X_P^\top)^{-1})\mathbb{E}(\eta\eta^\top)).$$

The distribution of η is normal with mean 0 and covariance $Var(\eta) = Var(\epsilon) + Var(X_{P^c}\theta_{P^c}) = \sigma^2 + \|\theta_{P^c}\|^2$, so

$$\mathbb{E}(\eta\eta^\top) = (\sigma^2 + \|\theta_{P^c}\|^2) I_n.$$

On the other hand, the matrix $P^{-1} = (X_P X_P^T)$, with $X_P \sim \mathcal{N}(0, I_p)$ is a scatter matrix, and has a Wishart distribution $\mathcal{W}_n(p, I)$. Hence, the distribution of $P = (X_P X_P^\top)^{-1} \sim \mathcal{W}_n^{-1}(p - n - 1, I)$ is inverse-Wishart with identity scale matrix I_n and $p - n - 1$ degrees of freedom. Hence

$$\mathbb{E}(P) = \begin{cases} \frac{1}{p-n-1} I_n & if\ p \geq n+2 \\ \infty & if\ p \in n, n+1 \end{cases}$$

We conclude that

$$\mathbb{E}(\|X_P^\top (X_P X_P^\top)^{-1}\eta\|^2) = \begin{cases} (\|\theta_P\|^2 + \sigma^2)\frac{n}{p-n-1} & if\ p \geq n+2 \\ \infty & if\ p \in n, n+1 \end{cases}$$

combining the first and second terms gives the expression for the risk.

8.2 Proof of Theorem 4.3

This proof is taken from (Ba et al. 2020).

Following the bias-variance decomposition 14, we have

$$\begin{aligned}
V(\hat{\theta}) &= \mathbb{E}_{x,\epsilon}\left[\left(\hat{\theta}^\top \phi(Wx) - \mathbb{E}_\epsilon\left[\hat{\theta}^\top \phi(Wx)\right]\right)^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\left((\Phi(XW^\top)^\dagger y)^\top \phi(Wx) - \mathbb{E}_\epsilon\left[(\Phi(XW^\top)^\dagger y)^\top \phi(Wx)\right]\right)^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\left(y^\top(\Phi(XW^\top)^\dagger)^\top \phi(Wx) - \mathbb{E}_\epsilon\left[y^\top(\Phi(XW^\top)^\dagger)^\top \phi(Wx)\right]\right)^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\left((y^\top - \mathbb{E}_\epsilon(y^\top))(\Phi(XW^\top)^\dagger)^\top \phi(Wx)\right)^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\left((F(X)^\top + \epsilon^\top - \mathbb{E}_\epsilon(F(X)^\top + \epsilon^\top))(\Phi(XW^\top)^\dagger)^\top \phi(Wx)\right)^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\left((F(X)^\top + \epsilon^\top - F(X)^\top - \mathbb{E}_\epsilon(\epsilon^\top))(\Phi(XW^\top)^\dagger)^\top \phi(Wx)\right)^2\right] \\
&= \mathbb{E}_{x,\epsilon}\left[\left((\epsilon^\top - \mathbb{E}_\epsilon(\epsilon^\top))(\Phi(XW^\top)^\dagger)^\top \phi(Wx)\right)^2\right] \\
&= \sigma^2 \mathrm{tr}\left(\Phi(XW^\top)^\dagger \Phi(WX^\top)^\dagger \mathbb{E}_x\left[\phi(Wx)\phi(Wx)^\top\right]\right) \\
&= \mathrm{tr}\left(\Phi(XW^\top)^\dagger \Phi(WX^\top)^\dagger K_W\right)
\end{aligned}$$

Where we used in the above computation the following facts: $\hat{\theta} = \Phi(XW^\top)^\dagger y$, $\Phi(X^\top W)$ and $F(X)$ are independent of ϵ, $y = F(X) + \epsilon$, σ is omitted (normalized) and define the expected non-linear Gram matrix $K_W \in \mathbb{R}^{p\times p}$ as:

$$K_W = \mathbb{E}_x\left[\phi(Wx)\phi(Wx)^\top\right].$$

We now that the Gram matrix of a nonlinear activation is almost surely full-rank, as specified in the following lemma from (Pennington and Worah 2017).

Lemma 8.2 *Suppose ϕ is not linear. Then the smallest singular value of $\Phi = \phi(X^\top W) \in \mathbb{R}^{n\times p}$ is of order $O(\sqrt{n})$. To be precise, consider the empirical spectral density $\mu_{\Phi\Phi^\top/n}(d\lambda)$. When $n, d, h \to \infty$ with $d/n \to \gamma$ and $p/n \to \psi$, we have*

$$\mu_{\Phi\Phi^\top/n}(d\lambda) \to [1 - \gamma_2^{-1}]_+ \delta_0(\lambda) + \mu_+(d\lambda),$$

where $\mu_+(d\lambda)$ has non-negative support $[\rho, \infty)$ with $\rho > 0$.

We therefore consider two scenarios: Φ is full column rank (Case I) or full row rank (Case II).

Case 1: $(p < n)$. In this case, the variance simplifies to:

$$V = \mathrm{tr}\left(\left(\phi(XW^\top)\phi(WX^\top)\right)^{-1} K_W\right) = \lim_{\xi\to 0^-} \mathrm{tr}\left(\left(\Phi\Phi^\top - \xi I\right)^{-1} K_W\right) = \lim_{\xi\to 0^-} V_\xi,$$

where the continuity and boundedness of V_ξ at $\xi = 0^-$ is guaranteed by Lemma 8.2 when $\psi = \frac{p}{n} \neq 1$.

A ridge ξ is added to make use of ((Louart et al. 2018) Theorem 1), which derived the asymptotic

equivalent of the resolvent: $\left(\Phi\Phi^\top - \xi I\right)^{-1}$. It follows that as $n, d, p \to \infty$,

$$\begin{aligned}
&\operatorname{tr}\left(p^{-1}\left(\Phi\Phi^\top - \xi I\right)^{-1}\left(\frac{n}{p}\frac{K_W}{1+p^{-1}\operatorname{tr}\left(p\left(\Phi\Phi^\top - \xi I\right)^{-1}K_W\right)}\right) - p^{-1}I\right) \\
&\leq \frac{1}{h}\left\|\left(\Phi\Phi^\top - \xi I\right)^{-1} - \left(\frac{n}{p}\frac{K_W}{1+p^{-1}\operatorname{tr}\left(p\left(\Phi\Phi^\top - \xi I\right)^{-1}K_W\right)}\right)^{-1}\right\|_F \cdot \left\|\frac{n}{p}\frac{K_W}{1+p^{-1}\operatorname{tr}\left(p\left(\Phi\Phi^\top - \xi I\right)^{-1}K_W\right)}\right\|_2 \\
&\leq \frac{1}{p}O(n^{-1/2+\epsilon})O(n^{1/2}) \to 0.
\end{aligned}$$

Where we have used the inequality $\operatorname{tr}(AB) \leq \|A\|_F\|B\|_2$, and equivalently, by taking $\xi \to 0$, we get

$$\lim_{\xi\to 0}\lim_{n,d,p\to\infty}\frac{n}{p}\frac{\operatorname{tr}\left(\left(\Phi\Phi^\top - \xi I\right)^{-1}K_W\right)}{1+\operatorname{tr}\left(\left(\Phi\Phi^\top - \xi I\right)^{-1}K_W\right)} = 1.$$

Therefore,

$$\lim_{\xi\to 0}\lim_{n,d,p\to\infty}\frac{n}{p}\cdot\operatorname{tr}\left(\left(\Phi\Phi^\top - \xi I\right)^{-1}K_W\right) = \frac{\psi}{1-\psi}.$$

Note that $\frac{\partial V_\xi}{\partial \xi} = \operatorname{tr}\left(\xi\left(\Phi\Phi^\top - \xi I\right)^{-2}K_W\right)$ is bounded around the neighborhood of $\xi = 0$, hence, following the same argument as ((Hastie et al. 2020), Theorem 4), we can exchange the limits of $\xi \to 0$ and $n, d, p \to \infty$:

$$V \to \frac{\psi}{1-\psi}.$$

Case 2: $(p > n)$. Instead of calculating the variance, (Ba et al. 2020) analyse a modified d quantity V_ξ and then take $\xi \to 0$, which can be connected to the trace of the resolvent of a matrix $\tilde{A}$ defined later; this translates the calculation of V_ξ into the calculation of the Stieltjes transform of $\tilde{A}$. The proof is done in two steps:

step 1 : The variance can be written as:

$$V = \operatorname{tr}(\Phi^\top(\Phi\Phi^\top)^{-2}\Phi K_W)$$

we take $S = \Phi/\sqrt{n}$, then by the same continuity argument provided by lemma 8.2, we have

$$V = \lim_{\xi\to 0}\frac{1}{n}\left[\operatorname{tr}\left(S^\top(SS^\top - \xi I)^{-2}SK_W\right)\right] = \lim_{\xi\to 0}V_\xi.$$

with $\xi \in \mathbb{C}$ and $\operatorname{Im}(\xi) > 0$ or $\xi < 0$.

We decompose the normalized feature matrix $S = \frac{\phi(XW^\top)}{\sqrt{n}}$ as $S = U\Sigma V^\top$, where $\Sigma = \operatorname{diag}(\phi_1, \ldots, \phi_n) \in \mathbb{R}^{n\times p}$ is a tall diagonal matrix, and $U = [u_1, \ldots, u_p] \in \mathbb{R}^{n\times n}$ is the set of orthogonal eigenvectors of $SS^\top = \frac{\phi(XW^\top)\phi(WX^\top)}{n}$, and $V \in \mathbb{R}^{p\times p}$ is the set of orthogonal eigenvectors of $S^\top S = \frac{\phi(WX^\top)\phi(XW^\top)}{n}$.

Now the variance can be written as

$$V_\xi = \frac{1}{n}\operatorname{tr}\left(\left(\Sigma V^\top\right)^\top\left(\Sigma\Sigma^\top + \xi I_n\right)^{-2}\Sigma V^\top K_W\right).$$

By the same argument as in Lemma 13 of (Hastie et al. 2020), one can show that when $n, d, p \to \infty$,

$$V_\xi = \frac{1}{n}\mathrm{tr}\left((\Sigma V^\top)^\top \left(\Sigma\Sigma^\top + \xi I_n\right)^{-2} \Sigma V^\top K_W\right) \to \frac{1}{n}\mathrm{tr}\left((\Sigma V^\top)^\top \left(\Sigma\Sigma^\top + \xi I_n\right)^{-2} \Sigma V^\top \tilde{K}_W\right)$$

in which $\tilde{K}_W$ is the approximation of K_W defined in Lemma **??**. Writing the trace explicitly (denote eigenvalues $\lambda_i = \phi_i^2$, and $\phi_{n+1} = \cdots = \phi_p = 0$), we have

$$V_\xi \to \frac{1}{n}\mathrm{tr}\left((\Sigma V^\top)^\top \left(\Sigma\Sigma^\top + \xi I_n\right)^{-2} U\Sigma\tilde{K}_W\right) = \frac{\psi}{p}\sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+\xi)^2} v_i^\top \tilde{K}_W v_i.$$

Since the positive support of the spectrum λ is lower bounded and the density at 0 is $1 - \psi^{-1}$, we have

$$V_\xi \to \psi \frac{1}{p}\lim_{p,d,n\to\infty} \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+\xi)^2} u_i^\top \tilde{K}_W u_i = \psi \int \frac{\lambda}{(\lambda+\xi)^2}\mu_\infty(d\lambda) = \psi \int_{\lambda>\rho} \frac{\lambda}{(\lambda+\xi)^2}\mu_\infty^+(d\lambda),$$

where we define $\mu_n(x) = \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i}(x) u_i^\top \tilde{K}_W u_i$ and its positive part $\mu_n^+(x)$. Hence, we have

$$V = \lim_{\xi\to 0} V_\xi = \lim_{\xi\to 0} \psi \int_{\lambda>\rho} \frac{\lambda}{(\lambda+\xi)^2}\mu_\infty^+(d\lambda) = \psi \int_{\lambda>\rho} \frac{1}{\lambda}\mu_\infty^+(d\lambda).$$

We define the following matrix $\tilde{A}_n(\rho, \varsigma, \tau) \in \mathbb{R}^{N\times N}$, where $N = n + p$:

$$\tilde{A}_n(\rho,\varsigma,\tau) = \begin{bmatrix} \rho I_p + \varsigma \mathbf{1}_p\mathbf{1}_p^\top + \tau Q & S^\top \\ S & 0_n \end{bmatrix}.$$

Denote the Stieltjes transform of $\tilde{A}_n$ as:

$$\tilde{m}_n(\xi,\rho,\varsigma,\tau) = \frac{1}{n}\mathrm{tr}\left(\left(\tilde{A}_n(\rho,\varsigma,\tau) - \xi I_N\right)^{-1}\right). \tag{57}$$

Then, following the definition of $\tilde{K}_W$, one can show that:

$$\tilde{m}_n(\xi, r_x, s_x, t_x) = \frac{1}{n}\mathrm{tr}\begin{bmatrix} \tilde{K}_W x - \xi I_p & S^\top \\ S & -I_n \end{bmatrix}^{-1}. \tag{58}$$

Taking the matrix derivative gives:

$$\begin{aligned}
\frac{\partial}{\partial x}\tilde{m}_n(\xi, r_x, s_x, t_x)\big|_{x=0} &= \frac{1}{n}\mathrm{tr}\left(\begin{bmatrix} -\xi I_p & S^\top \\ S & -I_n \end{bmatrix}^{-1}\begin{bmatrix} \tilde{K}_W & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} -\xi I_p & S^\top \\ S & -I_n \end{bmatrix}^{-1}\right) \\
&= \frac{1}{n}\mathrm{tr}\left(\begin{bmatrix} \xi^2 I_p + S^\top S & 0 \\ 0 & I_n + SS^\top \end{bmatrix}^{-1}\begin{bmatrix} \tilde{K}_W & 0 \\ 0 & 0 \end{bmatrix}\right) \\
&= \frac{1}{n}\mathrm{tr}\left(U(\Sigma\Sigma^\top + \xi^2 I_p)^{-1}U^\top \tilde{K}_W\right),
\end{aligned}$$

where U and Σ are the eigenvector and eigenvalue matrices of $SS^\top$. Explicitly, this becomes:

$$= \frac{1}{n}\sum_{i=1}^{p}\frac{1}{\lambda_i+\xi^2}v_i^\top \tilde{K}_W v_i.$$

Denote the limit $\tilde{m}(\xi,\rho,\varsigma,\tau) = \lim_{n,h,d\to\infty}\tilde{m}_n(\xi,\rho,\varsigma,\tau)$, and the derivative is given as:

$$-\frac{\partial}{\partial x}\tilde{m}(\xi,r_x,s_x,t_x)\big|_{x=0} = \lim_{h,d,n\to\infty}\frac{1}{n}\sum_{i=1}^{p}\frac{1}{\lambda_i+\xi^2}v_i^\top \tilde{K}_W v_i.$$

Simplifying further:

$$= \psi\int_{\lambda\geq 0}\frac{1}{\lambda+\xi^2}\mu_\infty(d\lambda) = \gamma_2\left(-\frac{1}{\xi^2}\right) + \gamma_2\int_{\lambda>\rho}\frac{1}{\lambda+\xi^2}\mu_\infty^+(d\lambda).$$

For simplicity, we define the following function on ξ:

$$q(\xi) = -\frac{\partial}{\partial x}\tilde{m}(\xi,r_x,s_x,t_x)\big|_{x=0},$$

and also denote:

$$q^+(\xi) = q(\xi) - \frac{\psi-1}{\xi^2} = \gamma_2\int_{\lambda>\rho}\frac{1}{\lambda+\xi^2}\mu_\infty^+(d\lambda).$$

Hence

$$V = \lim_{\xi\to 0} q^+(\xi).$$

Step 2: Calculating $q(\xi)$ and $m_n(\xi,\rho,\varsigma,\tau)$.

We define A_n by subtracting the off-diagonal entries of the upper-left block of $\tilde{A}_n$:

$$A_n(\rho,\tau) = \begin{bmatrix} \rho I_n + \tau Q & S^\top \\ S & 0_n \end{bmatrix}, \tag{63}$$

where $S = \tilde{S} - a_0 I_{n\times p}$, i.e., $S_{ik} = \phi(x_i^\top w_k) - a_0 = \varphi(x_i^\top w_k)$, with $a_0 = \mathbb{E}[\phi(x)]$.

The Stieltjes transform of A_n is given by:

$$m_n(\xi,\rho,\tau) = \frac{1}{n}\mathrm{tr}\left(A_n(\rho,\tau) - \xi I_N\right)^{-1}.$$

The following lemma shows that $\tilde{m}_n$ and m_n have the same limit.

Lemma 8.3 *When $n\to\infty$, and for $Im(\xi)>0$ or $\xi<0$, we have:*

$$m_n(\xi,\rho,\tau) \to \tilde{m}_n(\xi,\rho,\varsigma,\tau).$$

Proof: See the proof in (Ba et al. 2020).

To calculate the Stieltjes transform m_n, we take advantage of the block structure of A_n:

$$m_{1,n}(\xi,\rho,\tau) = \frac{1}{p}\mathrm{tr}\left((A_n(\rho,\tau)-\xi I_N)^{-1}\big|_{[1..p,1..p]}\right),$$

$$m_{2,n}(\xi,\rho,\tau) = \frac{1}{n}\mathrm{tr}\left((A_n(\rho,\tau)-\xi I_N)^{-1}\big|_{[p+1..p+n,p+1..p+n]}\right).$$

One can observe that:

$$m_n(\xi,\rho,\tau) = \psi m_{1,n}(\xi,\rho,\tau) + m_{2,n}(\xi,\rho,\tau).$$

In the following equations, we omit the subscript n, as well as the dependency on ρ,ς,τ. Following (Hastie et al. 2020), we rewrite the matrix $A_n = A$ as:

$$A = \begin{bmatrix} A^* & a^\top \\ a & 0 \end{bmatrix},$$

where A^* is a $(N-1)\times(N-1)$ matrix with the last column and row of A removed, and a is the activation vector:

$$A^* = \begin{bmatrix} \rho I_p + \tau Q & S^{*\top} \\ S^* & 0_{n-1} \end{bmatrix}, \quad a = \begin{bmatrix} \phi(W^\top x_n) & 0_{n-1} \end{bmatrix} = \begin{bmatrix} s & 0_{n-1} \end{bmatrix}.$$

Hence, by the block matrix inverse formula:

$$(A-\xi I_N)^{-1} = \begin{bmatrix} (A^*-\xi I_{N-1})^{-1} & * \\ * & \left[-\xi - a(A^*-\xi I_{N-1})^{-1}a^\top\right]^{-1} \end{bmatrix}.$$

Plugging this back into the Stieltjes transform, we have:

$$m_{2,n}(\xi,\rho,\tau) = \frac{1}{n}\mathrm{tr}\left((A_n(\rho,\varsigma,\tau)-\xi I_N)^{-1}\big|_{[p+1,N]}\right) = \mathbb{E}_a\left[\left(-\xi - a(A^*-\xi I_{N-1})^{-1}a^\top\right)^{-1}\right].$$

To obtain an asymptotic description of $m_{2,n}$, we perform the orthonormal decomposition on the nonlinearity φ introduced in (Cheng and Singer 2012):

$$\varphi(x) = a_1 x + \varphi^\perp(x),$$

where $a_1 = \mathbb{E}_{x\sim N(0,1)}[x\varphi(x)]$.

For each w_i (where $1 \le i \le p$), we perform the following orthonormal decomposition along the direction of x_n and the direction of $\tilde{w}_i$ perpendicular to x_n:

$$w_i = \frac{w_i^\top x_n}{\|x_n\|}\, x_n + \tilde{w}_i = \eta_i \frac{x_n}{\|x_n\|} + \tilde{w}_i.$$

We can thus simplify the activation vector a as:

$$\begin{aligned} a &= \left[\tfrac{1}{\sqrt{n}}\varphi(\|x_n\|\eta_1) \quad \cdots \quad \tfrac{1}{\sqrt{n}}\varphi(\|x_n\|\eta_p) \quad \underbrace{0\cdots 0}_{n-1}\right] \\ &= \underbrace{\left[\tfrac{1}{\sqrt{n}}a_1\|x_n\|\eta_1 \quad \cdots \quad \tfrac{1}{\sqrt{n}}a_1 x_n\|\eta_p \quad \underbrace{0\cdots 0}_{n-1}\right]}_{a_1=[s_1,0]} \\ &\quad + \underbrace{\left[\tfrac{1}{\sqrt{n}}\varphi_\top(\|x_n\|\eta_1) \quad \cdots \quad \tfrac{1}{\sqrt{n}}\varphi_\top(\|x_n\|\eta_p) \quad \underbrace{0\cdots 0}_{n-1}\right]}_{a_2=[s_2,0]}. \end{aligned}$$

For $1 \le i \ne j \le p$, $1 \le k \le n-1$, we define the matrix Q_{ij} as follows:

$$Q_{ij} = \left(\eta_i \frac{x_n}{\|x_n\|} + \tilde{w}_i\right)^\top \left(\eta_j \frac{x_n}{\|x_n\|} + \tilde{w}_j\right) = \eta_i\eta_j + \underbrace{\tilde{w}_i^\top \tilde{w}_j}_{\tilde{Q}_{ij}}.$$

Similarly, we decompose the activation function in S:

$$\begin{aligned} S_{ki} &= \frac{1}{\sqrt{n}}\varphi\left(\eta_i \frac{x_n^\top x_k}{\|x_n\|} + \tilde{w}_i^\top x_k\right) = \frac{1}{\sqrt{n}}a_1\eta_i \frac{x_n^\top x_k}{\|x_n\|} + \frac{1}{\sqrt{n}}a_1\tilde{w}_i^\top x_k + \frac{1}{\sqrt{n}}\varphi^\perp\left(\eta_i \frac{x_n^\top x_k}{\|x_n\|} + \tilde{w}_i^\top x_k\right) \\ &= \underbrace{\frac{1}{\sqrt{n}}\varphi(\tilde{w}_i^\top x_k)}_{\tilde{S}_{ki}} + \underbrace{\frac{1}{\sqrt{n}}a_1\eta_i \frac{x_n^\top x_k}{\|x_n\|}}_{a_1\eta_i u_k} + \underbrace{\frac{1}{\sqrt{n}}\left(\varphi^\perp\left(\eta_i \frac{x_n^\top x_k}{\|x_n\|} + \tilde{w}_i^\top x_k\right) - \varphi^\perp(\tilde{w}_i^\top x_k)\right)}_{E_{ki}}. \end{aligned}$$

We thus have an equivalent expression for the matrix A^*:

$$\begin{aligned} A^* &= \begin{bmatrix} \rho I_p + \tau\tilde{Q} & \tilde{S}^{*\top} \\ \tilde{S}^* & 0_{n-1} \end{bmatrix} + \begin{bmatrix} t\eta\eta^\top & a_1\eta u^\top \\ a_1 u\eta^\top & 0_{n-1} \end{bmatrix} + \begin{bmatrix} E_0 & E_1^\top \\ E_1 & 0_{n-1} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \rho I_p + \tau\tilde{Q} & \tilde{S}^{*\top} \\ \tilde{S}^* & 0_{n-1} \end{bmatrix}}_{\tilde{A}^*} + \underbrace{\begin{bmatrix} \eta & 0_p \\ 0_{n-1} & u \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \tau a_1 & a_1 \\ a_1 & 0 \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} \eta & 0_p \\ 0_{n-1} & u \end{bmatrix}^\top}_{U^\top} + \underbrace{\begin{bmatrix} E_0 & E_1 \\ E_1^\top & 0_{n-1} \end{bmatrix}}_{E}. \\ &= \tilde{A}^* + UCU^\top + E. \end{aligned}$$

By an argument similar to ((Hastie et al. 2020), B.1.2), the matrix E diminishes to 0 as $n \to \infty$ with respect to the Frobenius norm. Therefore, by the Woodbury's identity and the expression for $m_{2,n}$, we

have:

$$
\begin{aligned}
m_{2,n}(\xi,\rho,\tau) &= \mathbb{E}_a\left[\left(-\xi - a(A^* - \xi I_{N-1})^{-1}a^\top\right)^{-1}\right] \\
&\to \mathbb{E}_a\left[\left(-\xi - a\left(\tilde{A}^* - \xi I_{N-1} + UCU^\top\right)^{-1}a^\top\right)^{-1}\right] \\
&= \mathbb{E}_a\left[\left(-\xi - \underbrace{a\left(\tilde{A}^* - \xi I_{N-1}\right)^{-1}a^\top}_{u} + \underbrace{a(\tilde{A}^* - \xi I_{N-1})^{-1}U}_{v^\top}\right.\right. \\
&\qquad \underbrace{\left(C^{-1} + U^\top(\tilde{A}^* - \xi I_{N-1})^{-1}U\right)^{-1}}_{S} \\
&\qquad \left.\left.\underbrace{U(\tilde{A}^* - \xi I_{N-1})^{-1}a^\top}_{v}\right)^{-1}\right].
\end{aligned}
$$

We now bound each term u, v, and S to compute $m_{2,n}$. For u

$$\mathbb{E}_a u = E_a\left[a(A^{\tilde{*}} - \xi I_{N-1})^{-1}a^\top\right] = \operatorname{tr}\left[\mathbb{E}_s[ss^\top](\tilde{A}^* - \xi I_{N-1})^{-1}[1..p, 1..p]\right] = b\psi m_{1,n}(\xi,\rho,\tau),$$

where $b = \mathbb{E}_{x\sim N(0,1)}\left[\varphi(x)^2\right] = \mathbb{E}_{x\sim N(0,1)}\left[(\varphi(x) - E[\varphi(x)])^2\right] = r$.

From a standard concentration of measure argument, we have that as $n, p, d \to \infty$:

$$u \to \mathbb{E}_a u = r\psi m_{1,n}(\xi,\rho,\tau).$$

For v: Note that U depends on a. We compute:

$$\mathbb{E}_a v^\top = \mathbb{E}_a\left[a^\top(A^{\tilde{*}} - \xi I_{N-1})^{-1}U\right] = \mathbb{E}_a\left[\left[s^\top, 0_{n-1}^\top\right](\tilde{A}^* - \xi I_{N-1})^{-1}\begin{bmatrix}\eta & 0_p \\ 0_{n-1} & u\end{bmatrix}\right]$$

$$= \left[s^\top(\tilde{A}^\top - \xi I_{N-1})^{-1}[1..p, 1..p]\,\eta \quad 0\right] = \left[\operatorname{tr}\left((\tilde{A}^* - \xi I_{N-1})^{-1}[1..p, 1..p]\,\mathbb{E}_s[\eta s^\top]\right) \quad 0\right].$$

where

$$v = \operatorname{tr}\left((\tilde{A}^* - \xi I_{N-1})^{-1}[1..p, 1..p]\,\mathbb{E}_s[\eta s^\top]\right) = a_1\sqrt{\frac{\psi^2}{\gamma}}m_{1,n}(\xi,\rho,\tau),$$

which gives:

$$v \to \mathbb{E}_a v = \left[a_1\sqrt{\tfrac{\psi^2}{\gamma}}m_{1,n}(\xi,\rho,\tau) \quad 0\right].$$

as $n, d, p \to \infty$. And finally for S,

$$
\begin{aligned}
\mathbb{E}_a S^{-1} &= \mathbb{E}_a\left[\left(C^{-1} + U^\top(\tilde{A}^* - \xi I_{N-1})^{-1}U\right)\right] \\
&= \begin{bmatrix} \tau & a_1 \\ a_1 & 0 \end{bmatrix}^{-1} + \mathbb{E}_a\left[\left(\begin{bmatrix} \eta & 0_p \\ 0_{n-1} & u \end{bmatrix}^\top (\tilde{A}^* - \xi I_{N-1})^{-1} \begin{bmatrix} \eta & 0_p \\ 0_{n-1} & u \end{bmatrix}\right)\right] \\
&= \begin{bmatrix} 0 & \frac{1}{a_1} \\ \frac{1}{a_1} & -\frac{\tau}{a_1^2} \end{bmatrix} + \mathbb{E}_a\left[\begin{bmatrix} \eta^\top(\tilde{A}^* - \xi I_{N-1})^{-1}[1..p]\eta & 0 \\ 0 & u^\top(\tilde{A}^* - \xi I_{N-1})^{-1}[p+1..p+n-1]u \end{bmatrix}\right] \\
&= \begin{bmatrix} 0 & \frac{1}{a_1} \\ \frac{1}{a_1} & -\frac{\tau}{a_1^2} \end{bmatrix} + \begin{bmatrix} \frac{\psi}{\gamma m_{1,n}(\xi,\rho,\tau)} & 0 \\ 0 & m_{2,n}(\xi,\rho,\tau) \end{bmatrix} \\
&= \begin{bmatrix} \psi\gamma^{-1}m_{1,n}(\xi,\rho,\tau) & \frac{1}{a_1} & \\ \frac{1}{a_1} & m_{2,n}(\xi,\rho,\tau) & -\frac{\tau}{a_1^2} \end{bmatrix}.
\end{aligned}
$$

and hence as $n, d, p \to \infty$,

$$
S^{-1} \to \mathbb{E}_a S^{-1} = \begin{bmatrix} \psi\gamma^{-1}m_{1,n}(\xi,\rho,\tau) & \frac{1}{a_1} & \\ \frac{1}{a_1} & m_{2,n}(\xi,\rho,\tau) & -\frac{\tau}{a_1^2} \end{bmatrix}.
$$

Therefore, we arrive at the following expression on $m_{2,n}$,

$$
\begin{aligned}
m_{2,n}(\xi,\rho,\tau) &\to \mathbb{E}_a\left[\left(-\xi - u + v^\top S v\right)^{-1}\right] \\
&\to \left(-\xi - u + v^\top S v\right)^{-1} \\
&\to \left(-\xi - r\psi m_{1,n} + \left(a_1\sqrt{\frac{\psi^2}{\gamma}}m_{1,n}\right)^2\right) \\
&\left(\begin{bmatrix} \gamma^{-1}\psi m_{1,n}(\xi,\rho,\tau) & \frac{1}{a_1} \\ \frac{1}{a_1} & m_{2,n}(\xi,\rho,\tau) - \frac{\tau}{a^2} \end{bmatrix}^{-1}\right) \\
&= \left(-\xi - r\psi m_{1,n} + \frac{\psi a_1^2 m_{1,n}^2\left(a_1^2 m_{2,n} - \tau\right)}{m_{1,n}\left(a_1^2 m_{2,n} - \tau\right) - \gamma\psi^{-1}}\right)^{-1}.
\end{aligned}
$$

and for $m_{1,n}$,

$$
\begin{aligned}
&m_{1,n}(\xi,\rho,\tau) \to \\
&\left(-\xi - \rho - \gamma^{-1}\psi\tau^2 m_{1,n} - r m_{2,n} + \frac{\tau^2\gamma^{-1}\psi m_{1,n}^2\left(a_1^2 m_2 - \tau\right) - 2\tau a_1^2 m_{1,n}m_{2,n} + a_1^2 m_{1,n}m_2^2}{m_1\left(a_1^2 m_{2,n} - \tau\right) - \gamma\psi^{-1}}\right)^{-1}.
\end{aligned}
$$

The proof for the uniqueness of $m_{1,n}$ and $m_{2,n}$ is given in Lemma 11 of (Hastie et al. 2020).

8.3 Proof of Theorem 4.5

This proof is taken from (Chen and Schaeffer 2021) with some details added.

In the **under-parameterized regime**, we use the triangle inequality and then bound the three terms

in the inequality. First we consider the following rank 1 decomposition of $\Phi^*\Phi$

$$\frac{1}{n}\Phi^*\Phi = \frac{1}{n}\sum_{\ell=1}^{n} X_l X_l^*$$

with X_ℓ is the ℓ-the column of A^* Then, we have:

$$\begin{aligned}\left|\frac{1}{n}\lambda_k(\Phi^*\Phi) - 1\right| &\leq \left\|\frac{1}{n}\Phi^*\Phi - I_p\right\|_2 \\ &\leq \frac{1}{n}\sum_{\ell=1}^{n}\|X_\ell X_\ell^* - \mathbb{E}_x\left[X_\ell X_\ell^*\right]\|_2 + \frac{1}{n}\sum_{\ell=1}^{n}\|\mathbb{E}_x\left[X_\ell X_\ell^*\right] - \mathbb{E}_{x,\omega}\left[X_\ell X_\ell^*\right]\|_2 + \|L\|_2,\end{aligned}$$

Where $L = \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{E}_{x,\omega}\left[X_\ell X_\ell^*\right] - I_p = \frac{1}{n}\mathbb{E}_{x,\omega}\left[\Phi^*\Phi\right] - I_p$.
To bound this term, we notice that L is symmetric with $L_{jk} = \frac{1}{n}\mathbb{E}_{x,\omega}\left[(\Phi^*\Phi)_{jk}\right] - I_{p_{jk}} = \frac{1}{n}\mathbb{E}_{x,\omega}\left[\sum_{r=1}^{n}\exp(i\langle x_r, w_k - w_j\rangle)\right] - I_{p_{jk}}$. Hence, $L_{jj} = 0$, and for $j \neq k$ we have

$$L_{jk} = \frac{1}{n}\sum_{r=1}^{n}\mathbb{E}_{x,\omega}\left[\Pi_{l=1}^{d}\exp(ix_r^l(\omega_k^l - \omega_j^l))\right] = \frac{1}{n}\sum_{r=1}^{n}\mathbb{E}_{x,\omega}\left[\Pi_{l=1}^{d}\exp(ix_r^l\omega_k^l)\exp(-ix_r^l\omega_j^l)\right]$$

Since $\omega_k^l \sim \mathcal{N}(0,\sigma^2)$, the expectation over w_l is:

$$\mathbb{E}_{\omega_l}\left[e^{ix_j\omega_j}\right] = e^{-\frac{1}{2}\sigma_\omega^2 x_j^2}.$$

Because the ω_k^l's are independent for $l = 1...d$, the expectation over ω is:

$$\mathbb{E}_w\left[e^{i\langle x_r,\omega_k\rangle}\right] = \prod_{l=1}^{d} e^{-\frac{1}{2}\sigma^2 x_r^{l^2}} = e^{-\frac{1}{2}\sigma^2\|x_r\|^2}.$$

Now, we compute:

$$\mathbb{E}_x\left[e^{-\frac{1}{2}\sigma^2\|x_r\|^2}e^{-\frac{1}{2}\sigma^2\|x_r\|^2}\right] = e^{-\sigma^2\|x_r\|^2},$$

where $x_r \sim \mathcal{N}(0,\gamma^2 I_d)$.

Since $\|x_r\|^2$ follows a chi-squared distribution scaled by σ, specifically:

$$\|x_r\|^2 \sim 2\sigma^2\chi_d^2,$$

where χ_d^2 has d degrees of freedom.

The moment-generating function of a scaled chi-squared random variable $\sigma^2\chi_d^2$ is:

$$\mathbb{E}_{x,\omega}\left[e^{i\langle x_r,(\omega_k-\omega_j)\rangle}\right] = \left(1 + 2\gamma^2\sigma^2\right)^{-d/2}.$$

Hence, we get:

$$L_{jk} = \left(1 + 2\gamma^2\sigma^2\right)^{-d/2}$$

for $j \neq k$. Now, let z be a unit vector, then By Hölder's inequality, we have:

$$
\begin{aligned}
|\langle Lz, z\rangle| &= \left| \sum_{j,k=1, j\neq k}^{p} \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{-\frac{d}{2}} z_j \bar{z}_k \right| \\
&\leq \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} \|z\|_1^2. \\
&\leq p\left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{\frac{d}{2}} \\
&\leq p\left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{\frac{d}{4}} \\
&\leq \sqrt{\delta}\eta
\end{aligned}
$$

Where we used the assumption 19 in the last inequality. Therefore, the term $\|L\|_2$ is bounded by $\sqrt{\delta}\eta \leq \eta$. For the second term, define the random variable Z by

$$
Z(\omega_1, ..., \omega_p) = \|\mathbb{E}_x(X_\ell X_\ell^*) - \mathbb{E}_{x,\omega}(X_\ell X_\ell^*)\|_2
$$

Since $\mathbb{E}_x(X_\ell X_\ell^*)$ depends only on $\{\omega_k\}_{k\in[p]}$, the random variable Z is independent of $\{x_i\}_{i\in[n]}$. The second moment of Z_ℓ is bounded by

$$
\begin{aligned}
\mathbb{E}_\omega\left(Z(\omega_1, \ldots, \omega_p)^2\right) &= \mathbb{E}_\omega \|\mathbb{E}_x(X_\ell X_\ell^*) - \mathbb{E}_{x,\omega}(X_\ell X_\ell^*)\|_2^2 \\
&\leq \mathbb{E}_\omega \|\mathbb{E}_x(X_\ell X_\ell^*) - \mathbb{E}_{x,\omega}(X_\ell X_\ell^*)\|_F^2 \\
&= \sum_{j,k=1, j\neq k}^{N} \mathbb{E}_\omega \left(\exp\left(-\frac{\gamma^2}{2}\|\omega_j - \omega_k\|_2^2\right) - \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{d/2}\right)^2 \\
&= \sum_{j,k=1, j\neq k}^{N} \mathbb{E}_\omega \exp\left(-\gamma^2\|\omega_j - \omega_k\|_2^2\right) - \left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{d} \\
&\leq N^2\left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{d/2}.
\end{aligned}
$$

by Markov inequality and the assumption 19, we have

$$
\begin{aligned}
\mathbb{P}_\omega\left(Z(w) \geq \eta\right) &\leq \mathbb{P}_\omega\left(Z(w) \geq \frac{p}{\sqrt{\delta}}\left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{(d/4)}\right) \\
&\leq \frac{\delta(4\gamma^2\sigma^2+1)}{p^2}\mathbb{E}_\omega\left[Z(\omega)^2\right] \\
&\leq \delta.
\end{aligned}
$$

Hence, with probability at least $1-\delta$ over $\{w_k\}_{k\in[p]}$, we have $Z(\omega_1, ..., \omega_p) = \|\mathbb{E}_x(X_\ell X_\ell^*) - \mathbb{E}_{x,\omega}(X_\ell X_\ell^*)\|_2 \leq \eta$.

For the last term, we take $Y_\ell = X_\ell X_\ell - \mathbb{E}_x\left[X_\ell X_\ell\right]$. Then we have,

$$
Y_{j,k} = \exp(i\langle x_\ell, \omega_k - \omega_j\rangle) - \exp(-\gamma^2\|\omega_k - \omega_j\|^2/2).
$$

And for $k = j$, we have $(Y_\ell)_{jk} = 0$. The norms are bounded by

$$\begin{aligned}\|Y\|_2 &\leq \max_{j\in[p]} \sum_{k=1,k\neq j} \left|\exp(i\langle x_\ell, \omega_k - \omega_j\rangle) - \exp(-\gamma^2\|\omega_k - \omega_j\|^2/2)\right| \\ &\leq 2p \qquad\qquad \forall \ell \in [n]\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\left\|\sum_{\ell=1}^n \mathbb{E}_x(Y_\ell^2)\right\|_2 &\leq \sum_{\ell=1}^n \|\mathbb{E}_x(Y_\ell^2)\|_2 \\ &\leq \sum_{\ell=1}^n \|\mathbb{E}_x((X_\ell X_\ell^*)^2) + (\mathbb{E}_x X_\ell X_\ell^*)^2\|_2 \\ &\leq \sum_{\ell=1}^n \|p\mathbb{E}_x(X_\ell X_\ell^*) + (\mathbb{E}_x X_\ell X_\ell^*)^2\|_2 \\ &\leq \sum_{\ell=1}^n p\|\mathbb{E}_x(X_\ell X_\ell^*)\|_2 + \|\mathbb{E}_x(X_\ell X_\ell^*)\|_2^2 \\ &\leq n[p(1+2\eta) + (1+2\eta)^2].\end{aligned}$$

noting that $X_\ell^* X_\ell = \sum_{k=1}^p |X_\ell^k|^2 = \sum_{k=1}^p |\exp(i\langle x_k, w_\ell\rangle)|^2 = p$, and

$$\begin{aligned}\|\mathbb{E}_x X_\ell X_\ell^*\|_2 &\leq \|\mathbb{E}_x X_\ell X_\ell^* - \mathbb{E}_{x,\omega} X_\ell X_\ell^*\|_2 + \|\mathbb{E}_{x,\omega} X_\ell X_\ell^*\|_2 \\ &\leq \eta + \|I_p + L\|_2 \\ &\leq 2\eta + 1\end{aligned}$$

The next step makes use of the following Lemma.

Lemma 8.4 *Let $\{X_\ell\}_{\ell\in[n]}$ be a set of $\mathbb{C}^{p\times p}$ independent mean-zero self-adjoint random matrices, assume that $\|X_\ell\|^2 \leq K$ almost surely for all $\ell \in [n]$, and define $\sigma^2 := \left\|\sum_{\ell=1}^n \mathbb{E}(X_\ell^2)\right\|_2$.*
Then for all $t > 0$, we have

$$P\left(\sum_{\ell=1}^n \|X_\ell\|_2 \geq t\right) \leq 2p\exp\left(-\frac{t^2}{2\sigma^2 + \frac{2Kt}{3}}\right).$$

Proof: See (Corollary 8.15 from (Foucart and Rauhut 2013))
By setting the parameters in the Lemma above to be $\sigma^2 = n(\frac{5}{3}p + \frac{25}{9})$ and $K = 2p$ and assuming that $\eta \leq 1/3$, we get

$$\mathbb{P}\left(\frac{1}{n}\left\|\sum_{\ell=1}^n (X_\ell X_\ell^* - \mathbb{E}X_\ell X_\ell^*)\right\|_2 \geq \eta\right) = \mathbb{P}\left(\left\|\sum_{\ell=1}^n Y_\ell\right\|_2 \geq n\eta\right) \leq 2p\exp\left(-\frac{n\eta^2}{\frac{10}{3}p + \frac{50}{9} + \frac{4p\eta}{3}}\right).$$

If $n \geq 4p\eta^{-2}\log(\frac{2p}{\delta})$, (assuming $p > 25$), then with probability at least $1-\delta$ with respect to the draw

of $\{x_j\}_{j\in[n]}$ and the weights $\{\omega_k\}_{k\in[p]}$, we have

$$\frac{1}{n}\left\|\sum_{\ell=1}^{n}(X_\ell X_\ell^* - \mathbb{E}X_\ell X_\ell^*)\right\|_2 \leq \eta.$$

Hence,

$$\begin{aligned}\left|\lambda_k\left(\frac{1}{n}\Phi^*\Phi\right)\right| &\leq \frac{1}{n}\left\|\sum_{\ell=1}^{n}(X_\ell X_\ell^* - \mathbb{E}X_\ell X_\ell^*)\right\|_2 + \frac{1}{n}\left\|\sum_{\ell=1}^{n}(\mathbb{E}X_\ell X_\ell^* - \mathbb{E}_{x,\omega}X_\ell X_\ell^*)\right\|_2 + \|L\|_2 \\ &\leq 3\eta.\end{aligned}$$

For the **over-parameterized regime**, it is worth noting that the feature matrix Φ and its conjugate Φ^* have the same distribution since the data and weights have symmetric distribution around 0. Therefore, the proof for this regime is essentially the same as before with the roles of x and ω reversed.

Now we consider the **Interpolation point** $n = p$, the Gram matrix $\frac{1}{p}\Phi^*\Phi$ can be written as the following 1-rank decomposition

$$\frac{1}{p}\Phi^*\Phi = \sum_{\ell=1}^{p} X_\ell X_\ell^*.$$

where X_ℓ is a $\ell$the column of Φ^*. We can use the Rayleigh quotient to bound the maximum eigenvalue of the Gram matrix

$$\lambda_{\max} \geq \frac{1}{p}\langle\Phi^*\Phi v, v\rangle \quad \forall v \in \mathbb{C}^{p\times p} \quad \text{with} \quad \|v\|_1 = 1.$$

Setting $v = \frac{1}{\sqrt{N}}X_1$ gives

$$\begin{aligned}\lambda_{\max}\left(\frac{1}{p}\Phi^*\Phi\right) &\geq \frac{1}{p^2}\sum_{\ell=1}^{p} X_1^* X_\ell X_\ell^* X_1 \\ &= \frac{1}{p^2}\left(p^2 + \sum_{\ell=2}^{p} X_1^* X_\ell X_\ell^* X_1\right) \\ &= 1 + \frac{1}{p^2}\sum_{\ell=2}^{p}\sum_{j,k=1}^{p}\exp(i\langle x_1 - x_\ell, \omega_j - \omega_k\rangle) \\ &= 1 + \frac{p(p-1)}{p^2}\frac{1}{p^2}\sum_{\ell=2}^{p}\sum_{j,k=1 j\neq k}^{p}\exp(i\langle x_1 - x_\ell, \omega_j - \omega_k\rangle).\end{aligned}$$

Taking the expectation on both sides yields

$$\mathbb{E}\lambda_{\max}\left(\frac{1}{N}\Phi^*\Phi\right) \geq 2 - \frac{1}{N} + \frac{(N-1)^2}{N}\frac{1}{\sqrt{4\gamma^2\sigma^2+1}^d} \geq 2 - \frac{1}{N},$$

Similarly, for the smallest eigenvalue, we have

$$\lambda_{\min}\left(\frac{1}{N}\Phi^*\Phi\right) \leq \frac{1}{N}\langle\Phi^*\Phi v, v\rangle \quad \forall v \in \mathbb{C}^N, \|v\|_2 = 1.$$

Since $\{X_\ell\}_{\ell\in[p]}$ are vectors in $\mathbb{C}^p$, we can find a unit vector $u = [u_1, \ldots, u_p]^T \in \mathbb{C}^p$ such that u is

orthogonal to span$\{X_\ell\}_{\ell\in[p-1]}$, and thus

$$\begin{aligned}\lambda_{\min}\left(\frac{1}{p}\Phi^*\Phi\right) &\leq \frac{1}{p}\langle\Phi^*\Phi u, u\rangle \\ &\leq \frac{1}{p}u^* X_p X_p^* u \\ &= \frac{1}{p}\sum_{j,k=1}^{p}\bar{u}_j u_k \exp(i\langle x_p, \omega_k - \omega_j\rangle).\end{aligned}$$

The vector u is linearly independent of $\{X_\ell\}_{\ell\in[p-1]}$, but since the columns $\{X_\ell\}_{\ell\in[p]}$ are linearly independent (Φ is full rank almost surely see Lemma 8.2), u is dependent of the vector X_p we find

$$\begin{aligned}\mathbb{E}\lambda_{\min}\left(\frac{1}{p}\Phi^*\Phi\right) &\leq \frac{1}{p} + \frac{1}{p}\mathbb{E}\sum_{j\neq k}\bar{u}_j u_k \exp\left(i\langle x_p, \omega_k - \omega_j\rangle\right) \\ &= \frac{1}{p} + \frac{1}{p}\mathbb{E}_{\omega_p, x_p}\sum_{j\neq k}\bar{u}_j u_k \mathbb{E}_{\omega_1,\ldots,\omega_{p-1},x_1,\ldots x_{p-1}}\left[\exp\left(i\langle x_p, \omega_k - \omega_j\rangle\right)\right] \\ &= \frac{1}{p} + \frac{1}{p}\mathbb{E}_{\omega_p, x_p}\sum_{j\neq k}\bar{u}_j u_k \mathbb{E}_{\omega_1,\ldots,\omega_{p-1}}\left[\exp\left(i\langle x_p, \omega_k - \omega_j\rangle\right)\right] \\ &= \frac{1}{p} + \frac{1}{p}\mathbb{E}_{\omega_p, x_p}\sum_{j\neq k}\bar{u}_j u_k \exp\left(-\frac{\sigma^2}{2}\|x_p\|^2\right) \\ &\leq \frac{1}{p} + \frac{1}{p}\mathbb{E}_{\omega_p, x_p}\exp\left(-\gamma^2\|x_p\|^2\right) \\ &\leq \frac{1}{p} + \frac{1}{p}\left(\frac{1}{2\gamma^2\sigma^2+1}\right)^{d/2} \\ &\leq \frac{1}{p} + \left(1-\frac{1}{p}\right)^{1/2}\left(\frac{1}{4\gamma^2\sigma^2+1}\right)^{d/4}.\end{aligned}$$

where we used the Cauchy–Schwarz inequality to get to the 5th line and the fact that for $p \geq 2$ we have $1 - \frac{1}{p} \geq \frac{1}{p}$ and $1 - \frac{1}{p} \leq \sqrt{1-\frac{1}{p}}$. This completes the proof.

8.4 Proof of Theorem 5.10

We start by stating the following theorem by (Singh, Lucchi, et al. 2022).

Theorem 8.5 *Under the assumption 5.5 and taking the limit of external regularization $\lambda \to 0$ and $k = 1$, we obtain the following lower bound on the population risk at the minimum $\theta^\star$:*

$$\mathcal{L}(\hat{\theta}) \geq \tilde{\mathcal{L}}_S(\hat{\theta}) + \frac{1}{n+1}\frac{\sigma^2_{min}\alpha\lambda_{min}\left(C_f^{\tilde{D}}(\hat{\theta})\right)}{\lambda_r\left(H_{\mathcal{L}}^S(\hat{\theta})\right)}$$

Here, we consider the convention $\lambda_1 \geq \cdots \geq \lambda_r$ for the eigenvalues, with $r := \operatorname{rank}(H_S^{\mathcal{L}}(\hat{\theta}))$, and σ^2_{min} denotes the minimum $\sigma^2(x,y) = \|\nabla_f \ell(f_{\theta^\star}(x), y)\|^2$ over $\tilde{\mathcal{D}}$, i.e., $(x,y) \sim \mathcal{D} : \sigma^2(x,y) > 0$ (subset of $\mathcal{D}$ where $\sigma^2(x,y) > 0$).

Proof: See the proof in (Singh, Lucchi, et al. 2022).

For the MSE loss ℓ and $k = 1$, we have $\nabla_f^2 \ell = 1$, and by using the assumption 5.6, the expression 24 of

$H_S^{\mathcal{L}}$ for $\hat{\theta}$ becomes:

$$H_{\mathcal{L}}^S(\hat{\theta}) = H_o^S(\hat{\theta}) + H_f^S(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta f_\theta(x_i) \nabla_\theta f_\theta(x_i)^\top$$
$$= C_f^S(\hat{\theta})$$

which is a rank-1 decomposition for the matrix $C_f^S(\hat{\theta})$. Using the assumption 5.7, we can bound bound the minimum eigenvalue λ_r of $C_f^S(\hat{\theta})$ by the one at the initialization resulting in.

$$\mathcal{L}(\hat{\theta}) \geq \tilde{\mathcal{L}}_S(\hat{\theta}) + \frac{1}{n+1} \frac{A\sigma_{\min}^2 \alpha \lambda_{\min}\left(C_f^{\tilde{D}}(\theta^0)\right)}{B\lambda_r\left(C_f^S(\theta^0)\right)}.$$

For the next step, we state the following theorem by (Vershynin 2011).

Theorem 8.6 *((Vershynin 2011) (Theorem 5.58)).*
Let A be an $N \times n$ matrix whose columns are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^N$. Then the following holds as $N, n \to \infty$ but their ratio $\frac{n}{N} \in]0, 1]$.

$$\lambda_{\min}\left(\frac{1}{N}AA^\top\right) \to \left(1 - c\sqrt{\frac{n}{N}}\right)^2 \quad \text{and} \quad \lambda_{\max}\left(\frac{1}{N}AA^\top\right) \to \left(1 + c\sqrt{\frac{n}{N}}\right)^2 \quad a.s.$$

where c depends only on the sub-Gaussian norm of the columns.

The aim is to apply this theorem on the matrix $C_f^S(\theta^0) = \frac{1}{n} Z_S(\theta^0) Z_S(\theta^0)^\top$. However, This result holds for random matrices whose columns are isotropic random vectors. Therefore, we use a **whitening transformation** to turn it into a random matrix with isotropic sub-Gaussian columns by taking our whitening matrix to be $W = \left(C_f^{\mathcal{D}}(\theta^0)\right)^{-\frac{1}{2}}$, which is possible since $C_f^{\mathcal{D}}(\theta^0)$ is symmetric semi-definite and its spectrum is bounded away from zero as a typical assumption in most analysis. Hence, the matrix $\tilde{Z}_S = \left(C_f^{\mathcal{D}}(\theta^0)\right)^{-\frac{1}{2}} Z_S(\theta^0)$ is a random matrix with independent isotropic sub-Gaussian columns. we apply the theorem to $\frac{1}{n}\tilde{Z}\tilde{Z}^\top$ and obtain by setting $A = \frac{1}{\sqrt{n}}(C_f^S(\theta^0))^{\frac{1}{2}}$, $B = \frac{1}{\sqrt{n}}(C_f^S(\theta^0))^{-\frac{1}{2}} Z_S(\theta^0)$ and using the fact that $s_{\min}(AB) \leq \|A\|_2 s_{\min}(B)$ where s is the singular value.

$$\begin{aligned}
\lambda_r(ABB^\top A^\top) &= \lambda_r\left(\frac{1}{n} C_f^S(\theta^0)^{1/2} C_f^S(\theta^0)^{-1/2} Z_S(\theta^0) Z_S(\theta^0)^\top (C_f^S(\theta^0)^{-1/2})^\top (C_f^S(\theta^0))^{1/2})^\top\right) \\
&= \lambda_r(C_f^S(\theta^0)) = s_{\min}^2(AB) = s_{\min}^2\left(\frac{1}{n}(C_f^S(\theta^0))^{1/2}(C_f^S(\theta^0))^{-1/2} Z_S(\theta^0)\right) \\
&\leq \|(C_f^S(\theta^0))\|_2 s_r^2\left(\frac{1}{n}(C_f^S(\theta^0))^{-1/2} Z_S\right) \\
&\leq \|(C_f^S(\theta^0))\|_2 \lambda_r\left(\frac{1}{n}\tilde{Z}\tilde{Z}^\top\right) \\
&\leq \|(C_f^S(\theta^0))\|_2 \left(1 - c\sqrt{\frac{p}{n}}\right)^2
\end{aligned}$$

Hence, the final result

$$\mathcal{L}(\hat{\theta}) \geq \tilde{\mathcal{L}}_S(\hat{\theta}) + \frac{1}{n+1} \frac{A\sigma_{\min}^2 \alpha \lambda_{\min}\left(C_f^{\tilde{D}}(\theta^0)\right)}{B\|C_f^S(\theta^0)\|_2 \left(1 - c\sqrt{\frac{p}{n}}\right)^2}.$$

8.5 Proof of Theorem 6.1

This proof is taken from (Heckel and Yilmaz 2020).

We start by decomposing the difference of the two risks into

$$\left|R(\theta^t - \bar{R}(\tilde{\theta}))\right| \leq \left|R(\theta^t) - \bar{R}(\tilde{\theta})\right| + \left|R(\theta^t - \bar{R}(\tilde{\theta}))\right|. \tag{34}$$

and bound the two terms separately.

The first term can be bounded using the following lemma.

Lemma 8.7 *(Heckel and Yilmaz 2020) Define $\tilde{X}$ so that $X = \tilde{X}\Sigma^{1/2}$. Suppose that*

$$\|I - \tilde{X}^T\tilde{X}\| \leq \kappa, \quad \text{with } \kappa \leq \frac{\min_i \eta_i\sigma_i^2}{2\max_i \eta_i\sigma_i^2}.$$

Then,

$$\|R(\theta^t) - R(\tilde{\theta}^t)\| \leq (1 - (1 - \min_i \frac{\eta_i\sigma_i^2}{2})^t)^2 \frac{\max_i \eta_i^2\sigma_i^4}{\min_i \eta_i^2\sigma_i^4} 8\kappa^2 \max_{\ell\in\{1,\dots,t\}} \|\Sigma\tilde{\theta}_\ell - \Sigma\theta^*\|_2^2.$$

To apply this lemma, we first need to verify its condition. Since the columns of our design matrix X are drawn from $\mathcal{N}(0,\Sigma)$, the columns of the matrix $\tilde{X}$ given by $X = \tilde{X}\Sigma^{1/2}$ are iid $\mathcal{N}(0, \frac{1}{\sqrt{n}}I_d)$. Using the following concentration inequality from (Foucart and Rauhut 2013), for any $\beta \in]0,1[$,

$$\mathbb{P}\left[\left\|I - \tilde{X}^\top\tilde{X} \geq \beta\right\|\right] \leq e^{-\frac{n\beta^2}{15}+4d}$$

Taking $\beta = \sqrt{\frac{75d}{n}}$ we get that, with probability at least $1 - e^{-d}$,

$$\left\|I - \tilde{X}^\top\tilde{X}\right\| \leq \sqrt{75\frac{d}{n}}$$

Lemma 8.8 *Provided that $\eta_i\sigma_i^2 \leq 1$ for all i, with probability at least $1 - 2d(e^{-\beta^2/2} + e^{-n/8})$, we have:*

$$\max_\ell \left\|\Sigma\tilde{\theta}^\ell - \Sigma\theta^*\right\|_2^2 \leq 2\|\Sigma\theta^*\|_2^2 + \frac{4d}{n}\sigma^2\beta^2.$$

Applying the lemma with $\beta^2 = 10\log(d)$, we obtain that with probability at least $1 - 2d^{-5} - 2de^{-n/8} - e^{-d}$:

$$\left|R(\theta^t) - R(\tilde{\theta}^t)\right| \leq 8\frac{\max_i \eta_i^2\sigma_i^4}{\min_i \eta_i^2\sigma_i^4} \cdot \left(2\|\Sigma\theta^*\|_2^2 + \frac{4d}{n}\sigma^2 \cdot 10\log(2d)\right).$$

Lemma 8.9 *With probability at least $1 - 4e^{-\beta^2/8}$, we have:*

$$\left|R(\tilde{\theta}_t) - \overline{R}(\tilde{\theta}_t)\right| \leq \frac{\sigma^2}{n}3\beta\sqrt{d}.$$

with $\bar{R}(\tilde{\theta}^t)$ as defined in 27.

The proofs of the three lemmas used in this proof can be found in (Heckel and Yilmaz 2020).