DECLARATION OF AUTHORSHIP

FHS MATHEMATICS AND STATISTICS PART C

Candidate number:	1015294
Title of Dissertation (in capitals): **Double Descent Curve in Linear Regression**	
Word count: _____**7905**____	

Please tick to confirm the following:

I have read and understood the University's disciplinary regulations concerning conduct in examinations and, in particular, the regulations on plagiarism (*The University Student Handbook* Section 7.8; available at https://www.ox.ac.uk/students/academic/student-handbook).	✓
I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1.	✓
The dissertation I am submitting is entirely my own work except where otherwise indicated.	✓
It has not been submitted, either partially or in full, either for this Honour School or qualification or for another Honour School or qualification of this University (except where the Special Regulations for the subject permit this), or for a qualification at any other institution.	✓
I have clearly indicated the presence of all material I have quoted from other sources, including any diagrams, charts, tables or graphs.	✓
I have clearly indicated the presence of all paraphrased material with appropriate references.	✓
I have acknowledged appropriately any assistance I have received in addition to that provided by my supervisor.	✓
I have not copied from the work of any other candidate.	✓
I have not used the services of any agency providing specimen, model or ghostwritten work in the preparation of this thesis/dissertation/extended essay/assignment/project/other submitted work. (See also section 2.4 of Statute XI on University Discipline under which members of the University are prohibited from providing material of this nature for candidates in examinations at this University or elsewhere: http://www.admin.ox.ac.uk/statutes/352-051a.shtml).	✓
I agree to retain an electronic version of the work until the publication of my final examination result.	✓

I agree that this electronic copy may be used should it be necessary to confirm my word count or to check for plagiarism.	✓

Double Descent Curve in Linear Regression

Part C Mathematics and Statistics TT 2020

Candidate Number 1015294

Abstract

We derive an explicit formula for the limit of the generalization risk for the minimum l_2-norm least squares estimator in linear regression. This formula has double descent shape which is a new way to qualitatively describe the generalization risk. The asymptotic setting we consider is when the number of observations n and the number of parameters p go to infinity such that $p/n \to \gamma$, where γ is positive and finite. The main result of the dissertation is to show that the asymptotic risk depends only on the signal strength r^2, the noise σ^2 and the aspect ratio γ. It is also shown that the risk shoots up when p is close to n, but decreases as p increases beyond n.

The statements of all results in this dissertation have appeared in the literature in an identical, equivalent, or closely related form. References have been appropriately provided. The selection and choice of presentation of material are my own work. Also the proofs of almost all **Propositions** are entirely based on my own ideas, where they are not - proper references are given. None of the **Theorems** and their proofs provided in the dissertation are novel, but some of the proofs have been written much more comprehensively and this is essentially product of my own contribution.

Section 1 is based on [8]. Section 3 briefly describes crucial theoretical results from random matrix theory and complex analysis and discusses how we shall need this theory in the subsequent sections. Sections 2, 4, 5 and 8 are based almost entirely on [1] but filling in and explaining some missing significant details and especially missing calculations are done by my own efforts. In essence, all computations from these four sections which are in the dissertation but not in [1] are mine unless otherwise stated. Section 6 is based on [16].
The word count **estimate** given below was found by using version 3.1.1 of the online counter at `http://app.uio.no/ifi/texcount/online.php`. Sections 1 - 8 were included in the count. The Appendix was counted as advised by the Dissertation Guideline provided by the Mathematical Institute. It is possible that the counter provides an overestimate by including certain LaTeX commands and formulae.
Word Count: 7905

Contents

1 Introduction

In this dissertation we study the generalization risk in ridgeless and ridge regression under different models for the features $x_i \in \mathbb{R}^p$. The regression problem we consider has a vector-matrix form $y = X\beta + \epsilon$, where we have response variable $y \in \mathbb{R}^n$, feature matrix $X \in \mathbb{R}^{n \times p}$, parameter $\beta \in \mathbb{R}^p$ and noise term $\epsilon \in \mathbb{R}^n$. For the least squares estimator $\hat{\beta} \in \mathbb{R}^p$ we examine the shape of its generalization risk $\mathbb{E}((x_0^T\hat{\beta} - x_0^T\beta)^2|X)$, where $x_0 \in \mathbb{R}^p$ is a test sample unseen in training. The main focus of the dissertation is to derive an explicit expression for the risk in high-dimensional asymptotic regime where the number of observations n and the number of parameters p go to infinity such that $p/n \to \gamma \in (0, \infty)$. Most of the material presented in the dissertation is following [1]. The aim is to gain a better understanding about the mathematics behind the derivation of the risk function formula for the curve. Many missing calculations in [1] and some further insights for the curve are examined here.

In this section we present the stimulus behind the conducted research in [8]. We describe the double descent phenomenon [8] for large models where the best generalization performance is obtained in overparametrized models. In [8] Belkin, Hsu, Ma and Mandal propose this new double descent style of learning which gives possible explanation for why very large models such as neural networks work so well in practice.

1.1 Classical U-curve

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" [2]. This is what the great physicist Enrico Fermi told Freeman Dyson when Freeman wanted Fermi's blessing on his findings in the high-energy theory of pseudoscalar mesons. Whether back in the days they knew about overfitting or not, it is clear that the idea of having a large number of parameters will interpolate data is highly intuitive. Avoiding overfitting is a must in classical statistics and often goes with the quote *"Memorizing is not learning"*. The main aim in machine learning is to minimize the *generalization error*, i.e the error on previously unseen observations. It can be decomposed into two error terms called *bias* and *variance*. The conventional approach to this optimization problem is to search for this *sweet spot* in model complexity that balances the *bias* and *variance* errors, Figure 1.

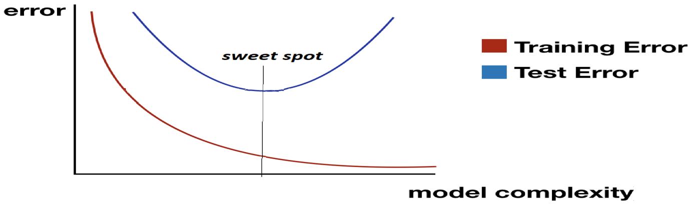

Figure 1: The classical U-curve. Model complexity, equivalently model capacity, is the number of parameters in the model.

Balancing between *bias* and *variance*, also known as *bias-variance trade-off*, is the same as balancing between underfitting and overfitting. Indeed, the *variance* is typically increasing as the number of parameters increases and is overfitting data points. On the other hand, the bias term is increasing when the model complexity is small and represents the error from underfitting data points.

1.2 Double descent curve

The classical statistical perspective described above contradicts with the current approach in real-life deep learning models. The high complexity of modern neural networks with many parameters does not seem to overfit data and even sometimes provides better generalization performance. For instance, GoogLeNet has 6.7 million parameters [5], AlexNet has 60 million parameters [6] and both neural networks are very successful. The classical statistical approach was challenged with more empirical evidence in [7], where the authors give example of convolutional neural networks which easily interpolate random labelling of the training data, i.e pure noise. Conventional statistics suggests that learning completely randomised data should be impossible or at least reduce the convergence rate of the training procedure significantly but in fact the learning method appears to be unaffected by this transformation to pure noise. This conceptual challenge is heuristically explained by the *double descent curve* proposed by Belkin, Hsu, Ma and Mandal [8], Figure 2.

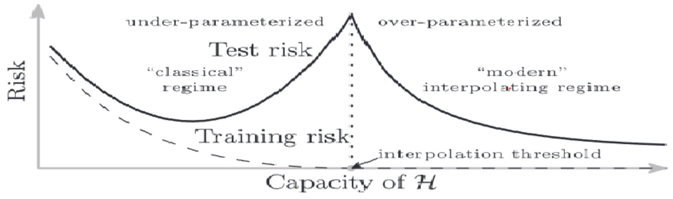

Figure 2: Taken from [8]. The double descent curve for the generalization error.

In this paper the authors describe several large deep learning models which exhibit double descent loss. One of the models in interest is the Random Fourier Features, which is equivalent to a two-layer neural network with constant weights in the first layer. The hypothesis space $\mathcal{H}_\mathcal{N}$ of the RFF family of functions consists of maps $f : \mathbb{R}^d \to \mathbb{C}$ such that:

$$f(x) = \sum_{n=1}^{\mathcal{N}} a_n \sigma(x; v_n) \quad \text{with Fourier features} \quad \sigma(x, v_n) = \cos(x^T v_n) + i \times sin(x^T v_n)$$

where $v_1, \ldots, v_n$ are i.i.d with normal distribution $N(0, I)$. The authors observe the double descent phenomenon when they fit RFF model on the famous MNIST dataset. The model capacity of $\mathcal{H}_\mathcal{N}$ is determined by the size of $\mathcal{N}$ which controls the number of parameters of $f(x)$. Interestingly, the classical U-curve is observed as the number of parameters $\mathcal{N}$ is less than the number of observations and the peak of the curve, also called interpolation threshold, is reached when $\mathcal{N}$ is equal to the number of observations. After this peak the model achieves nearly zero training loss and by conventional statistics is expected that will generalize poorly. However, the opposite scenario is observed. The accuracy of the model is improved as the number of parameters is increased and even shows better performance from the predictor corresponding to the bottom of the U-shaped curve. This unusual phenomenon demonstrates a counter-intuitive result that larger models can generalize better than smaller ones. A heuristic explanation why this is true is that for the class of functions in $\mathcal{H}_\mathcal{N}$ with fixed $\mathcal{N}$ we cannot be sure that the best solution in terms of both lowest training error and lowest norm is contained in $\mathcal{H}_\mathcal{N}$. But as we increase $\mathcal{N}$ we allow a larger family of solutions $\mathcal{H}_\mathcal{N}$ to be considered which can result into a solution with smaller norm, while we can still achieve small training error (near zero) by being after the interpolation threshold. The key idea in the paper [8] is that allowing a model to have higher number of parameters does not necessarily mean having more complicated model but rather it can lead to choosing small norm interpolating predictor with smooth structure. In Figure 3 below we can see that the larger model with 100 times more parameters has better generalizing performance. Although, it has more wiggling movements it looks simpler than the model with the smaller number of parameters. That is because the larger learner has smaller norm. With this example Belkin, Hsu, Ma and Mandal show that interpolation does not contradict generalization as previously thought by classical statistics.

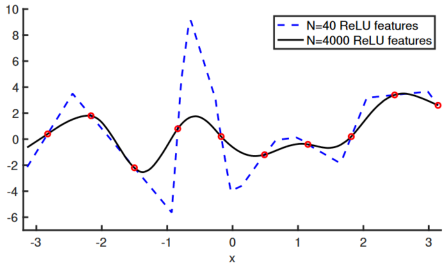

Figure 3. Taken from [8]: Two univariate models are fitted on 10 data points (red circles) using Random ReLU feature map $\phi(x; (w,b)) = \max(w^T x + b, 0)$. The blue dashed line is the model with 40 features and the black solid line is the model with 4000 features.

To sum up, the key statements in [8] are the following:

1. When the number of parameters equals the number of observations the training risk is almost zero but the test risk shoots up. We call this the interpolation threshold.

2. Below the interpolation threshold we observe the classical U-shaped curve.

3. After the interpolation threshold we observe a zero training loss and a monotonically decreasing test risk.

2 Ridgeless Regression

In this section we introduce the simplest model - linear ridgeless regression, as in [1]. We will derive the min-norm least squares estimator, also known as the solution with smallest l_2-norm of $\arg\min_\beta ||y - X\beta||_2^2$. The reason why such a simple model as linear regression is considered is because the risk function is tractable and there exist some partial parallels between linear models and shallow neural networks. We describe these parallels in the next subsection where we describe the setup in detail.

2.1 Setup

Suppose we have n i.i.d. training samples $(x_i, y_i) \in \mathbb{R}^{p\times 1} \times \mathbb{R}, i = 1, 2, \ldots, n$, where each sample is generated independently from the following model ***(1)***:

- Draw $x_i \sim F_x$ and $\epsilon_i \sim F_\epsilon$ independently from some probability distributions F_x and F_ϵ.
- Assume ϵ_i is independent homoscedastic noise such that $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$.
- Assume feature vectors x_i with mean 0, covariance matrix $\text{Cov}(x_i) = \Sigma$ and can be represented as $x_i = \Sigma^{1/2} z_i$, where z_i is a random vector with i.i.d. entries with mean 0 and variance 1.
- Obtain response $y_i = x_i^T \beta + \epsilon_i$ for some $\beta \in \mathbb{R}^p$ which we wish to learn. In vector-matrix notation we have $y = X\beta + \epsilon$.

Note that in the third assumption above we applied a linear transform to the feature vector z_i and derived a new feature vector x_i. We did this because it is related to the setting in shallow neural networks. There the aim is to learn a nonlinear function $f(z_i;\theta)$ which associates output variables y_i and input variables z_i through the parameter θ. Assume we initialize the parameter variable with θ_0. After Taylor expansion for $f(z_i;\theta)$ around θ_0 we obtain:

$$f(z_i;\theta) = f(z_i;\theta_0) + \bigtriangledown f(z_i;\theta_0)^T(\theta - \theta_0) + o((\theta - \theta_0))$$

In neural networks we usually have so many parameters and the loss curve is so complicated that when we train the model we do not move too far away from the choice of initialization θ_0. Therefore we can ignore the terms of order $o((\theta - \theta_0))$. Further if we assume that we chose the initialization such that $f(z_i;\theta_0) = 0$ and set $\beta = \theta - \theta_0$, then we have:

$$f(z_i;\theta) \approx \bigtriangledown f(z_i;\theta_0)^T \beta$$

The last equation says that the target function $f(z_i;\theta)$ we want to learn is linear in the transformed features $x_i = \bigtriangledown f(z_i;\theta_0)$. In our case we consider a linear transformation $x_i = \Sigma^{1/2} z_i$ which is as if we have a one layer neural network without an activation function. The above setup also reminds us of the Neural Tangent Kernel Method where we do gradient descent in the space of functions f.

Although working with a linear model provides a complete expression for the risk function, a linear setting does not capture all properties in neural networks. In linear regression we do not *learn* the values of the parameters, we just *observe* them. On the other hand, in neural networks we progressively update our parameters and the regression function as we optimize the loss function. Another concept which linear regression fails to capture about neural networks is the distinction between the dimensions of the input/feature space $x \in \mathbb{R}^p$ and the parameter space β (or θ) $\in \mathbb{R}^p$; they are both equal to p. We call this situation **ambiguity of** p. In neural networks the number of parameters is different from the number of inputs. When we create deeper network we increase the dimension of the parameter space, but the dimension of the input space is not changed. This **ambiguity of** p will lead to some peculiar results for the risk in linear regression described in Section 5.

2.2 Prediction risk

Consider model *(1)* and assume we have an estimator $\hat{\beta}$ depending on the training data X and y. Assume also we have drawn unseen test sample (x_0, y_0) identically and independently to the training data in model *(1)*. In Section 1 we said that we are interested in the generalization risk $\mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X]$. However, the goal of machine learning is to minimize the expected difference between the predicted response $\hat{y}_0 = x_0^T\hat{\beta}$ and the true response y_0: $\mathbb{E}[(\hat{y}_0 - y_0)^2|X]$. In this subsection we show that considering either of the two quantities would not change the shape of the risk curve. Note that both of the above **conditional** expectations are taken with respect to all random quantities in the expression - the training response variables y and the fresh sample (x_0, y_0) .

Proposition 2.1. *Assume we have model (1). Then the generalization risk satisfies:* $\mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X] = \mathbb{E}[(\hat{y}_0 - y_0)^2|X] - \sigma^2$.

Proof. The goal of this proposition is to "extract" the noise ϵ_0, coming from the fresh sample (x_0, y_0), out of the risk function. We do this by expanding the expression inside the expectation.

$$\begin{aligned}
\mathbb{E}[(\hat{y}_0 - y_0)^2|X] &= \mathbb{E}[(x_0^T\hat{\beta} - (x_0^T\beta + \epsilon_0))^2|X] \\
&= \mathbb{E}[((x_0^T\hat{\beta} - x_0^T\beta) - \epsilon_0)^2|X] \\
&= \mathbb{E}\Big[(x_0^T\hat{\beta} - x_0^T\beta)^2|X\Big] - 2\mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)\epsilon_0|X] + \mathbb{E}[\epsilon_0^2|X] \\
&= \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X] - 2\mathbb{E}\Big[\mathbb{E}\big\{(x_0^T\hat{\beta} - x_0^T\beta)\epsilon_0|X, x_0, y\big\}\Big] + \sigma^2 \\
&= \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X] - 2\mathbb{E}\Big[(x_0^T\hat{\beta} - x_0^T\beta)\mathbb{E}\big\{\epsilon_0|x_0, y\big\}|X\Big] + \sigma^2 \\
&= \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X] + \sigma^2
\end{aligned}$$

where in the fourth equation we applied the Law of Total Probability on the cross term by additionally conditioning on x_0 and y so that we can take $(x_0^T\hat{\beta} - x_0^T\beta)$ out of the inner expectation (note $\hat{\beta}$ depends on X and y). In the last equation we used that the error ϵ_0 is independent from x_0 and y and has mean 0. □

One of the assumptions in model *(1)* is that the error terms have constant variance σ^2. Hence, $\mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X]$ is the same as $\mathbb{E}[(y_0 - \hat{y}_0)^2|X]$ just shifted by a constant term σ^2 which does not change the shape of the curve. This leads to the following definition of the risk:

Definition 2.2. *Assume model (1). Define the generalization risk $R_X(\hat{\beta}; \beta)$ for an estimator $\hat{\beta}$ and unseen data sample (x_0, y_0) in the following way:*

$$R_X(\hat{\beta}; \beta) = \mathbb{E}[(x_0^T\hat{\beta} - x_0^T\beta)^2|X]$$

For the rest of the dissertation when we use a term which is a combination of two words w_1 and w_2 separated by a space symbol, where $w_1 \in$ {generalization, test, prediction, out-of-sample} and $w_2 \in$ {error, risk} we shall refer to the risk in Definition 2.2. Note that the generalization risk depends only on the observed features X as we take the expectation conditional on X, not on the training responses y. The reason for that is because the response variables y contain noise ϵ and we want to treat this noise as random quantity rather than a given quantity. We can further modify the prediction risk so that later we derive its *bias* $-$ *variance* decomposition.

Proposition 2.3. *Assume we have model (1). Then the test error satisfies:* $R_X(\hat{\beta};\beta) = \mathbb{E}[||\hat{\beta}-\beta||_\Sigma^2|X]$*, where we define the* Σ*-norm* $||x||_\Sigma^2 = x^T\Sigma x$*.*

Proof.

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

In the fourth equation we used the Law of Total Probability for expectations. Note we additionally conditioned on y because the estimator $\hat{\beta}$ depends on the training data X and y. □

Note that if $\Sigma = I$, then we have the usual l_2-norm. By Proposition 2.3 the test error is just the expected Euclidean distance between the parameter's estimator and the parameter itself. Next we derive the *bias-variance trade-off* decomposition of the generalization error.

Proposition 2.4. *Assume we have model (1). The bias-variance decomposition of the test error for an arbitrary estimator* $\hat{\beta}$ *is:*

$$R_X(\hat{\beta};\beta) = B_X(\hat{\beta};\beta) + V_X(\hat{\beta};\beta) \tag{2.1}$$

$$B_X(\hat{\beta};\beta) = ||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2 \;\; \textit{and} \;\; V_X(\hat{\beta};\beta) = \textit{trace}(\textit{Cov}(\hat{\beta}|X)\Sigma) \tag{2.2}$$

Proof. The proof of this proposition is given in Appendix 8.1. □

2.3 Minimum norm Least Squares Estimator

In the previous subsection we defined the out-of-sample error for a general estimator $\hat{\beta}$. Here we will consider an exact expression for the estimator; it will be the solution with smallest norm of the least squares regression problem. The next proposition determines this solution.

Proposition 2.5. *The classical least squares problem without a regularization term* $\arg\min_\beta\{||y - X\beta||_2^2\}$ *is minimized by* $\hat{\beta} = (X^TX)^\dagger X^Ty$*, called the least squares estimator, where* $(X^TX)^\dagger$ *is the Moore-Penrose pseudoinverse of* X^TX*. Note* $\arg\min_\beta\{||y - X\beta||_2^2\}$ *need not have a unique minimizer. However, the proposed* $\hat{\beta}$ *is the solution with the smallest* l_2*-norm among all minimizers and that is why we call it* ***the minimum norm least squares estimator****.*

Proof. The proof of this proposition is given in Appendix 8.2. □

Definition 2.6. *For the rest of the dissertation we shall use interchangeably the terms minimum l_2-norm least squares estimator and least squares estimator to refer to $\hat{\beta} = (X^TX)^\dagger X^Ty$. Also we will simply write $\hat{\beta}$ to indicate this estimator.*

We proceed with deriving the bias-variance decomposition of the prediction error of $\hat{\beta} = (X^TX)^\dagger X^Ty$.

Definition 2.7. *Assume model (1). Let $\hat{\Sigma} = n^{-1}X^TX$ be the sample covariance matrix. Let also $\Pi = I - \hat{\Sigma}^\dagger\hat{\Sigma}$.*

Proposition 2.8. *Consider model (1). The bias-variance decomposition of the test error for $\hat{\beta} = (X^TX)^\dagger X^Ty$ is given by:*

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

$$B_X(\hat{\beta};\beta) = \beta^T\Pi\Sigma\Pi\beta \quad and \quad V_X(\hat{\beta};\beta) = \frac{\sigma^2}{n} trace(\hat{\Sigma}^\dagger\Sigma)$$

Proof. The proof of this proposition is given in Appendix 8.3 □

3 Theory

In this section we provide important results from random matrix theory and complex analysis which we shall need in the derivation of the asymptotic prediction risk. A random matrix is a matrix with random entries. For example, the feature matrix X, on which the generalization error depends, is random. Random matrix theory is a very useful tool to characterize the probabilistic behaviour of the eigenvalues of high-dimensional random symmetric matrices such as the sample covariance matrix $\hat{\Sigma} \in \mathbb{R}^{p\times p}$. In particular we are interested in the empirical spectral distribution of the eigenvalues of $\hat{\Sigma}$. An interesting question is to analyze the limiting performance of extreme eigenvalues as the dimension of the matrix becomes large. And of course we cannot go without calculus. We introduce a few results from complex analysis in order to be able to compute expectations (integrals) with respect to more sophisticated probability distributions.

3.1 Random Matrix Theory

Definition 3.1. *The **empirical spectral distribution** $F_{\hat{\Sigma}}(x)$ for a symmetric matrix $\hat{\Sigma} \in \mathbb{R}^{p\times p}$ is defined as follows:*

$$F_{\hat{\Sigma}}(x) = \frac{1}{p}\sum_{i=1}^{p} I\{\lambda_i(\hat{\Sigma}) \leq x\} \qquad for\ x \in [0,\infty)$$

where $\lambda_i(\hat{\Sigma})$ are the eigenvalues of $\hat{\Sigma}$ and I is the indicator function.

Explicit computation of the eigenvalues for large dimensional random matrices is often not possible. Thus, we would want to figure out their asymptotic properties. In order to do that we need to introduce some basic convergence tools in probability theory.

Definition 3.2. *Let $S, S_1, S_2, \dots, S_n$ be a sequence of random variables. Then we say $S_n \to S$ almost surely as $n \to \infty$ if:*

$$\mathbb{P}(S_n \to S \text{ as } n \to \infty) = 1$$

Definition 3.3. *Let $S, S_1, S_2, \dots, S_n$ be a sequence of random variables with corresponding distributions $F, F_1, F_2, \dots, F_n$. Then we say $S_n \to S$ weakly, (or in distribution) as $n \to \infty$ if for every x such that F is continuous at x:*

$$F_n(x) \to F(x) \text{ as } n \to \infty$$

Remark 3.4. *Almost sure convergence implies convergence in distribution.*

The next theorem is another useful convergence theorem in probability. It relates convergence in distribution of random variables with convergence of their expectations. This will basically give us convergence theorems for integrals.

Theorem 3.5. *(Portmanteau theorem [13])The following two convergence of measures statements are equivalent:*
1. S_n converges weakly to S as $n \to \infty$.
2. For all bounded continuous functions f that are non-zero only on a compact set we have: $\mathbb{E}_{S_n}(f) \to \mathbb{E}_S(f)$ as $n \to \infty$.

As we mentioned in the beginning we are interested in examining the limiting behaviour of eigenvalues of large matrices. Then it will be helpful to consider the limit of the empirical spectral distribution.

Definition 3.6. *The limit $F(x)$ of the empirical spectral distribution $F_{\hat{\Sigma}}(x)$ as $p \to \infty$ is called **limiting spectral distribution**. If $\hat{\Sigma}$ is a random quantity then we take the limit almost surely.*

Next we state a famous result, given by Bai-Yin [12], about the extreme eigenvalues of a particular sample covariance matrix $\hat{\Sigma}$.

Theorem 3.7. *(Bai-Yin [12]) Let Z be a $n \times p$ random matrix whose entries are i.i.d. random variables from a probability distribution P_z which has mean 0, variance 1 and a finite fourth moment. Let the sample covariance matrix be $\hat{\Sigma} = n^{-1}Z^TZ$. Assume that $n, p \to \infty$ in a proportional regime such that $p/n \to \gamma$, where $\gamma \in (0, 1)$. Then the smallest eigenvalue $\lambda_{min}(\hat{\Sigma})$ and the largest $\lambda_{max}(\hat{\Sigma})$ for $\hat{\Sigma}$ satisfy:*

$$\lim \lambda_{min}(\hat{\Sigma}) = (1 - \sqrt{\gamma})^2 \qquad \textit{almost surely}$$
$$\lim \lambda_{max}(\hat{\Sigma}) = (1 + \sqrt{\gamma})^2 \qquad \textit{almost surely}$$

The next theorem we describe is a well known result for the limiting spectral distribution $F_{\hat{\Sigma}}$ and is an important fact which we shall need later when we compute the explicit form of the asymptotic prediction risk.

Theorem 3.8. *(Marčenko-Pastur [9]) Let Z be a $n \times p$ random matrix whose entries are i.i.d. random variables from a probability distribution P_z which has mean 0 and finite variance $\sigma_z^2 < \infty$. We define the sample covariance to be $\hat{\Sigma} = n^{-1}Z^TZ$. Assume* *that $n, p \to \infty$ in a proportional regime such that $p/n \to \gamma$, where $\gamma \in (0,1)$. Then $F_{\hat{\Sigma}}$ converges to the Marčenko-Pastur law F_γ which has a probability density function:*

$$f_\gamma(x) = \begin{cases} \dfrac{1}{2\pi\sigma_z^2} \dfrac{\sqrt{(x-a)(b-x)}}{\gamma x} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

where $a = \sigma_z^2(1-\sqrt{\gamma})^2$ and $b = \sigma_z^2(1+\sqrt{\gamma})^2$. Further if γ is greater than 1 and finite, then the limiting distribution is the same as above but it has an additional point mass $\frac{\gamma-1}{\gamma}$ at $x = 0$.

Note that empirical spectral distribution $F_{\hat{\Sigma}}$ depends on the random matrix $\hat{\Sigma}$ whereas the limiting distribution F_γ is a deterministic quantity depending on γ. We will see later that $\gamma < 1$ and $\gamma > 1$ correspond to different learning regimes, in particular, the generalization risk will depend heavily on γ.

3.2 Complex Analysis

Now suppose we are interested in calculating the expectation of the inverse S^{-1} of a certain random variable S which has the spectral empirical distribution $F_{\hat{\Sigma}}$. This computation might not be feasible because we would be integrating over $F_{\hat{\Sigma}}$, where $\hat{\Sigma}$ is random, and thus, we decide to compute the expectation under its limiting spectral distribution F_γ (the Marčenko-Pastur law). Under certain conditions Theorem 3.5 will guarantee us that the expectation with respect to the empirical spectral distribution $F_{\hat{\Sigma}}$ will converge to the expectation with respect to the limiting spectral distribution F_γ. Now the calculation of the latter expectation (essentially an integral) will be possible with the use of Stieltjes transformation $m_\gamma(z)$ of the limiting spectral distribution F_γ:

Definition 3.9. *The Stieltjes transform $m_\gamma(z)$, with complex variable z, of a probability density function $dF_\gamma(x)$ with support $(0, \infty)$ has the following form:*

$$m_\gamma(z) = \int_0^\infty \frac{dF_\gamma(x)}{x - z}, \qquad z \in \mathbb{C} \setminus (0, \infty)$$

The idea is to use this transformation and turn our integral from one defined on the real line into an integral over the complex plane. Having an extra dimension will really make a difference when we need to integrate. Rather than just integrating over an interval in $\mathbb{R}$ we will consider integrating over a *path/contour* in $\mathbb{C}$. We will address one of the main consequences of a major theorem in complex analysis: Cauchy's Theorem [14], which informally says that the integral around a simple closed curve, of a function which is holomorphic in and on that curve, is 0. Further, if we deform the path (but still remain closed) to include a region where the integrand is still holomorphic, then we do not change the integral. As a result, the only contributing points to the integral's value come from the integrand's singularities. The contributions from these singularity points are also known as *residues*. The next proposition shows how to easily compute a residue at a simple pole for a particular type of complex function $f(z)$.

Proposition 3.10. *(Residues at Simple Poles). Suppose $f(z)$ has a simple pole a and $f(z) = g(z)/h(z)$, where g and h are holomorphic at a and h has a simple zero at a, then the residue of $f(z)$ at a is:*

$$res(f(z); a) = \frac{g(a)}{h'(a)}$$

Remark 3.11. *The family of functions $f(z)$ considered above is quite general and often appears in complex analysis problems. Here we are interested in computing the integral in Definition 3.9, which one can see is also from the same family.*

The next result is an important corollary of Cauchy's Theorem and relates the integral of a complex-valued function on a closed contour C with its residue values. This will be our main instrument from complex analysis to compute the integral in Definition 3.9.

Theorem 3.12. *(Cauchy's Residue Theorem [14]) Let f be a complex valued function which is holomorphic inside and on a unit circle C except at points $a_1, \dots, a_n$ inside C. Then:*

$$\int_C f(z)dz = 2\pi i \sum_{i=1}^{n} res(f(z); a_i)$$

Remark 3.13. *Note that in Cauchy's Residue Theorem we are integrating over a path in $\mathbb{C}$, whereas the integral in Definition 3.9 is over an interval on the real line $\mathbb{R}$. However, we can reparametrize to modify the integration to be over a contour in $\mathbb{C}$. We will see a specific example of this in the proof of Theorem 4.1.*

4 Underparametrized regime

We are interested in the asymptotic risk when $n, p \to \infty$ such that $p/n \to \gamma$ where γ is positive and finite. We shall consider two different regimes of "learning": underparametrized when $\gamma < 1$ and overparametrized when $\gamma > 1$. The interpolation threshold is when $\gamma = 1$. In this section we will derive the asymptotic prediction risk in the underparametrized regime. This result has been proven in [10] with the use of random matrix theory. Here we shall examine carefully the proof provided in [1]. We will fill in some important details which are omitted in the paper and will clearly indicate where each of the assumptions of the theorem are used.

Theorem 4.1. *Consider model **(1)** with the following assumptions:*

- *$x \sim F_x$ and $x = \Sigma^{1/2} z$ for a random vector z with i.i.d., isotropic entries and a finite fourth moment.* ***(i)***

- *Σ is a positive definite matrix and its smallest eigenvalue $\lambda_{min}(\Sigma)$ is greater than some positive constant c.* ***(ii)***

- *n and $p \to \infty$ such that $p/n \to \gamma < 1$.* ***(iii)***

Then the prediction risk for $\hat{\beta} = (X^TX)^{\dagger}X^Ty$ satisfies:

$$R_X(\hat{\beta};\beta) \to \sigma^2 \frac{\gamma}{1-\gamma} \qquad \textit{almost surely}$$

Proof. We will divide the proof in several smaller steps to make it more understandable for the reader.

Step 1. In this step we analyze the asymptotic behaviour of the bias term $B_X(\hat{\beta};\beta) \overset{\text{Prop.2.8}}{=} \beta^T\Pi\Sigma\Pi\beta$. Note that in the underparametrized regime we have that $p < n$ and if the columns of X are linearly independent then we have that X^TX is invertible. Hence, the least squares estimator is unbiased because $\mathbb{E}((X^TX)^{-1}X^Ty|X) = (X^TX)^{-1}X^TX\beta = \beta$. This motivates us to seek to prove that $\hat{\beta} = (X^TX)^{\dagger}X^Ty$ is almost surely unbiased as we are looking for a probabilistic statement of the asymptotic bias.
We will prove that $\hat{\Sigma}^{\dagger} = n^{-1}X^TX$ is almost surely invertible from where we will get that $\Pi = I - \hat{\Sigma}^{\dagger}\hat{\Sigma}$ will be almost surely 0 and the result follows.
Note that $n^{-1}X^TX$ is a symmetric matrix and if we show that its smallest eigenvalue is almost surely strictly positive, then $n^{-1}X^TX$ will be almost surely positive definite and hence, almost surely invertible. Denote the smallest eigenvalue value of $n^{-1}X^TX$ with $\lambda_{min}((n^{-1}X^TX))$. From ***(i)*** we can express $X = Z\Sigma^{1/2}$ where Z is the matrix with rows the vectors z_i^T with i.i.d isotropic entries. Now we will prove that $\lambda_{min}(n^{-1}X^TX) \geq \lambda_{min}(n^{-1}Z^TZ)\lambda_{min}(\Sigma)$. We will use the following well known fact from Linear Algebra for the minimum eigenvalue of a positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$:

$$\lambda_{min}(\Sigma) = \min_{v\in\mathbb{R}^p\setminus\{0\}} \frac{<\Sigma v, v>}{<v,v>} \tag{4.1}$$

where we use the traditional vector inner product $<v,u> = v^Tu$. This is a result which follows directly from the min-max theorem [11].
Then we have:

$$\lambda_{min}(n^{-1}X^TX) = \lambda_{min}(n^{-1}\Sigma^{1/2}Z^TZ\Sigma^{1/2}) \overset{\text{by (4.1)}}{=} \min_{v\in\mathbb{R}^p\setminus\{0\}} n^{-1}\frac{<\Sigma^{1/2}Z^TZ\Sigma^{1/2}v, v>}{<v,v>}$$
$$= \min_{v\in\mathbb{R}^p\setminus\{0\}} n^{-1}\frac{<Z^TZ\Sigma^{1/2}v, \Sigma^{1/2}v>}{<v,v>} = \min_{v\in\mathbb{R}^p\setminus\{0\}} n^{-1}\frac{<Z^TZ\Sigma^{1/2}v, \Sigma^{1/2}v>}{<v,v>}\frac{<\Sigma v, v>}{<\Sigma^{1/2}v, \Sigma^{1/2}v>}$$
$$= \min_{v\in\mathbb{R}^p\setminus\{0\}} n^{-1}\frac{<\Sigma v, v>}{<v,v>}\frac{<Z^TZ\Sigma^{1/2}v, \Sigma^{1/2}v>}{<\Sigma^{1/2}v, \Sigma^{1/2}v>} \overset{\text{by (4.1)}}{\geq} \lambda_{min}(\Sigma) \min_{v\in\mathbb{R}^p\setminus\{0\}} n^{-1}\frac{<Z^TZ\Sigma^{1/2}v, \Sigma^{1/2}v>}{<\Sigma^{1/2}v, \Sigma^{1/2}v>}$$
$$\overset{\text{by (4.1)}}{=} \lambda_{min}(\Sigma)\lambda_{min}(n^{-1}Z^TZ)$$

By the assumption for the smallest eigenvalue of Σ ***(ii)*** we have:

$$\lambda_{min}(\Sigma)\lambda_{min}(n^{-1}Z^TZ) \geq c\lambda_{min}(n^{-1}Z^TZ).$$

Further by Theorem 3.7 (uses assumption ***(i)***) we have that:

$$\lambda_{min}(\Sigma)\lambda_{min}(n^{-1}Z^TZ) \geq c(1-\sqrt{\gamma})^2 \quad \text{almost surely as } n,p\to\infty, p/n\to\gamma<1$$

Thus for $n,p\to\infty$ and $p/n\to\gamma<1$ by ***(iii)***:

$$\lambda_{min}(n^{-1}X^TX) \geq \lambda_{min}(\Sigma)\lambda_{min}(n^{-1}Z^TZ) \geq c(1-\sqrt{\gamma})^2 \tag{4.2}$$

The right hand side of the inequality is strictly positive, thus by the discussion in the beginning we have proven that $\hat{\Sigma}$ is almost surely invertible and $B_X(\hat{\beta};\beta)$ is almost surely 0.

Step 2. In this step we simplify $V_X(\hat{\beta};\beta)$ so that it depends on the sum of the reciprocals of eigenvalues of $n^{-1}Z^TZ$. This way we will be able to express the variance $V_X(\hat{\beta};\beta)$ in terms of an expectation with respect to the empirical spectral distribution of $n^{-1}Z^TZ$. That will give us the opportunity to apply Marčenko-Pastur Theorem 3.8.

$$\begin{aligned}
V_X(\hat{\beta};\beta) &\overset{\text{Prop.2.8}}{=} \frac{\sigma^2}{n}\text{trace}(\hat{\Sigma}^{\dagger}\Sigma) \overset{\text{a.s.}}{=} \frac{\sigma^2}{n}\text{trace}(\hat{\Sigma}^{-1}\Sigma) = \frac{\sigma^2}{n}\text{trace}((\frac{X^TX}{n})^{-1}\Sigma)\\
&= \frac{\sigma^2}{n}\text{trace}(\Sigma^{-1/2}(\frac{Z^TZ}{n})^{-1}\Sigma^{-1/2}\Sigma) = \frac{\sigma^2}{n}\text{trace}((\frac{Z^TZ}{n})^{-1}\Sigma^{-1/2}\Sigma\Sigma^{-1/2})\\
&= \frac{\sigma^2}{n}\text{trace}((\frac{Z^TZ}{n})^{-1}) = \frac{\sigma^2}{n}\text{trace}(UD^{-1}U^T) = \frac{\sigma^2}{n}\text{trace}(D^{-1}U^TU) = \frac{\sigma^2}{n}\text{trace}(D^{-1})\\
&= \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{1}{\lambda_i(\frac{Z^TZ}{n})} = \frac{\sigma^2 p}{n}\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i(\frac{Z^TZ}{n})} = \frac{\sigma^2 p}{n}\int_0^{\infty}\frac{1}{x}dF_{\frac{Z^TZ}{n}}(x)
\end{aligned}$$

where in the third line we used the eigenvalue decomposition of $n^{-1}Z^TZ = U^TDU$ and noted that D is a diagonal matrix with the eigenvalues $\lambda_i(n^{-1}Z^TZ)$, $i = 1,\dots,p$, of $n^{-1}Z^TZ$ and U is an orthogonal matrix with the corresponding eigenvectors.

Step 3. It is not possible to compute the last integral expression for the variance because of the randomness of $n^{-1}Z^TZ$. We need to apply Marčenko-Pastur Theorem 3.8 so that we obtain an expectation with respect to the Marčenko-Pastur law 3.8 which depends only on γ. From Theorem 3.8 with $\sigma_z^2 = 1$ and ***(i)*** we have that $F_{n^{-1}Z^TZ}$ converges almost surely (hence weakly as well) to the Marčenko-Pastur law F_γ. If we set $f(x) = \frac{1}{x}I(x \geq a)$, where $a = (1-\sqrt{\gamma})^2$, in Portmanteau theorem (3.5) then we have that as $n, p \to \infty$ with $p/n \to \gamma < 1$ ***(iii)***:

$$\begin{gathered}
\mathbb{E}_{F_{\frac{Z^TZ}{n}}}(f(x)) \to \mathbb{E}_{F_\gamma}(f(x)) \quad \text{i.e.}\\
\int_a^{\infty}\frac{1}{x}dF_{\frac{Z^TZ}{n}}(x) \to \int_a^{\infty}\frac{1}{x}dF_\gamma(x)
\end{gathered}$$

We also have that $\int_a^\infty \frac{1}{x}dF_\gamma(x) = \int_a^b \frac{1}{x}dF_\gamma(x)$, where $b = (1+\sqrt{\gamma})^2$, because by the Marčenko-Pastur Theorem 3.8 we know that the support of $F_\gamma(x)$ is $[a, b]$.

Hence, so far we have obtained that as $n, p \to \infty$ with $p/n \to \gamma < 1$ ***(iii)***:

$$V_X(\hat{\beta};\beta) \to \sigma^2\gamma\int_a^b \frac{1}{x}dF_\gamma(x) \qquad \text{almost surely} \tag{4.3}$$

Step 4. In this step we compute the last integral using the Stieltjes transform (Definition 3.9) $m_\gamma(-z)$ of the Marčenko-Pastur law F_γ and then we will send $z \to 0^+$ for real z. Consider:

$$m_\gamma(-z) = \int_a^b \frac{1}{x+z}dF_\gamma(x) = \int_a^b \frac{1}{x+z}\frac{1}{2\pi x\gamma}\sqrt{(b-x)(x-a)}dx \qquad z \in \mathbb{C}\setminus(0,\infty) \tag{4.4}$$

such that $a = (1-\sqrt{\gamma})^2 = 1 - 2\sqrt{\gamma} + \gamma$ and $b = (1+\sqrt{\gamma})^2 = 1 + 2\sqrt{\gamma} + \gamma$.

Claim: We have the following closed form for $m_\gamma(-z) = \dfrac{-(1-\gamma+z) + \sqrt{(1+\gamma+z)^2 - 4\gamma}}{2\gamma z}$

$$\tag{4.5}$$

where $\gamma \in (0, \infty)$. A proof for the above claim is provided in Appendix 8.4

Step 5. Finally, we derive the variance $V_X(\hat{\beta};\beta)$ from (4.3) by sending $z \to 0^+$ in $m_\gamma(-z)$ (4.5):

$$\begin{aligned} V_X(\hat{\beta};\beta) &\to \sigma^2\gamma \lim_{z\to 0^+} \frac{-(1-\gamma+z) + \sqrt{(1+\gamma+z)^2 - 4\gamma}}{2\gamma z} \\ &= \sigma^2\gamma \lim_{z\to 0^+} \frac{-1 + \frac{1+\gamma+z}{\sqrt{(1+\gamma+z)^2-4\gamma}}}{2\gamma} \qquad \text{by L'Hopital's rule} \\ &= \sigma^2\gamma \lim_{z\to 0^+} \frac{-1 + \frac{1+\gamma}{1-\gamma}}{2\gamma} = \sigma^2 \frac{\gamma}{1-\gamma} \end{aligned}$$

As the bias $B_X(\hat{\beta};\beta)$ is almost surely 0, the result for $R_X(\hat{\beta};\beta)$ follows. □

5 Overparametrized regime

In this section we analyze the asymptotic behaviour of the prediction error when $n, p \to \infty$ while $p/n \to \gamma > 1$. We examine only the case with **isotropic** features, i.e. $\Sigma = I$. We will see that the derivation of the limiting variance error is similar to the one in the underparametrized case, however, the limiting bias term will not be almost surely 0 anymore because X^TX will not be almost surely invertible. In order to compute the limiting bias we will need a generalized version of the Marčenko-Pastur law. In the end we will calculate the limit of the expected l_2-norm of $\hat{\beta}$, i.e. we will compute the limit of $\mathbb{E}(||\hat{\beta}||_2^2|X)$, and will see that it has a very similar form as the test risk $R_X(\hat{\beta};\beta)$. This will give a theoretical evidence why overparametrized solutions might work well. Indeed the fact that the test error $\hat{\beta}$ and its norm follow similar formulas will indicate that estimators based on highly parametrized models with small risk will have small norms. As a consequence, when we increase the number of parameters we do not choose a more *complex* solution, but we allow the model to choose a *simpler* one in terms of having smaller norm.

5.1 Generalized Marčenko-Pastur Theorem

In this subsection we briefly describe a spectral convergence theorem for random matrices proposed by Rubio and Mestre in [15]. Firstly, we explain why we would need this theorem when deriving the asymptotic bias of the test risk. Secondly, we give a simplified version of Rubio and Mestre Theorem [15] to ease the reader from heavy notation.

Recall that the limiting bias in underparametrized regime is almost surely equal to 0 because X^TX is almost surely invertible. However, in overparametrized regime we do not have that and thus, we will rely on the next crucial definition for Moore-Penronse pseudoinverse of X^TX. It "fixes" the singularity of the non-invertible X^TX by adding a non-singular diagonal matrix zI.

Definition 5.1. *(Limiting definition of pseudoinverse) Let $X \in \mathbb{R}^{n\times p}$. Then the pseudoinverse of X can be characterized as:*

$$X^\dagger = \lim_{z\to 0^+} (X^TX + zI)^{-1}X^T \quad \textit{equivalently,} \tag{5.1}$$

$$(X^TX)^\dagger X^T = \lim_{z\to 0^+} (X^TX + zI)^{-1}X^T \tag{5.2}$$

The equivalence relation in the definition follows directly from the fact that $X^\dagger = (X^TX)^\dagger X^T$, which was proved in Proposition 2.5. Equation (5.2) is very intuitive because what we do is that if X^TX is not invertible, then we add a non-singular term zI to make it invertible and then we send the variable z to 0^+. The next result gives a limiting characterization to the term $\hat{\Sigma}^\dagger\hat{\Sigma}$ which is part of the bias $B_X(\hat{\beta};\beta)$.

Proposition 5.2. *Let $\hat{\Sigma} = n^{-1}X^TX$ be the sample variance. Then:*

$$\hat{\Sigma}^\dagger\hat{\Sigma} = \lim_{z\to 0^+} (\hat{\Sigma} + zI)^{-1}\hat{\Sigma}$$

Proof. From Definition 5.1 we have $(X^TX)^\dagger X^T = \lim_{z\to 0^+}(X^TX + zI)^{-1}X^T$. Right multiply by X to get $(X^TX)^\dagger X^TX = \lim_{z\to 0^+}(X^TX + zI)^{-1}X^TX$ and hence, $\hat{\Sigma}^\dagger\hat{\Sigma} = \lim_{z\to 0^+}(\hat{\Sigma} + zI)^{-1}\hat{\Sigma}$. □

The above representation of $\hat{\Sigma}^\dagger\hat{\Sigma}$ involves the term $(\hat{\Sigma}+zI)^{-1}$. We are interested in the proportional limit of this term as $n, p \to \infty$, where $p/n \to \gamma > 1$. This motivates us to use the next spectral convergence theorem proposed by Rubio and Mestre [15]. It relates the random quantity trace$((\hat{\Sigma}+zI)^{-1})$ with deterministic quantities $\Theta_n, c_n(z)$.

Theorem 5.3. *(Rubio and Mestre theorem[15]) Assume the features x have finite moment of order $8+\mu$ and $\hat{\Sigma} = n^{-1}X^TX$ is the sample covariance. For any $z \in \mathbb{C}\setminus(0,\infty)$ and any deterministic sequence of matrices $\Theta_n \in \mathbb{R}^{p\times p}, n = 1, 2, 3, \dots$ with uniformly bounded trace norm $||\Theta||_{tr} = trace((\Theta^T\Theta)^{1/2})$, it is true with probability 1 that as $n, p \to \infty$, where $p/n \to \gamma > 1$ that:*

$$trace\Big(\Theta_n((\hat{\Sigma} + zI)^{-1} - c_n(z)I)\Big) \to 0$$

for a deterministic sequence $c_n(z), n = 1, 2, 3, \dots$ which depends on Θ_n, p/n, $\hat{\Sigma}$ and z. Explicit definition of $c_n(z)$ is given in [15].

The above theorem is often called the ***Generalized Marčenko-Pastur Theorem*** because for $\Theta_n = p^{-1}I$ we can derive the Marčenko-Pastur Theorem 3.8. Indeed, if we substitute $\Theta_n = p^{-1}I$ in Theorem 5.3, then we have:

$$\text{trace}\left(\frac{1}{p}(\hat{\Sigma} + zI)^{-1} - \frac{1}{p}c_n(z)I\right) \to 0$$

$$\text{trace}\left(\frac{1}{p}(\hat{\Sigma} + zI)^{-1}\right) - \text{trace}\left(\frac{1}{p}c_n(z)I\right) \to 0 \tag{5.3}$$

The first term is equal to the Stieltjes transform $m_{\hat{\Sigma}}(z)$:

$$\text{trace}(\frac{1}{p}(\hat{\Sigma}+zI)^{-1}) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i(\hat{\Sigma})+z} = \int\frac{dF_{\hat{\Sigma}}(x)}{x+z} = m_{\hat{\Sigma}}(-z)$$

where $\lambda_i(\hat{\Sigma})$ are the eigenvalues of $\hat{\Sigma}$. The second term in (5.3) is just $c_n(z)$ and one can derive from its definition given in [15] that $c_n(z) \to m_\gamma(-z)$. Hence, from (5.3) follows $m_{\hat{\Sigma}}(-z) \to m_\gamma(-z)$. This is equivalent to Marčenko-Pastur Theorem 3.8 because the limit of $F_{\hat{\Sigma}}$ is uniquely characterized by that of $m_{\hat{\Sigma}}$ [17].
The main application of the Generalized Marčenko-Pastur Theorem is that we are able to analyze the asymptotic behaviour of a random quantity depending on $\hat{\Sigma}$ using a deterministic quantity depending on the aspect ratio γ. For example, we already used in Theorem 4.1 the important consequence for convergence of Stieltjes transformations (expectations) $m_{\hat{\Sigma}}(z) \to m_\gamma(z)$ as $n, p \to \infty$. This was very useful because $m_{\hat{\Sigma}}(z)$ is random whereas $m_\gamma(z)$ is deterministic. With the help of the Generalized Marčenko-Pastur Theorem we would be able to consider convergence of expectations to more general distributions than the empirical spectral distribution $F_{\hat{\Sigma}}$ (Remark 5.5).

5.2 Asymptotic risk

Now we can derive the asymptotic risk for overparametrized regime with **isotropic** features. The proof of the next theorem replicates the proof of Lemma 2 and Lemma 3 given in [1]. We will clearly indicate where each of the assumptions of the theorem is used.

Theorem 5.4. *Consider model* ***(1)*** *with the following assumptions:*

- *the feature vectors x has i.i.d with $\Sigma = I$ and a finite moment of order $8+\mu$, for some $\mu > 0$.* ***(i)***
- *$||\beta||_2^2 = r^2$ for all n, p* ***(ii)***
- *$n, p \to \infty$, such that $p/n \to \gamma > 1$* ***(iii)***

Then the generalization error of the estimator $\hat{\beta} = (X^TX)^\dagger X^Ty$ satisfies:

$$R_X(\hat{\beta};\beta) \to r^2\left(1-\frac{1}{\gamma}\right)+\frac{\sigma^2}{\gamma-1} \qquad \textit{almost surely}$$

Proof. Firstly, we derive the limit of $V_X(\hat{\beta};\beta)$. From Proposition 2.8 with $\Sigma = I$ by ***(i)***:

$$V_X(\hat{\beta};\beta) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i(\frac{X^TX}{n})} \tag{5.4}$$

where $\lambda_i(n^{-1}X^TX)$, $i = 1, \ldots, n$ are the **non-zero** eigenvalues of $n^{-1}X^TX$. Now consider the eigenvalues $\mu_i(p^{-1}XX^T)$, $i = 1, \ldots, p$ of the Gram matrix $p^{-1}XX^T$.

Consider the SVD of X, $X = U^T D V$, where U is $n \times n$ orthogonal matrix, V is $p \times p$ orthogonal matrix and D is a diagonal matrix. Then:

$$\begin{aligned} X^T X &= V^T D^T U U^T D V = V^T D^T D V \\ X X^T &= U^T D V V^T D^T U = U^T D D^T U \end{aligned}$$

From both equations we see that the eigenvalues of $X^T X$ and XX^T are the diagonal entries of $D^T D$ and DD^T, which are the same. Therefore, we can write $\lambda_i(n^{-1}X^T X) = \frac{p}{n}\mu_i(p^{-1}XX^T)$ for $i = 1, \ldots, n$. Now we substitute in (5.4):

$$V_X(\hat{\beta};\beta) = \frac{\sigma^2}{p}\sum_{i=1}^{n}\frac{1}{\mu_i(\frac{XX^T}{p})} = \frac{\sigma^2 n}{p}\int \frac{1}{x} dF_{XX^T/p}(x)$$

where $F_{XX^T/p}(x)$ is the empirical spectral distribution of $p^{-1}XX^T$. Observe that in the overparametrized case we have $n/p < 1$ and computing the above integral is essentially doing the same calculations as we did for the variance term in the underparametrized regime. By ***(iii)*** $n/p \to \rho = \gamma^{-1} < 1$ as $n, p \to \infty$ and from Theorem 4.1 we have:

$$V_X(\hat{\beta};\beta) \to \sigma^2 \frac{\rho}{1-\rho} = \frac{\sigma^2}{\gamma - 1}$$

Now we wish to compute the asymptotic bias. We rewrite $B_X(\hat{\beta};\beta)$ using the limiting characterization of $\hat{\Sigma}^{\dagger}\hat{\Sigma}$, from Proposition 5.2.

$$\begin{aligned} B_X(\hat{\beta};\beta) &= \beta^T \Pi\Sigma\Pi\beta = \beta^T(I - \hat{\Sigma}^{\dagger}\hat{\Sigma})\beta \quad \text{ using } \Sigma = I \text{ and } \Pi\Pi = \Pi \\ &= \lim_{z\to 0^+} \beta^T(I - (\hat{\Sigma} + zI)^{-1}\hat{\Sigma})\beta \qquad \text{by Proposition 5.2} \\ &= \lim_{z\to 0^+} \beta^T(I - (\hat{\Sigma} + zI)^{-1}(\hat{\Sigma} + zI) + (\hat{\Sigma} + zI)^{-1} zI)\beta \\ &= \lim_{z\to 0^+} \beta^T((\hat{\Sigma} + zI)^{-1} zI)\beta = \lim_{z\to 0^+} z\beta^T(\hat{\Sigma} + zI)^{-1}\beta \end{aligned} \tag{5.5}$$

Now we can use the Generalized Marčenko-Pastur Theorem 5.3. Note that this theorem requires the features x to have finite moment of order $8 + \mu$ ***(i)***. If we choose $\Theta_n = \beta\beta^T$ in Theorem 5.3 then we have $c_n(z) \to m_\gamma(-z)$ by definition of $c_n(z)$ in [15] and we obtain that as $n, p \to \infty$ and $p/n \to \gamma > 1$ ***(iii)***:

$$\begin{aligned} \text{trace}\left(\beta\beta^T((\hat{\Sigma} + zI)^{-1} - m_\gamma(-z)I)\right) &\to 0 \implies \\ \text{trace}\left(\beta^T(\hat{\Sigma} + zI)^{-1}\beta - \beta^T m_\gamma(-z)I\beta\right) &\to 0 \implies \\ \text{trace}\left(\beta^T(\hat{\Sigma} + zI)^{-1}\beta\right) &\to m_\gamma(-z)\text{trace}(\beta^T\beta) \implies \\ \beta^T(\hat{\Sigma} + zI)^{-1}\beta &\to r^2 m_\gamma(-z) \qquad \text{by } \textit{\textbf{(ii)}} \end{aligned} \tag{5.6}$$

Taking the proportional limit of the bias term from (5.5) gives:

$$\begin{aligned}\lim_{\substack{p,n\to\infty\\p/n\to\gamma>1}} B_X(\hat{\beta};\beta) &= \lim_{\substack{p,n\to\infty\\p/n\to\gamma>1}} \lim_{z\to 0^+} z\beta^T(\hat{\Sigma}+zI)^{-1}\beta\\ &= \lim_{z\to 0^+} \lim_{\substack{p,n\to\infty\\p/n\to\gamma>1}} z\beta^T(\hat{\Sigma}+zI)^{-1}\beta\\ &= \lim_{z\to 0^+} zr^2 m_\gamma(-z) = r^2 \lim_{z\to 0^+} zm_\gamma(-z) \qquad \text{by (5.6)}\\ &= r^2 \lim_{z\to 0^+} z\frac{-(1-\gamma+z)+\sqrt{(1+\gamma+z)^2-4\gamma}}{2\gamma z} \qquad \text{by (4.5)}\\ &= r^2\frac{-1+\gamma+\gamma-1}{2\gamma} = r^2(1-\frac{1}{\gamma})\end{aligned}$$

where the exchange of limits in the second line requires some care; details for this step are given in the proof of Lemma 2 in [1]. The result for the test risk follows immediately. □

In the next remark we give some insight behind the Generalized Marčenko-Pastur Theorem 5.3.

Remark 5.5. *In the proof of Theorem 5.4 when we derived the bias term we used the Generalized Marčenko-Pastur Theorem 5.3 with $\Theta_n = \beta\beta^T$. We derived in (5.6) that*

$$\beta^T(\hat{\Sigma}+zI)^{-1}\beta \to r^2 m_\gamma(-z) \tag{5.7}$$

This can be interpreted as a convergence in Stieltjes transformations/expectations statement. Indeed the above convergence is equivalent to:

$$m_{G_{\hat{\Sigma}}}(-z) \to m_\gamma(-z) \qquad \textit{almost surely as } n,p\to\infty, p/n\to\gamma>1 \tag{5.8}$$

where we used the ***generalized spectral empirical distribution*** *defined the following way:*

$$G_{\hat{\Sigma}}(x) = \sum_{i=1}^{p} |w_i|^2 I\{\lambda_i(\hat{\Sigma})\le x\} \tag{5.9}$$

where $\boldsymbol{w} = (w_1,\dots,w_p)$ is given by $\boldsymbol{w} = U\beta r^{-1}$ with U having the orthonormal eigenvectors of $\hat{\Sigma}$ and $r^2 = ||\beta||_2^2$. Note that the sum of all $|w_i|^2$ for $i = 1,\dots,p$ adds to 1, therefore, $G_{\hat{\Sigma}}(x)$ is well-defined. Indeed, $\sum|w_i|^2 = ||\boldsymbol{w}||_2^2 = ||U\beta r^{-1}||_2^2 = ||\beta r^{-1}||_2^2 = 1$. The result of the above equivalence in convergences is derived in Appendix 8.5.

Now we can summarize the results from Theorem 4.1 and Theorem 5.4 to obtain full specification for the generalization risk in **isotropic** case.

Theorem 5.6. *Consider model **(1)** with the assumptions in Theorem 5.4. Then the limit $R(\gamma)$ as $n,p\to\infty, p/n\to\gamma$, of the generalization risk $R_X(\hat{\beta};\beta)$ for the min-norm least squares estimator depends only on the aspect ratio γ, the signal strength r^2 and the noise σ^2.*

$$R(\gamma) = \begin{cases} \sigma^2\dfrac{\gamma}{1-\gamma} & \textit{for } \gamma<1\\ r^2\left(1-\dfrac{1}{\gamma}\right)+\sigma^2\dfrac{1}{\gamma-1} & \textit{for } \gamma>1\end{cases}$$

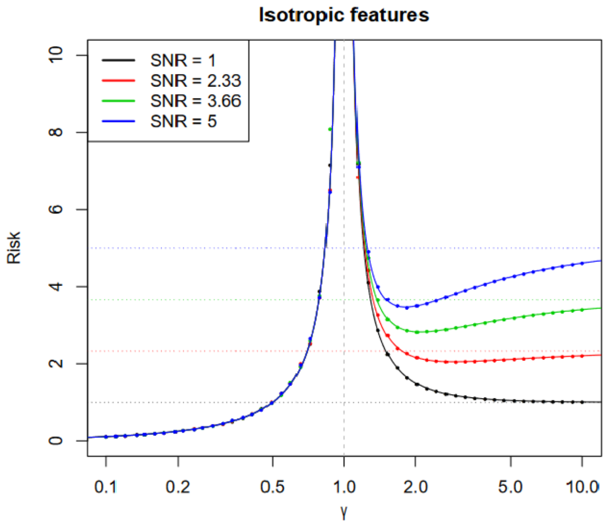

Figure 4: Taken from [8]. The solid lines are asymtotic risk cures in Theorem 5.6 for the least squares estimator, when the signal r^2 varies from 1 to 5, and the noise is 1. In the underparametrized regime the curve is the same for different signal stregths as it does not depend on r^2.

One can easily see that the global minimum of the asymptotic test risk is attained trivially at $\gamma = 0$, Figure 4. That is because we are in a linear regression setting where $\gamma = 0$ implies that the number of features we have is $p = 0$. Then the bias and variance will be both 0. Hence, no features - no generalization error. This is a consequence of the **ambiguity of** p which we already mentioned in Section 2. As a result, linear regression models in overparametrized regime will never achieve a smaller asymptotic risk than their counterparts in underparametrized regime. Still the result for the limit of the generalization risk curve is impressive because we observe decreasing risk after the interpolation threshold, Figure 4. This defies the traditional understanding in statistics that having too many parameters will overfit.

Theorem 5.6 shows that the asymptotic test error is pure variance when $\gamma < 1$. It also goes to infinity when γ approaches 1. There is no bias because the number of parameters p we wish to learn is smaller than the number of observations n. This means we have enough data to match the true parameters.

The more interesting regime is when $\gamma > 1$. One can observe that the bias-variance trade-off is controlled by the signal strength r^2 and the noise σ^2. Therefore, the shape of the asymptotic generalization risk depends on the signal-to-noise ratio (SNR) r^2/σ^2. We can summarize the main observations in overparametrized regime in the next statements:

1. If SNR > 1, then the smallest test error we can achieve is at the **local** minimum $\gamma = \frac{\sqrt{SNR}}{\sqrt{SNR}-1}$ (set $dR(\gamma)/d\gamma = 0$).

2. At $\gamma = \frac{\sqrt{SNR}}{\sqrt{SNR}-1}$, the expected l_2-norm of the least squares estimator is not optimal and is rather high (Proposition 5.7).

3. If SNR$<$ 1 we observe a *monotonically decreasing* test error, i.e the more parameters the better.

4. If SNR$<$ 1, then we have a better learning performance as we increase the number of parameters because the least squares estimator will have a smaller norm (Proposition 5.7. We are able to choose a simpler solution. Intuitively, if $SNR > 1$ with large noise σ^2 then there will be a lot of randomness and we need as many parameters as possible

to learn successfully.
The last statement could give motivation to explain why neural networks work so well in overparametrized regime. Deep Neural Networks try to learn complicated functions with complex loss curve. Initialization near true values of the parameters is based on pure luck. Thus initialization step involves a lot of randomness. By Statement 4 models with large number of parameters will perform better.

5.3 Limiting norm

In this section we derive the asymptotic expected l_2-norm of the estimator $\hat{\beta} = (X^T X)^\dagger X^T y$ for isotropic feature matrix X with i.i.d entries. This result was given in Section 3.4 in [1] but a proof was not provided. Here we will prove the result.

Proposition 5.7. *Consider the setup in Theorem 5.4. If $n, p \to \infty$, such that $p/n \to \gamma$, then almost surely we have:*

$$\mathbb{E}(||\hat{\beta}||_2^2|X) \to \begin{cases} r^2 + \sigma^2 \dfrac{\gamma}{1-\gamma} & \text{for } \gamma < 1 \\ \\ r^2 \dfrac{1}{\gamma} + \sigma^2 \dfrac{1}{\gamma - 1} & \text{for } \gamma > 1 \end{cases}$$

Proof. We will to express $\mathbb{E}(||\hat{\beta}||_2^2|X)$ as a sum of two terms which resemble the bias-variance decomposition of the prediction error $R_X(\hat{\beta}; \beta)$. Our strategy is to split the expected norm into a sum of two terms, one depending on ϵ (the variance) and one which does not (the bias). We do this by expressing y as $y = X\beta + \epsilon$.

$$\begin{aligned}
\mathbb{E}(||\hat{\beta}||_2^2|X) &= \mathbb{E}(||(X^T X)^\dagger X^T y||_2^2|X) = \mathbb{E}(y^T X(X^T X)^\dagger (X^T X)^\dagger X^T y|X) \\
&= \mathbb{E}(\text{trace}\left[y^T X(X^T X)^\dagger (X^T X)^\dagger X^T y\right]|X) \\
&= \mathbb{E}(\text{trace}\left[yy^T X(X^T X)^\dagger (X^T X)^\dagger X^T\right]|X) \qquad \text{trace trick} \\
&= \text{trace}\left[\mathbb{E}(yy^T|X) X(X^T X)^\dagger (X^T X)^\dagger X^T\right] \qquad \text{trace and } \mathbb{E} \text{ are both linear} \\
&= \text{trace}\left[(X\beta\beta^T X^T + \mathbb{E}(\epsilon\epsilon^T|X)) X(X^T X)^\dagger (X^T X)^\dagger X^T\right] \\
&= \text{trace}\left[X\beta\beta^T X^T X(X^T X)^\dagger (X^T X)^\dagger X^T + \mathbb{E}(\epsilon\epsilon^T) X(X^T X)^\dagger (X^T X)^\dagger X^T\right] \\
&= \text{trace}\left[\beta\beta^T X^T X(X^T X)^\dagger (X^T X)^\dagger X^T X\right] + \text{trace}\left[\sigma^2 (X^T X)^\dagger (X^T X)^\dagger X^T X\right] \\
&= \text{trace}\left[\beta^T \hat{\Sigma}\hat{\Sigma}^\dagger \hat{\Sigma}^\dagger \hat{\Sigma}\beta\right] + \frac{\sigma^2}{n}\text{trace}\left[\hat{\Sigma}^\dagger \hat{\Sigma}^\dagger \hat{\Sigma}\right] \\
&= \beta^T (\hat{\Sigma}\hat{\Sigma}^\dagger)^T \hat{\Sigma}^\dagger \hat{\Sigma}\beta + \frac{\sigma^2}{n}\text{trace}\left[\hat{\Sigma}^\dagger \hat{\Sigma}\hat{\Sigma}^\dagger\right] \qquad \text{Pseudo inverse property } \hat{\Sigma}\hat{\Sigma}^\dagger = (\hat{\Sigma}\hat{\Sigma}^\dagger)^T \\
&= \beta^T \hat{\Sigma}^\dagger \hat{\Sigma}\hat{\Sigma}^\dagger \hat{\Sigma}\beta + \frac{\sigma^2}{n}\text{trace}\left[\hat{\Sigma}^\dagger\right] \qquad \text{Pseudo inverse property } \hat{\Sigma}^\dagger \hat{\Sigma}\hat{\Sigma}^\dagger = \hat{\Sigma}^\dagger \\
&= \underbrace{\beta^T \hat{\Sigma}^\dagger \hat{\Sigma}\beta}_{bias} + \underbrace{\frac{\sigma^2}{n}\text{trace}\left[\hat{\Sigma}^\dagger\right]}_{variance} \qquad \text{Pseudo inverse property } \hat{\Sigma}^\dagger \hat{\Sigma}\hat{\Sigma}^\dagger = \hat{\Sigma}^\dagger
\end{aligned} \tag{5.10}$$

Observe that the last equation resembles the bias and variance errors in Proposition 2.8. Indeed, $n^{-1}\sigma^2\text{trace}[\hat{\Sigma}^\dagger]$ is exactly the same as the variance error when we set $\Sigma = I$,

and we have already derived its asympotics in underparametrized and overparametrized regime to be $\sigma^2 \dfrac{\gamma}{1-\gamma}$ (Theorem 4.1) and $\sigma^2 \dfrac{1}{\gamma - 1}$ (Theorem 5.4), respectively.
As for the first term $\beta^T \hat{\Sigma}^\dagger \hat{\Sigma} \beta$ in (5.10), we know from **Step 1** in the proof of Theorem 4.1 that for $\gamma < 1$ we have $\hat{\Sigma}^\dagger$ is almost surely invertible, hence, $\beta^T \hat{\Sigma}^\dagger \hat{\Sigma} \beta$ is almost surely $\beta^T \beta = r^2$. For $\gamma > 1$ in Theorem 5.4 we have shown that $B_X(\hat{\beta}; \beta) = \beta^T (I - \hat{\Sigma}^\dagger \hat{\Sigma}) \beta \to r^2 (1 - \dfrac{1}{\gamma})$ almost surely. Note the LHS is $\beta^T \beta - \beta^T \hat{\Sigma}^\dagger \hat{\Sigma} \beta = r^2 - \beta^T \hat{\Sigma}^\dagger \hat{\Sigma} \beta$. Therefore, $\beta^T \hat{\Sigma}^\dagger \hat{\Sigma} \beta$ converges almost surely to $\dfrac{r^2}{\gamma}$. QED. □

Observe that the limit of the expected l_2-norm $\hat{\beta}$ (called *the asymptotic norm*, Proposition 5.7) has a similar behaviour as the limit of its generalization risk (called *the asymptotic risk*, Theorem 5.6). The asymptotic norm depends on $\hat{\beta}$ and the asymptotic risk depends on β and $\hat{\beta}$. They have exactly the same variance terms because the variances does not depend on β. We can make the following observations:
1. When $\gamma < 1$ the variance for the asymptotic risk and the asymptotic norm both increase as γ increase.
2. When $\gamma > 1$ the variance for the asymptotic risk and the asymptotic norm both decrease as γ increases.
3. The main difference between the asymptotic norm and the asymptotic risk is their bias.
4. When $\gamma < 1$, the asymptotic norm has a bias equal to the signal strength r^2 and the asymptotic risk has bias equal to 0.
5. When $\gamma > 1$ the asymptotic norm has a decreasing bias but the asymptotic risk has an increasing bias.
6. Interestingly, the sum of the biases of the asymptotic risk and the asymptotic norm is always $r^2 + 0 = \gamma^{-1} r^2 + r^2 (1 - \gamma^{-1})$ and does not depend on γ. This means that we cannot have it both ways - either the asymptotic norm or the asymptotic risk should have large bias.
7. If the noise σ^2 is significantly larger than the signal strength r^2, then the asymptotic risk and the asymptotic norm will behave similarly. In this case choosing more parameters leads to choosing a simpler solution with smaller asymptotic norm.

6 Ridge Regression

In this section we derive the asymptotic predictive risk of the least squares estimator for the ridge regression problem. The loss is $L(\beta; y, x) = n^{-1} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2$. In Section 5 we examined only the case with isotropic features $\Sigma = I$. Here we will consider a general Σ. Note if we send the regularization parameter $\lambda \to 0$, then we will be in the ridgless case we considered in Sections 2-5. In order to make the calculations tractable, we work in a Bayesian setup by imposing a prior belief on the parameter β we want to learn. Observe that the generalization risk $R_X(\hat{\beta}_\lambda; \beta)$ depends on β which is a random variable now. Thus we will consider the **Bayes risk** which is just the *expected* generalization risk, where expectation is taken with respect to β.

6.1 Setup

Suppose we have n i.i.d training samples
$(x_i, y_i) \in \mathbb{R}^{p\times 1} \times \mathbb{R}, i = 1, 2, \ldots, n$, where each sample is generated independently from the following model ***(2)***:

- Draw feature x_i, noise ϵ_i and response y_i exactly the same way as in model **(1)**.

- Assume a Bayesian framework with a prior on the parameter $\beta \sim P_\beta$ satisfying $\mathbb{E}(\beta) = 0$ and $Cov(\beta) = \frac{r^2}{p}I$, where $r^2 = \mathbb{E}(||\beta||_2^2)$ is the expected signal strength.

Assuming that we have random regression coefficients β partially resembles the initialization procedure in neural networks where we often set an initial value to the parameters using a certain distribution. This distribution acts on the parameters the same way the prior does in Bayesian regression learning.

Proposition 6.1. *Let $X \in \mathbb{R}^{n\times p}$ be the feature matrix. The ridge regression estimator $\hat\beta_\lambda = (X^TX + n\lambda I)^{-1}X^Ty$ minimizes the loss $L(\beta; y, x)$ in the ridge regression problem given by $L(\beta; y, x) = n^{-1}||y - X\beta||_2^2 + \lambda||\beta||_2^2$.*

Proof. The loss $L(\beta; y, x) = n^{-1}||y - X\beta||_2^2 + \lambda||\beta||_2^2$ is convex and thus, if we set its derivative with respect to β to be 0 we will obtain a minimizing solution.

$$\left.\frac{dL(\beta; y, x)}{d\beta}\right|_{\beta=\hat\beta_\lambda} = 0 \iff -\frac{2}{n}X^T(y - X\hat\beta_\lambda) + 2\lambda\hat\beta_\lambda = 0 \iff$$
$$X^T(y - X\hat\beta_\lambda) = n\lambda\hat\beta_\lambda \iff X^Ty = (X^TX + n\lambda I)\hat\beta_\lambda \iff$$
$$\hat\beta_\lambda = (X^TX + n\lambda I)^{-1}X^Ty$$

□

Definition 6.2. *Assume model **(2)**. The **integrated** or **Bayes risk** for $\hat\beta_\lambda$ is $R_X(\hat\beta_\lambda) = \mathbb{E}_\beta(R_X(\hat\beta_\lambda; \beta))$, where $R_X(\hat\beta_\lambda; \beta)$ is defined as before in Definition 2.2.*

Proposition 6.3. *Assume model **(2)**. The bias-variance decomposition of the Bayes risk $R_X(\hat\beta_\lambda)$ for an estimator $\hat\beta_\lambda$ is:*

$$R_X(\hat\beta_\lambda) = \underbrace{\mathbb{E}_\beta(B_X(\hat\beta_\lambda; \beta))}_{B_X(\hat\beta_\lambda)} + \underbrace{\mathbb{E}_\beta(V_X(\hat\beta_\lambda; \beta))}_{V_X(\hat\beta_\lambda)} \quad \textit{where}$$
$$B_X(\hat\beta_\lambda; \beta) = ||\mathbb{E}(\hat\beta_\lambda|X) - \beta||_\Sigma^2 \;\; \textit{and} \;\; V_X(\hat\beta_\lambda; \beta) = \textit{trace}(Cov(\hat\beta_\lambda|X)\Sigma)$$

Proof. The result follows directly from Proposition 2.4. □

The above proposition defines the Bayesian bias $B_X(\hat\beta_\lambda)$ and the Bayesian variance $V_X(\hat\beta_\lambda)$ which note that depend only on $\hat\beta_\lambda$. In the next proposition we rewrite the Bayesian bias and variance in such a way that we would be able to apply the convergence theorem for random matrices proposed by Ledoit and Peche in [18].

Proposition 6.4. *Assume model (2). Let $r^2 = \mathbb{E}(||\beta||_2^2)$ be the expected signal strength. If we use the notation in Proposition 6.3, then the bias and variance terms of the Bayes risk for the estimator $\hat{\beta}_\lambda = (X^TX + \lambda nI)^{-1}X^Ty$ satisfy:*

$$B_X(\hat{\beta}_\lambda) = (\lambda n)^2 \frac{r^2}{p} trace(\Sigma(X^TX + \lambda nI)^{-2})$$
$$V_X(\hat{\beta}_\lambda) = \sigma^2 trace(\Sigma(X^TX + \lambda nI)^{-1}X^TX(X^TX + \lambda nI)^{-1})$$

Proof. The proof is provided in Appendix 8.6 □

Before we prove the theorem for the asymptotic risk in ridge regression we need the next definition.

Definition 6.5. *Assume $m_H(z)$ be the Stieltjes transform for a measure H. Then the companion Stieltjes transform v_H is given by:*

$$v_H(z) + 1/z = \gamma(m_H(z) + 1/z)$$

6.2 Asymptotic risk

In this section we derive the asymptotic generalization risk for the ridge estimator $\hat{\beta}_\lambda$. The main contribution for the proof is given by Dobriban and Wager in Theorem 2.1 [16] . Our formulation of the theorem slightly differs the one proposed in [16] because we consider different generalization risk. We use the risk given in Definition 6.2 whereas Dobriban and Wager consider a shifted by σ^2 risk, recall Proposition 2.1.

Theorem 6.6. *Consider model (2) with the following assumptions:*

- *$x \sim F_x$ and $x = \Sigma^{1/2}z$ for a random vector z with i.i.d isotropic entries and a finite 12th moment.* ***(i)***
- *Σ is deterministic and positive definite matrix such that its extreme eigenvalues satisfy $0 < c \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq C$, for all n, p and constants c, C.* ***(ii)***
- *As $n, p \to \infty$ the empirical spectral distribution F_Σ of Σ converges weakly to a measure H.* ***(iii)***
- *$n, p \to \infty$, such that $p/n \to \gamma > 1$.* ***(iv)***

Then the prediction risk for $\hat{\beta}_\lambda = (X^TX + \lambda nI)^{-1}X^Ty$ satisfies:

$$R_X(\hat{\beta}_\lambda) \to \frac{r^2}{\gamma}\left(\frac{1}{v(-\lambda)} - \frac{\lambda v'(-\lambda)}{v(-\lambda)^2}\right) + \sigma^2\left(\frac{v'(-\lambda)}{v(-\lambda)^2} - 1\right)$$

where $v = v_H$ is the companion Stieltjes transform for measure H. By $v'(-\lambda)$ we denote the derivative of the companion Stieltjes transform, $v'(z)$, evaluated at $z = -\lambda$.

Proof. As before assume the sample variance is $\hat{\Sigma} = n^{-1}X^TX$. Then we can plug in the bias and variance terms from Proposition 6.4 :

$$\begin{aligned}
V_X(\hat{\beta}_\lambda) &= \frac{\sigma^2}{n}\text{trace}\left(\Sigma(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\right)\\
&= \frac{\sigma^2}{n}\text{trace}\left(\Sigma(\hat{\Sigma}+\lambda I)^{-1}(\hat{\Sigma}+\lambda I)(\hat{\Sigma}+\lambda I)^{-1} - \lambda\Sigma(\hat{\Sigma}+\lambda I)^{-2}\right)\\
&= \sigma^2\frac{p}{n}\text{trace}\left(\frac{\Sigma(\hat{\Sigma}+\lambda I)^{-1}}{p}\right) - \sigma^2\lambda\frac{p}{n}\text{trace}\left(\frac{\Sigma(\hat{\Sigma}+\lambda I)^{-2}}{p}\right)\\
&\to \sigma^2\gamma\frac{1}{\gamma}\left(\frac{1}{\lambda v(-\lambda)}-1\right) - \sigma^2\lambda\gamma\frac{v(-\lambda)-\lambda v'(-\lambda)}{\gamma\lambda^2 v(-\lambda)^2} \qquad \text{as } n,p\to\infty,\ p/n\to\gamma\\
&= \sigma^2\left(\frac{1}{\lambda v(-\lambda)}-1\right) - \sigma^2\left(\frac{1}{\lambda v(-\lambda)} - \frac{v'(-\lambda)}{v(-\lambda)^2}\right)\\
&= \sigma^2\left(\frac{v'(-\lambda)}{v(-\lambda)^2}-1\right)
\end{aligned}$$

where in the fourth line we use two convergence results given in Section 6, [16]. The convergence of the first trace term involving functionals $\hat{\Sigma}, \Sigma$ follows directly from the theorem of Ledoit and Peche, [18]. Note that this theorem uses assumptions ***(i)*** and ***(iii)***. The convergence of the second trace term is shown in [16] and uses assumption ***(ii)***. We can derive the limit of the bias $B_X(\hat{\beta}_\lambda)$ following similar strategy as in the variance:

$$\begin{aligned}
B_X(\hat{\beta}_\lambda) &= (\lambda n)^2\frac{r^2}{p}\text{trace}\left(\Sigma(n\hat{\Sigma}+n\lambda I)^{-2}\right) \qquad \text{by Proposition 6.4}\\
&= \lambda^2 r^2\text{trace}\left(\frac{\Sigma(\hat{\Sigma}+\lambda I)^{-2}}{p}\right)\\
&\to \lambda^2 r^2\frac{v(-\lambda)-\lambda v'(\lambda)}{\gamma\lambda^2 v(-\lambda)^2} \qquad \text{as } n,p\to\infty,\ p/n\to\gamma;\ \text{using Theorem 2.1 [16]}\\
&= \frac{r^2}{\gamma}\left(\frac{1}{v(-\lambda)} - \frac{\lambda v'(-\lambda)}{v(-\lambda)^2}\right)
\end{aligned}$$

The result for the generalization risk follows. □

Remark 6.7. *If we send the regularization paramete $\lambda \to 0$, then we obtain the asymptotic risk for ridgeless regression also provided in Theorem 3. in [1].*

7 Conclusion

In our analyses we have seen the explicit derivation of the asymptotic risk for the least squares estimator in linear regression. We have observed the double descent curve as a function of the aspect ratio γ, essentially the model size. In [20], it is shown that the double descent curve is also observed as a function of training time measured in number of training epochs. This is remarkable because we can achieve better generalization risk by either increasing the number of parameters or by training longer. In conclusion, the double descent style learning curve can give possible explanation of the success of modern deep learning models where larger models are usually better.

8 Appendix

8.1 Proof of Proposition 2.4

By Proposition 2.3 we have that the test error satisfies:

$$
\begin{aligned}
R_X(\hat{\beta};\beta) &= \mathbb{E}[(\hat{\beta}-\beta)^T\Sigma(\hat{\beta}-\beta)|X] \\
&= \mathbb{E}[\beta^T\Sigma\beta|X] - \mathbb{E}[\beta^T\Sigma\hat{\beta}|X] - \mathbb{E}[\hat{\beta}^T\Sigma\beta|X] + \mathbb{E}[\hat{\beta}^T\Sigma\hat{\beta}|X] \\
&= \mathbb{E}[\beta^T\Sigma\beta|X] - \mathbb{E}[\beta^T\Sigma\hat{\beta}|X] - \mathbb{E}[\hat{\beta}^T\Sigma\beta|X] + \mathbb{E}[\hat{\beta}^T\Sigma\hat{\beta}|X] + \mathbb{E}[\hat{\beta}|X]^T\Sigma\mathbb{E}[\hat{\beta}|X] \\
&- \mathbb{E}[\hat{\beta}|X]^T\Sigma\mathbb{E}[\hat{\beta}|X] \qquad \text{add and subtract } \mathbb{E}[\hat{\beta}|X]^T\Sigma\mathbb{E}[\hat{\beta}|X] \\
&= \underbrace{\beta^T\Sigma\beta - \beta^T\Sigma\mathbb{E}[\hat{\beta}|X] - \mathbb{E}[\hat{\beta}^T|X]\Sigma\beta + \mathbb{E}[\hat{\beta}|X]^T\Sigma\mathbb{E}[\hat{\beta}|X]}_{\text{Bias}} \\
&+ \underbrace{\mathbb{E}[\hat{\beta}^T\Sigma\hat{\beta}|X] - \mathbb{E}[\hat{\beta}|X]^T\Sigma\mathbb{E}[\hat{\beta}|X]}_{\text{Variance}} \\
&= \underbrace{||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2}_{Bias} + \underbrace{\mathbb{E}[\text{trace}(\hat{\beta}\hat{\beta}^T\Sigma)|X] - \text{trace}\Big(\mathbb{E}[\hat{\beta}|X]\mathbb{E}[\hat{\beta}|X]^T\Sigma\Big)}_{Variance} \\
&\overset{*}{=} ||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2 + \text{trace}\Big\{\Big(\mathbb{E}[\hat{\beta}\hat{\beta}^T|X] - 2\mathbb{E}[\hat{\beta}|X]\mathbb{E}[\hat{\beta}|X]^T + \mathbb{E}[\hat{\beta}|X]\mathbb{E}[\hat{\beta}|X]^T\Big)\Sigma\Big\} \\
&= ||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2 + \text{trace}\Big\{\mathbb{E}\big[(\hat{\beta}-\mathbb{E}(\hat{\beta}))(\hat{\beta}-\mathbb{E}(\hat{\beta}))^T\Sigma|X\big]\Big\} \\
&= \underbrace{||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2}_{\text{Bias}} + \underbrace{\text{trace}(\text{Cov}(\hat{\beta}|X)\Sigma)}_{\text{Variance}}
\end{aligned}
$$

In equation * we used that expectation and trace operators are exchangeable.

8.2 Proof of Proposition 2.5

Before we derive $\hat{\beta}$ we state the Moore-Penrose conditions for the pseudoinverse of a general matrix M:

$$M^\dagger M M^\dagger = M^\dagger \tag{1}$$

$$M M^\dagger M = M \tag{2}$$

$$(M M^\dagger)^T = M M^\dagger \tag{3}$$

$$(M^\dagger M)^T = M^\dagger M \tag{4}$$

We have:

$$
\begin{aligned}
||y - X\beta||_2^2 &= ||y - XX^\dagger y + XX^\dagger y - X\beta||_2^2 = ||(I - XX^\dagger)y + XX^\dagger y - X\beta||_2^2 \\
&= ||(I - XX^\dagger)y||_2^2 + ||XX^\dagger y - X\beta||_2^2
\end{aligned} \tag{8.1}
$$

The last equation follows because the cross term cancels. Indeed,

$$
\begin{aligned}
(XX^\dagger y - X\beta)^T(I - XX^\dagger) &= (X^\dagger y - \beta)^T X^T(I - XX^\dagger) \\
&= (X^\dagger y - \beta)^T(X^T - X^T XX^\dagger) = 0
\end{aligned}
$$

where in the last inequality we used that $(X^T - X^T X X^\dagger) = 0$ which can be proven by taking the transpose in (2) and then applying (3). From equation (8.1) in order to minimize the problem with respect to β we require $X\hat{\beta} = XX^\dagger y$. Thus $\hat{\beta}$ lives in the range of $X^\dagger$ and all solutions can be written in the form:

$$\hat{\beta} = X^\dagger y + (I - X^\dagger X)v \qquad \text{for some vector } v \in \mathbb{R}^p.$$

where the first term is the range of $X^\dagger$ and the second is the projection on the kernel space of $X^\dagger$. Therefore, the solution with minimum l_2-norm is when we set $v = 0$ and then $\hat{\beta} = X^\dagger y$. It remains to show that $X^\dagger y = (X^T X)^\dagger X^T y$, i.e. $X^\dagger = (X^T X)^\dagger X^T$. In order to do that we will use the fact that the pseudoinverse of a general matrix M exists, is unique and is fully characterized by equations (1)-(4).
Denote $Y = (X^T X)^\dagger X^T$. Then:

$$YXY = \underbrace{(X^T X)^\dagger X^T X (X^T X)^\dagger}_{(X^T X)^\dagger \text{ by (1)}} X^T = (X^T X)^\dagger X^T = Y \qquad \text{verifies equation (1)}$$

$$(XY)^T = (X(X^T X)^\dagger X^T)^T = X(X^T X)^\dagger X^T = XY \qquad \text{verifies equation (3)}$$

$$(YX)^T = ((X^T X)^\dagger X^T X)^T = X^T X (X^T X)^\dagger$$

$$\overset{*}{=} X^T X (X^T X)^\dagger (X^T X)^\dagger X^T X \overset{**}{=} (X^T X)^\dagger X^T X = YX \qquad \text{verifies equation (4)}$$

In the equations which verify (3) and (4) we used that Moore-Penrose pseudoinverse commutes with transposing. In equation * we used that for symmetric matrix $S = X^T X$ we have $S^\dagger = S^\dagger S^\dagger S$ and in equation ** that $S^\dagger = S S^\dagger S^\dagger$. Indeed $S^\dagger = S^\dagger (S S^\dagger)^T = S^\dagger S^\dagger S$, where for the first equality we used equations (1) and (3), and for the second that pseudoinversion commutes with transposing. Deriving $S^\dagger = S S^\dagger S^\dagger$ is identical.
Finally, we verify that eq. (2) $XYX = Y$ holds. We will use the following result for any matrices M and N. If $M^T M N = 0$ then $MN = 0$ (*). Indeed, $M^T M N = 0 \implies N^T M^T M N = 0 \implies (MN)^T MN = 0 \implies MN = 0$. Now left multiply Y by $X^T X$:

$$X^T XY = X^T X (X^T X)^\dagger X^T = X^T X (X^T X)^\dagger X^T X X^\dagger = \underbrace{X^T X}_{X^T X (X^T X)^\dagger X^T X} X^\dagger \tag{8.2}$$

where the second equation uses $X^T = X^T X X^\dagger$ which follows by transposing (2) and then applying (3). Consider subtracting RHS from LHS in (8.2):

$$X^T X (X^T X)^\dagger X^T - X^T X X^\dagger = 0 \implies X^T X((X^T X)^\dagger X^T - X^\dagger) = 0 \overset{(*)}{\implies}$$
$$X((X^T X)^\dagger X^T - X^\dagger) = 0 \implies X(X^T X)^\dagger X^T = XX^\dagger \implies$$
$$X(X^T X)^\dagger X^T X = XX^\dagger X \implies XYX = X, \text{ verifies equation (2)}$$

Therefore, we conclude that the minimum l_2-norm least squares estimator is $\hat{\beta} = X^\dagger y = (X^T X)^\dagger X^T y$.

8.3 Proof of Proposition 2.8

By Proposition 2.4 we have that the test error satisfies:

$$R_X(\hat{\beta}; \beta) = B_X(\hat{\beta}; \beta) + V_X(\hat{\beta}; \beta)$$
$$B_X(\hat{\beta}; \beta) = ||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2 \text{ and } V_X(\hat{\beta}; \beta) = \text{trace}(\text{Cov}(\hat{\beta}|X)\Sigma)$$

Substitute $\hat{\beta} = (X^TX)^\dagger X^Ty$

$$\begin{aligned} B_X(\hat{\beta};\beta) &= ||\mathbb{E}(\hat{\beta}|X) - \beta||_\Sigma^2 = ||(X^TX)^\dagger X^TX\beta - \beta||_\Sigma^2 \\ &= ||(I - (X^TX)^\dagger X^TX)\beta||_\Sigma^2 = ||(I - n^{-1}\hat{\Sigma}^\dagger n\hat{\Sigma})\beta||_\Sigma^2 \\ &= ||\Pi\beta||_\Sigma^2 = \beta^T\Pi^T\Sigma\Pi\beta \\ &= \beta^T\Pi\Sigma\Pi\beta, \text{ as } \Pi^T = I - (\hat{\Sigma}^\dagger\Sigma)^T = I - \hat{\Sigma}^\dagger\Sigma = \Pi, \text{ by Moore-Penrose property in eq.(4).} \end{aligned}$$

For the variance term we have that $V_X(\hat{\beta};\beta) = \text{trace}(\text{Cov}(\hat{\beta}|X)\Sigma)$.

$$\begin{aligned} \text{Cov}(\hat{\beta}|X) &= \text{Cov}((X^TX)^\dagger X^Ty|X) = (X^TX)^\dagger X^T\text{Cov}(y|X)X(X^TX)^\dagger \\ &= (X^TX)^\dagger X^T\sigma^2 IX(X^TX)^\dagger = \sigma^2(X^TX)^\dagger X^TX(X^TX)^\dagger = \frac{\sigma^2}{n}\hat{\Sigma}^\dagger\hat{\Sigma}\hat{\Sigma}^\dagger = \frac{\sigma^2}{n}\hat{\Sigma}^\dagger \end{aligned}$$

The last equation applies Moore-Penrose property in eq. (1). Thus we derive $V_X(\hat{\beta};\beta) = \frac{\sigma^2}{n}\text{trace}(\hat{\Sigma}^\dagger\Sigma)$.

8.4 Proof of Claim in Step 4 in Theorem 4.1

The Stieltjes transform for the Marčenko-Pastur law for $\gamma < 1$ is:

$$m_\gamma(-z) = \int_a^b \frac{1}{x+z}\frac{1}{2\pi x\gamma}\sqrt{(b-x)(x-a)}dx \qquad z \in \mathbb{C}\setminus(0,\infty) \tag{8.3}$$

such that $a = (1-\sqrt{\gamma})^2 = 1 - 2\sqrt{\gamma} + \gamma$ and $b = (1+\sqrt{\gamma})^2 = 1 + 2\sqrt{\gamma} + \gamma$. It is natural to reparametrize $x = 1 + 2\sqrt{\gamma}\cos(\theta) + \gamma$ for $\theta \in [0,\pi]$. Then we have:

$$\begin{aligned} (b-x)(x-a) &= [(1+2\sqrt{\gamma}+\gamma) - (1+2\sqrt{\gamma}\cos(\theta)+\gamma)] \times [(1+2\sqrt{\gamma}\cos(\theta)+\gamma) - (1-2\sqrt{\gamma}+\gamma)] \\ &= 2\sqrt{\gamma}(1-\cos(\theta))2\sqrt{\gamma}(1+\cos(\theta)) = 4\gamma\sin^2(\theta) \end{aligned}$$

After reparametrization for the dx-term we have $dx = -2\sqrt{\gamma}\sin(\theta)d\theta$. Plugging in these two equations and the reparametrized x in (8.3) we obtain:

$$\begin{aligned} m_\gamma(-z) &= \int_0^\pi \frac{2}{\pi}\frac{\sin^2(\theta)}{(1+2\sqrt{\gamma}\cos(\theta)+\gamma)(1+2\sqrt{\gamma}\cos(\theta)+\gamma+z)}d\theta \\ &= \int_0^{2\pi} \frac{1}{\pi}\frac{(\frac{e^{i\theta}-e^{-i\theta}}{2i})^2}{(1+\sqrt{\gamma}(e^{i\theta}+e^{-i\theta})+\gamma)(1+\sqrt{\gamma}(e^{i\theta}+e^{-i\theta})+\gamma+z)}d\theta \\ &= \int_0^{2\pi} -\frac{1}{4\pi}\frac{e^{i\theta}-e^{-i\theta}}{(1+\sqrt{\gamma}(e^{i\theta}+e^{-i\theta})+\gamma)(1+\sqrt{\gamma}(e^{i\theta}+e^{-i\theta})+\gamma+z)}d\theta \\ &\overset{\psi=e^{i\theta}}{=} -\frac{1}{4\pi i}\int_{Circle:|\psi|=1} \frac{(\psi-\psi^{-1})^2}{\psi(1+\sqrt{\gamma}(\psi+\psi^{-1})+\gamma)(1+\sqrt{\gamma}(\psi+\psi^{-1})+\gamma+z)}d\psi \\ &= -\frac{1}{4\pi i}\int_{Circle:|\psi|=1} \frac{(\psi^2-1)^2}{\psi(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma))(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma+z))}d\psi \end{aligned} \tag{8.4}$$

The integrand function has 5 singularities and are all simple poles. The singularities are essentially the roots of the denominator of the integrand and are:

$\psi_0 = 0, \psi_1 = -\sqrt{\gamma}, \psi_2 = -\frac{1}{\sqrt{\gamma}}, \psi_{3,4} = \frac{-(1+\gamma+z) \pm \sqrt{(1+\gamma+z)^2 - 4\gamma}}{2\sqrt{\gamma}}.$

We use Proposition 3.10 to have a quick formula for the computation of the residues at these singularities. Denote the integrand in (8.4) to be equal to $f(\psi)$, its numerator to be $g(\psi)$ and its denominator to be $h(\psi)$. Then:

$$\begin{aligned} res(f(\psi);\psi_0) &= \frac{g(0)}{h'(0)} = \left.\frac{1}{(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma))(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma+z))}\right|_{\psi=0} \\ &= \frac{1}{\gamma} \end{aligned} \tag{8.5}$$

$$\begin{aligned} res(f(\psi);\psi_1) &= \frac{g(\psi_1)}{h'(\psi_1)} = \left.\frac{(\psi^2-1)^2}{\psi(2\sqrt{\gamma}\psi+(1+\gamma))(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma+z))}\right|_{\psi=-\sqrt{\gamma}} \\ &= \frac{(\gamma-1)^2}{-\sqrt{\gamma}(1-\gamma)(-\sqrt{\gamma}z)} = \frac{1-\gamma}{\gamma z} \end{aligned} \tag{8.6}$$

$$res(f(\psi);\psi_2) = \frac{g(\psi_2)}{h'(\psi_2)} = \left.\frac{(\psi^2-1)^2}{\psi(2\sqrt{\gamma}\psi+(1+\gamma))(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma+z))}\right|_{\psi=-\sqrt{\gamma}^{-1}} \tag{8.7}$$

$$= -\frac{1-\gamma}{\gamma z} \tag{8.8}$$

$$res(f(\psi);\psi_3) = \frac{g(\psi_3)}{h'(\psi_3)} = \left.\frac{(\psi^2-1)^2}{\psi(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma))(2\sqrt{\gamma}\psi+(1+\gamma+z))}\right|_{\psi=\psi_3} \tag{8.9}$$

We want to simplify $res(f(\psi);\psi_3)$. Assuming $a = 1+\gamma+z$, we have $\psi_3 = \frac{1}{2\sqrt{\gamma}}\left(-a+\sqrt{a^2-4\gamma}\right)$. We calculate the different components of (8.9) evaluated at $\psi = \psi_3$:

$$\left.(2\sqrt{\gamma}\psi+(1+\gamma+z))\right|_{\psi=\psi_3} = \left(-a+\sqrt{a^2-4\gamma}\right)+a = \sqrt{a^2-4\gamma} \tag{8.10}$$

$$\begin{aligned} \left.(\sqrt{\gamma}(\psi^2+1)+\psi(1+\gamma))\right|_{\psi=\psi_3} &= \sqrt{\gamma}\left(\frac{\left(-a+\sqrt{a^2-4\gamma}\right)^2}{4\gamma}+1\right)+(1+\gamma)\left(\frac{-a+\sqrt{a^2-4\gamma}}{2\sqrt{\gamma}}\right) \\ &= \sqrt{\gamma}\left(\frac{a^2-2a\sqrt{a^2-4\gamma}+a^2-4\gamma+4\gamma}{4\gamma}\right) \\ &+ (1+\gamma)\left(\frac{-a+\sqrt{a^2-4\gamma}}{2\sqrt{\gamma}}\right) \\ &= \sqrt{\gamma}\left(\frac{a^2-a\sqrt{a^2-4\gamma}}{2\gamma}\right)+(1+\gamma)\left(\frac{-a+\sqrt{a^2-4\gamma}}{2\sqrt{\gamma}}\right) \\ &= \frac{a\left(a-\sqrt{a^2-4\gamma}\right)+(1+\gamma)\left(-a+\sqrt{a^2-4\gamma}\right)}{2\sqrt{\gamma}} \\ &= \frac{1}{2\sqrt{\gamma}}\left(a-\sqrt{a^2-4\gamma}\right)z \quad \text{as } a = 1+\gamma+z \\ &= -\psi_3 z \end{aligned} \tag{8.11}$$

$$\begin{aligned}
(\psi^2-1)^2\big|_{\psi=\psi_3} &= \frac{1}{16\gamma^2}((\sqrt{a^2-4\gamma}-a)^2-4\gamma)^2 \\
&= \frac{1}{16\gamma^2}(a^2-4\gamma-2a\sqrt{a^2-4\gamma}+a^2-4\gamma)^2 \\
&= \frac{1}{4\gamma^2}(a^2-a\sqrt{a^2-4\gamma}-4\gamma)^2 \\
&= \frac{1}{4\gamma^2}((a^2-4\gamma)^2-2(a^2-4\gamma)a\sqrt{a^2-4\gamma}+a^2(a-4\gamma)) = \\
&= \frac{1}{4\gamma^2}(a^2-4\gamma)(a^2-4\gamma-2a\sqrt{a^2-4\gamma}+a^2) \\
&= \frac{1}{4\gamma^2}(a^2-4\gamma)\left(a-\sqrt{a^2-4\gamma}\right)^2 = \frac{\psi_3^2(a^2-4\gamma)}{\gamma} \qquad (8.12)
\end{aligned}$$

Now plug (8.10), (8.11), (8.12) in (8.9) to obtain:

$$res(f(\psi);\psi_3) = \frac{\psi_3^2(a^2-4\gamma)}{\gamma\psi_3(-\psi_3 z)\sqrt{a^2-4\gamma}} = -\left(\frac{\sqrt{a^2-4\gamma}}{\gamma z}\right) = -\left(\frac{\sqrt{(1+\gamma+z)^2-4\gamma}}{\gamma z}\right) \qquad (8.13)$$

In a similar fashion to the residue at ψ_3 we can obtain the residue at ψ_4 to be equal to:

$$res(f(\psi);\psi_4) = \frac{\sqrt{(1+\gamma+z)^2-4\gamma}}{\gamma z} \qquad (8.14)$$

By Cauchy's Residue Theorem 3.12 we can evaluate the Stieltjes transform (8.4) $m_\gamma(-z)$ of the Marčenko-Pastur law:

$$\begin{aligned}
m_\gamma(-z) &= -\frac{1}{4\pi i}2\pi i \times (\text{residues which are inside the unit circle}) \\
&= -\frac{1}{2}\left(res(f(\psi);\psi_0)+res(f(\psi);\psi_1)+res(f(\psi);\psi_3)\right) \\
&= -\frac{1}{2}\left(\frac{1}{\gamma}+\frac{1-\gamma}{\gamma z}-\frac{1}{\gamma z}\sqrt{(1+\gamma+z)^2-4\gamma}\right) \\
&= \frac{-(1-\gamma+z)+\sqrt{(1+\gamma+z)^2-4\gamma}}{2\gamma z} \qquad (8.15)
\end{aligned}$$

The above result is true for $\gamma < 1$. If $\gamma > 1$, then the Marčenko-Pastur law has a point mass $\frac{\gamma-1}{\gamma}$ at $x = 0$. Then $m_\gamma(-z)$ is the same as in (8.4) plus $(1-\gamma)/\gamma z$. When $\gamma > 1$ the residues which are inside the circle are ψ_1, ψ_2, ψ_3. Proceed as above to derive the same result.

8.5 Proof of Remark 5.5

We have that as $n, p \to \infty$, almost surely:

$$\beta^T(\hat{\Sigma}+zI)^{-1}\beta \to r^2 m_\gamma(-z) \iff$$
$$\left(\frac{\beta^T}{r}\right)(\hat{\Sigma}+zI)^{-1}\left(\frac{\beta}{r}\right) \to m_\gamma(-z)$$

where in the second line we write $r = ||\beta||_2$.

We can interpret the LHS as an expectation but with respect to the **generalized spectral empirical distribution** $G_{\hat{\Sigma}}(x)$ given in Remark 5.5. Consider the eigenvalue decomposition of $\hat{\Sigma} = UDU^T$ and observe that:

$$\begin{aligned}
\left(\frac{\beta^T}{r}\right)(\hat{\Sigma}+zI)^{-1}\left(\frac{\beta}{r}\right) &= \left(\frac{\beta^T}{r}\right)(UDU^T+zI)^{-1}\left(\frac{\beta}{r}\right)\\
&= \left(\frac{\beta^T}{r}\right)(U(D+zI)U^T)^{-1}\left(\frac{\beta}{r}\right)\\
&= \left(\frac{\beta^T}{r}\right)U^T(D+zI)^{-1}U\left(\frac{\beta}{r}\right)\\
&= \left(\frac{(U\beta)^T}{r}\right)(D+zI)^{-1}\left(\frac{U\beta}{r}\right) \quad \text{belongs to } \mathbb{R}^1\\
&= \text{trace}\left\{\left(\frac{(U\beta)^T}{r}\right)(D+zI)^{-1}\left(\frac{U\beta}{r}\right)\right\}\\
&= \text{trace}\left\{\boldsymbol{w}^T(D+zI)^{-1}\boldsymbol{w}\right\} \quad \boldsymbol{w} \text{ defined in Remark 5.5}\\
&= \sum_{i=1}^{p}\frac{1}{(\lambda_i(\hat{\Sigma})+z)|w_i|^2}\\
&= \int\frac{dG_{\hat{\Sigma}}(x)}{x+z} = m_{G_{\hat{\Sigma}}}(-z)
\end{aligned}$$

which derives $m_{G_{\hat{\Sigma}}} \to m_\gamma$ as $n, p \to \infty$, $p/n \to \gamma > 1$, and concludes our claim in Remark 5.5.

8.6 Proof of Proposition 6.4

We want to compute $B_X(\hat{\beta}_\lambda) = \mathbb{E}_\beta(B_X(\hat{\beta}_\lambda;\beta))$. We would compute the inner term $B_X(\hat{\beta}_\lambda;\beta)$ and then take the expectation with respect to β.

$$\begin{aligned}
B_X(\hat{\beta}_\lambda;\beta) &= ||\mathbb{E}(\hat{\beta}_\lambda|X)-\beta||_\Sigma^2 = ||(X^TX+n\lambda I)^{-1}X^TX\beta-\beta||_\Sigma^2\\
&= ||((X^TX+n\lambda I)^{-1}X^TX-I)\beta||_\Sigma^2\\
&= ||\left((X^TX+n\lambda I)^{-1}(X^TX+\lambda nI)-(X^TX+n\lambda I)^{-1}\lambda nI-I\right)\beta||_\Sigma^2\\
&= ||-\lambda n(X^TX+n\lambda I)^{-1}\beta||_\Sigma^2\\
&= (\lambda n)^2\beta^T(X^TX+n\lambda I)^{-1}\Sigma(X^TX+n\lambda I)^{-1}\beta \quad \text{belongs to } \mathbb{R}^1\\
&= (\lambda n)^2\text{trace}\left(\beta^T(X^TX+n\lambda I)^{-1}\Sigma(X^TX+n\lambda I)^{-1}\beta\right)\\
&= (\lambda n)^2\text{trace}\left(\beta\beta^T(X^TX+n\lambda I)^{-1}\Sigma(X^TX+n\lambda I)^{-1}\right)
\end{aligned}$$

$$\begin{aligned}
B_X(\hat{\beta}_\lambda) &= \mathbb{E}_\beta(B_X(\hat{\beta}_\lambda;\beta))\\
&= \mathbb{E}_\beta\left\{(\lambda n)^2\text{trace}\left(\beta\beta^T(X^TX+n\lambda I)^{-1}\Sigma(X^TX+n\lambda I)^{-1}\right)\right\}\\
&= (\lambda n)^2\text{trace}\left(\mathbb{E}_\beta(\beta\beta^T)(X^TX+n\lambda I)^{-1}\Sigma(X^TX+n\lambda I)^{-1}\right)\\
&= (\lambda n)^2\frac{r^2}{p}\text{trace}\left(\Sigma(X^TX+n\lambda I)^{-2}\right) \qquad \text{by trace trick and } \mathbb{E}_\beta(\beta\beta^T) = \frac{r^2}{p}
\end{aligned}$$

Next we derive that $V_X(\hat{\beta}_\lambda) = \sigma^2 \text{trace}\left(\Sigma(X^TX + n\lambda I)^{-1}X^TX(X^TX + n\lambda I)^{-1}\right)$. Firstly, note that $V_X(\hat{\beta}_\lambda; \beta)$ does not depend on β and we have:

$$V_X(\hat{\beta}_\lambda) = \mathbb{E}_\beta(V_X(\hat{\beta}_\lambda; \beta)) = \mathbb{E}_\beta(\text{trace}(\text{Cov}(\hat{\beta}_\lambda|X)\Sigma)) = \text{trace}(\text{Cov}(\hat{\beta}_\lambda|X)\Sigma)$$

Further, we can rewrite the covariance in the following way:

$$\begin{aligned}\text{Cov}(\hat{\beta}_\lambda|X) &= \text{Cov}\left((X^TX + \lambda nI)^{-1}X^Ty|X\right)\\ &= (X^TX + n\lambda I)^{-1}X^T\text{Cov}(y|X)X(X^TX + n\lambda I)^{-1}\\ &= \sigma^2(X^TX + n\lambda I)^{-1}X^TX(X^TX + n\lambda I)^{-1}\end{aligned}$$

where in the third line we used:

$$\text{Cov}(y|X) = \mathbb{E}\left([X\beta + \epsilon - \mathbb{E}(X\beta + \epsilon)][X\beta + \epsilon - \mathbb{E}(X\beta + \epsilon)]^T \,\middle|\, X\right) = \mathbb{E}(\epsilon\epsilon^T) = \sigma^2 I$$

Then we can obtain the desired result:

$$\begin{aligned}V_X(\hat{\beta}_\lambda) &= \text{trace}\left(\sigma^2(X^TX + n\lambda I)^{-1}X^TX(X^TX + n\lambda I)^{-1}\Sigma\right)\\ &= \sigma^2\text{trace}\left(\Sigma(X^TX + n\lambda I)^{-1}X^TX(X^TX + n\lambda I)^{-1}\right)\end{aligned}$$

References

[1] Hastie et al. ”Surprises in high-dimensional ridgeless least squares interpolation.” arXiv preprint arXiv:1903.08560 (2019).

[2] Mayer, Jürgen, Khaled Khairy, and Jonathon Howard. ”Drawing an elephant with four complex parameters.” American Journal of Physics 78.6 (2010): 648-649.

[3] Ruder, Sebastian. ”An overview of gradient descent optimization algorithms.” arXiv preprint arXiv:1609.04747 (2016).

[4] Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro. ”In search of the real inductive bias: On the role of implicit regularization in deep learning.” arXiv preprint arXiv:1412.6614 (2014).

[5] Szegedy, Christian, et al. ”Going deeper with convolutions.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[6] Hinton, Geoffrey E., Alex Krizhevsky, and Ilya Sutskever. ”Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems 25 (2012): 1106-1114.

[7] Zhang, Chiyuan, et al. ”Understanding deep learning requires rethinking generalization.” arXiv preprint arXiv:1611.03530 (2016).

[8] Belkin, Mikhail, et al. ”Reconciling modern machine learning and the bias-variance trade-off.” arXiv preprint arXiv:1812.11118 (2018).

[9] Marčenko, Vladimir A., and Leonid Andreevich Pastur. ”Distribution of eigenvalues for some sets of random matrices.” Mathematics of the USSR-Sbornik 1.4 (1967): 457.

[10] Vadim I. Serdobolskii.Multiparametric Statistics. Elsevier, 2007: p.254-p.284

[11] Teschl, Gerald. ”Mathematical methods in quantum mechanics.” Graduate Studies in Mathematics 99 (2009): 106.

[12] Bai, Zhi-Dong, and Yong-Qua Yin. ”Limit of the smallest eigenvalue of a large dimensional sample covariance matrix.” Advances In Statistics. 2008. 108-127.

[13] Billingsley, Patrick. Convergence of probability measures. John Wiley & Sons, 2013.

[14] Brown, James Ward, and Ruel Vance Churchill. Complex variables and applications. Boston: McGraw-Hill Higher Education,, 2009.

[15] Rubio, Francisco, and Xavier Mestre. ”Spectral convergence for a general class of random matrices.” Statistics probability letters 81.5 (2011): 592-602.

[16] Dobriban, Edgar, and Stefan Wager. ”High-dimensional asymptotics of prediction: Ridge regression and classification.” The Annals of Statistics 46.1 (2018): 247-279.

[17] Silverstein, Jack W. "The Stieltjes transform and its role in eigenvalue behavior of large dimensional random matrices." Random Matrix Theory and Its Applications. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap 18 (2009): 1-25.

[18] Ledoit, Olivier, and Sandrine Péché. "Eigenvectors of some large sample covariance matrix ensembles." Probability Theory and Related Fields 151.1-2 (2011): 233-264.

[19] Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." arXiv preprint arXiv:1803.03635 (2018).

[20] Nakkiran, Preetum, et al. "Deep double descent: Where bigger models and more data hurt." arXiv preprint arXiv:1912.02292 (2019).