Lecture Notes on High Dimensional Linear Regression

Alberto Quaini[1,2]

December 23, 2024 (first version: October 17, 2024)

[1]Department of Econometrics, Erasmus University of Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.

[2]Please, let me know if you find typos or mistakes at quaini@ese.eur.nl.

Introduction

These lecture notes were developed for a Master's course in *advanced machine learning* at Erasmus University of Rotterdam. The course is designed for graduate students in mathematics, statistics and econometrics. The content follows a proposition-proof structure, making it suitable for students seeking a formal and rigorous understanding of the statistical theory underlying machine learning methods.

At present, the notes focus on linear regression, with an in-depth exploration of the existence, uniqueness, relations, computation, and non-asymptotic properties of the most prominent estimators in this setting: least squares, ridgeless, ridge, and lasso.

Background

It is assumed that readers have a solid background in calculus, linear algebra, convex analysis, and probability theory. Some definitions and results from these fields, relevant for the course, are provided in the Appendix for reference.

Book-length references

The content of these lecture notes is inspired by a wide range of existing literature, but the presentation of topics follows my own interpretation and logical structure. Although most of the content can be traced back to established sources, certain sections reflect my perspective, and some material is original to this course. For those interested in more comprehensive, book-length discussions of related topics, the following key references are recommended: Hastie et al. [2009], Bühlmann and Van De Geer [2011], Hastie et al. [2015], and Wainwright [2019].

Disclaimer

Please note that despite my efforts, these lecture notes may contain errors. I welcome any feedback, corrections, or suggestions you may have. If you spot any mistakes or have ideas for improvement, feel free to contact me via email at `quaini@ese.eur.nl`.

Contents

Notation

- All random variables are defined on complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in a real Euclidean space.
- For a random variable (vector) [matrix] x ($\boldsymbol{x}$) [$\boldsymbol{X}$], the notation $x \in \mathbb{R}$ ($\boldsymbol{x} \in \mathbb{R}^n$) [$\boldsymbol{X} \in \mathbb{R}^{n \times p}$] means that x ($\boldsymbol{x}$) [$\boldsymbol{X}$] takes values in $\mathbb{R}$ ($\mathbb{R}^n$) [$\mathbb{R}^{n \times p}$].
- The symbols $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{d}$ denote convergence in probability and in distribution, respectively.
- Given a random variable x, its expectation is denoted $\mathbb{E}[x]$ and its variance $\mathbb{V}\text{ar}[x]$.
- For a vector $\boldsymbol{x} \in \mathbb{R}^n$, the $i-$th element is denoted x_i for $i = 1, \ldots, n$.
- For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, the $i, j-$th element is denoted $A_{i,j}$, the $j-$th column is denoted $\boldsymbol{A}_j$ and the $i-$th row is denoted $A_{(i)}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- The transpose of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ is denoted $\boldsymbol{A}'$.
- The Moore-Penrose inverse of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ is denoted $\boldsymbol{A}^+$.
- The l_p-norm $\|\cdot\|_p$ on $\mathbb{R}^n$ is defined for all $\boldsymbol{v} \in \mathbb{R}^n$ as $\|\boldsymbol{v}\|_p := \left(\sum_{i=1}^n |v_i|^p\right)^{1/p}$ when $p \in [1, +\infty)$, and $\|\boldsymbol{v}\|_p := \max_{i=1}^n |v_i|$ when $p = +\infty$.
- Given vector $\boldsymbol{\theta} \in \mathbb{R}^p$, the l_0-norm (which is not a norm!) $\|\boldsymbol{x}\|_0$ counts the number of nonzero elements of $\boldsymbol{\theta}$.
- $\text{argmin}_{x \in X} f(x)$ denotes the set of minimizers of f over set X.
- $\text{diag}(\boldsymbol{A})$ denotes the diagonal elements of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$
- $\text{diag}(a_1, \ldots, a_n)$ denotes a square matrix in $\mathbb{R}^{n \times n}$ that has diagonal elements given by $a_1, \ldots, a_n \in \mathbb{R}$ and that has zero elsewhere.
- Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, its rank is $\text{Rank}(\boldsymbol{A})$, its range is $\text{Range}(\boldsymbol{A})$, its kernel is $\text{Ker}(\boldsymbol{A})$, its trace is denoted $\text{Trace}(\boldsymbol{A})$.
- We denote P_S the orthogonal projection onto set $S \subset \mathbb{R}^{n \times p}$.
- Given a vector space V, the sum of two subsets $A, B \subset V$ is defined as $A + B := \{a + b \ : \ a \in A, b \in B\}$. The sum of a set $A \in V$ and a vector $b \in V$ is defined as $A + b := \{a + b \ : \ a \in A\}$.

- The symbol ∂ indicates the subdifferential.

Chapter 1

Linear Regression

Linear regression is a supervised learning technique aimed at predicting a *target* random variable y using a linear combination

$$\boldsymbol{x}'\boldsymbol{\theta} = x_1\theta_1 + \ldots + x_p\theta_p$$

of *explanatory variables* $\boldsymbol{x} = [x_1, \ldots, x_p]'$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $p \in \mathbb{N}$.[1] The target variable y is also referred to as *dependent* or *output* variable, while the explanatory variables $\boldsymbol{x}$ are also known as *independent variables*, *predictors* or *input* variables.

In typical applications, we observe only a sample of size $n \in \mathbb{N}$ of these random variables, represented by the pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for each i. Given a regression coefficient $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, a *statistical linear model*, or simply linear model, is expressed as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\theta}_0 + \varepsilon_{0i}, \quad i = 1, \ldots, n, \tag{1.1}$$

where ε_{0i} are real-valued *residual* random variables. Figure 1.1 depicts a linear model for $i = 1, \ldots, n$ observations y_i, with two predictors $\tilde{\boldsymbol{x}}_i = [1, x_i]'$ consisting of a unit constant and a variable x_i, a coefficient $\boldsymbol{\theta}_0 \in \mathbb{R}^2$ and the associated error terms ε_{0i}. The *Data Generating Process* (DGP), i.e., the joint distribution of the predictors $\boldsymbol{x}$ and the real-valued *residual* random variables $\boldsymbol{\varepsilon}_0 = [\varepsilon_{0i}]_{i=1}^n$, is subject to certain restrictions. Depending on the type of restrictions imposed on the DGP, different types of linear models are obtained. The two general forms of linear models are fixed and random design models, which are defined as follows.

Definition 1 (Fixed design model). In a *fixed design model*, the sequence $(\boldsymbol{x}_i)_{i=1}^n$ is fixed. The residuals ε_{0i} are independent and identically distributed.

[1] An intercept in $f(\boldsymbol{x}; \boldsymbol{\theta})$ can be introduced by adding a constant term to the predictors.

Definition 2 (Random design model). In a *random design model*, the pair $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a sequence of independent and identically distributed random variables.

The fixed design model is particularly suitable when the predictors are controlled by the analyst, such as the dose of medication administered to patients in the treatment group in a clinical trial. Conversely, the random design model is appropriate when the explanatory variables are stochastic, such as the wind speed observed at a specific time and location.

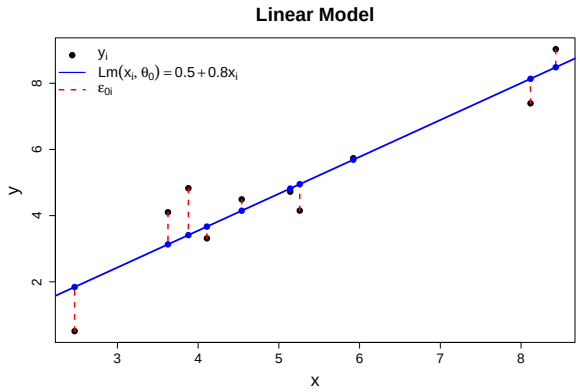

Figure 1.1: Statistical linear model $y_i = \tilde{\boldsymbol{x}}_i'\boldsymbol{\theta}_0 + \varepsilon_{0i}$ where $\tilde{\boldsymbol{x}}_i = [1, x_i]'$ and $\boldsymbol{\theta}_0 = [0.5, 0.8]'$.

We organize the observed values of the target variable in the vector $\boldsymbol{y} = [y_1, \ldots, y_n]' \in \mathbb{R}^n$, and the observations on the predictors in the *design matrix*

$$\boldsymbol{X} := \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix} \mathbb{R}^{n \times p}.$$

With this notation, the linear model (1.1) can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_0,$$

where $\boldsymbol{\varepsilon}_0 = [\varepsilon_{01}, \ldots, \varepsilon_{0n}]' \in \mathbb{R}^n$.

Example 1. A classic example of linear regression is found in the work of Nerlove et al. [1961], which examines returns to scale in the U.S. electricity power supply industry. In this study, the total cost y_i for firm i is predicted using a linear model based on the firm's output production x_{i1}, the wage rate x_{i2}, the price of fuel x_{i3}, and the rental price of capital x_{i4}, with data from $n = 145$ electric utility companies.

The rest of the chapter is organized as follows. First, we study the most basic linear regression approach, the method of least squares projection, and some of its recent machine learning extensions. Our study focuses on their existence, uniqueness, connections, geometric interpretation, and computation. Then, we study we cover both their finite- or small-sample properties, that are valid for any given sample size, and their asymptotic properties, that are useful approximations when the sample size is large enough.

1.0.1 Estimators and their properties

Definition 3 (Estimand). An *estimand* is a feature, or parameter, of interest of the population.

Definition 4 (estimator and estimate). An *estimator* is a function taking as input the data, and possibly other auxiliary variables, and outputting an *estimate*, which is a specific value assigned to the estimand.

For instance, in the context of the linear model (1.1), the coefficient $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ represents the estimand. An estimator is a function $\hat{\boldsymbol{\theta}}_n : \mathbb{R}^n \times \mathbb{R}^{n\times p} \to \mathbb{R}^p$ that takes as inputs the data $(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^n \times \mathbb{R}^{n\times p}$ and produces an estimate $\hat{\boldsymbol{\theta}}_n(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^p$. For simplicity, we use the same notation, $\hat{\boldsymbol{\theta}}_n$, to refer to both the estimator and the resulting estimate, although formally the estimate should be written as $\hat{\boldsymbol{\theta}}_n(\boldsymbol{y}, \boldsymbol{X})$. We may also write $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^p$ to indicate that an estimator outputs values in $\mathbb{R}^p$.

Definition 5 (Linear prediction). The quantity $\text{lm}(\boldsymbol{X}, \boldsymbol{\theta}) := \boldsymbol{X}\boldsymbol{\theta}$ denotes the *linear prediction* associated to coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^p$ and predictors $\boldsymbol{X} \in \mathbb{R}^{n\times p}$.

Let $\boldsymbol{M}^+$ denote the Moore-Penrose inverse of a generic real-valued matrix $\boldsymbol{M}$. We make extensive use of the following useful projections.[2]

Definition 6 (Useful projections). Given a fixed matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$:

- $\mathbf{P}_{\text{Range}(\boldsymbol{X}')} := \boldsymbol{X}^+\boldsymbol{X}$ is the orthogonal projector onto $\text{Range}(\boldsymbol{X}')$;
- $\mathbf{P}_{\text{Ker}(\boldsymbol{X})} := \boldsymbol{I} - \boldsymbol{X}^+\boldsymbol{X}$ is the orthogonal projector onto $\text{Ker}(\boldsymbol{X})$;
- $\mathbf{P}_{\text{Range}(\boldsymbol{X})} := \boldsymbol{X}\boldsymbol{X}^+$ is the orthogonal projector onto $\text{Range}(\boldsymbol{X})$;
- $\mathbf{P}_{\text{Ker}(\boldsymbol{X}')} := \boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^+$ is the orthogonal projector onto $\text{Ker}(\boldsymbol{X}')$.

[2] We use the notation Range (Ker) for the range (kernel) of a matrix. Details on these sets, the Moore-Penrose inverse and orthogonal projections are given in Appendix Section 2.1.

The next proposition demonstrates that, if we fix the design matrix $\boldsymbol{X}$, we can focus on regression coefficients in Range($\boldsymbol{X}'$). Indeed, coefficients in this set span all possible linear predictions that can be achieved through $\boldsymbol{X}$.

Proposition 1.0.1. *Given a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, for any $\boldsymbol{\theta} \in \mathbb{R}^p$:*

$$\text{lm}(\boldsymbol{X}, \boldsymbol{\theta}) = \text{lm}(\boldsymbol{X}, \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\, \boldsymbol{\theta}).$$

Proof. Using the identity $\boldsymbol{I} = \mathbf{P}_{\text{Range}(\boldsymbol{X}')} + \mathbf{P}_{\text{Ker}(\boldsymbol{X})}$, we have for any $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}(\mathbf{P}_{\text{Range}(\boldsymbol{X}')} + \mathbf{P}_{\text{Ker}(\boldsymbol{X})})\boldsymbol{\theta} = \boldsymbol{X}\, \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\, \boldsymbol{\theta}.$$

□

Finite-sample properties

Since an estimator is derived from data, it is a random variable. Intuitively, when comparing two estimators of the same estimand, we prefer the one whose probability distribution is "more concentrated" around the true value of the estimand. Formally, estimators are compared using several key properties.[3]

Definition 7 (Bias). The *bias* of an estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is the difference between the expected value of the estimator and the estimand:

$$\text{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) := \mathbb{E}[\hat{\boldsymbol{\theta}}_n] - \boldsymbol{\theta}_0.$$

Definition 8 (Estimation risk). The *estimation risk* of an estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ measures the difference between the estimate and the estimand as:

$$\text{ER}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) := \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2.$$

Definition 9 (MSE). The *Mean Squared Error* (MSE) of an estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is the expected estimation risk of the estimator:

$$\text{MSE}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) := \mathbb{E}[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2].$$

Definition 10 (Predictive risk). The *predictive risk* of an estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ measures the difference between the linear predictions of $\hat{\boldsymbol{\theta}}_n$ and those of $\boldsymbol{\theta}_0$:

$$\text{PR}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) := \|\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n) - \text{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0)\|_2^2/n.$$

[3]Some of these properties are defined by means of the l_2-norm. Note that this choice is typical, but arbitrary.

Definition 11 (Mean predictive risk). The *Mean Predictive Risk* (MPR) of an estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is the expected predictive risk of the estimator:

$$\mathrm{MPR}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) := \mathbb{E}[\|\mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n) - \mathrm{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0)\|_2^2/n].$$

As a corollary to Proposition 1.0.1, the predictive risk of an estimator is unchanged if both the estimator and the estimand are projected onto $\mathrm{Range}(\boldsymbol{X}')$.

Corollary 1.0.1. *Given a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, for any estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0 \in \mathbb{R}^p$:*

$$\mathrm{PR}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = \mathrm{PR}(\mathbf{P}_{\mathrm{Range}(\boldsymbol{X}')}\, \hat{\boldsymbol{\theta}}_n, \mathbf{P}_{\mathrm{Range}(\boldsymbol{X}')}\, \boldsymbol{\theta}_0).$$

Proof. The result follows from Proposition 1.0.1. □

The next proposition justifies the definition of mean predictive risk given in Definition 11.

Proposition 1.0.2. *Assume that the linear model (1.1) holds with $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \mathbf{0}$. Then, for any $\boldsymbol{\theta} \in \mathbb{R}^p$:*

$$\mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2/n] = \mathrm{MPR}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \mathbb{E}[\|\boldsymbol{\varepsilon}_0\|_2^2/n].$$

Proof. Since $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_0$, we have

$$\mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2/n] = \mathbb{E}[\|\boldsymbol{X}(\boldsymbol{\theta}_0 - \boldsymbol{\theta})\|_2^2/n] + \mathbb{E}[\|\boldsymbol{\varepsilon}_0\|_2^2/n] + 2(\boldsymbol{\theta}_0 - \boldsymbol{\theta})\mathbb{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}_0].$$

Then, the result follows since $\mathbb{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}_0] = \sum_{i=1}^n \mathbb{E}[\boldsymbol{X}_i\varepsilon_{0i}] = \mathbf{0}$, where $\boldsymbol{X}_i$ denotes the $i-$th row of $\boldsymbol{X}$. □

If our primary goal is to accurately predict the target variable, we seek a estimators $\hat{\boldsymbol{\theta}}$ with a low mean prediction risk $\mathbb{E}[\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\|_2^2/n]$. Since we cannot control the error term ε_0, Proposition 1.0.2 suggests that we should focus on estimators with a low mean predictive risk.

On the other hand, if our interest lies in understanding which predictors influence the target variable and how they do so, the true coefficient $\boldsymbol{\theta}_0$ becomes our focus. In this case, we might prefer unbiased estimators – those with zero bias – over biased ones. However, estimators with lower mean squared error (MSE) are generally favored, even if they feature some bias. The following proposition demonstrates that the MSE can be decomposed into a bias and a variance term.

Proposition 1.0.3 (Bias-variance decomposition of MSE)**.** *Given an estimator $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^p$ for $\boldsymbol{\theta}_0 \in \mathbb{R}^p$, the MSE can be decomposed as follows:*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = \|\mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\|_2^2 + \mathrm{Trace}(\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n]).$$

Proof. The result follows from

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) &= \mathbb{E}[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)'(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] \\
&= \mathbb{E}[\mathrm{Trace}\{(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)'\}] \\
&= \mathrm{Trace}(\mathbb{E}[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)']) \\
&= \mathrm{Trace}(\mathbb{E}[(\hat{\boldsymbol{\theta}}_n - \mathbb{E}[\hat{\boldsymbol{\theta}}_n] + \mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0))(\hat{\boldsymbol{\theta}}_n - \mathbb{E}[\hat{\boldsymbol{\theta}}_n] + \mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0))']) \\
&= \mathrm{Trace}(\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n] + \mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\,\mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)' + \\
&\qquad \mathbb{E}[\hat{\boldsymbol{\theta}}_n - \mathbb{E}[\hat{\boldsymbol{\theta}}_n]]\,\mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)' + \mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\mathbb{E}[\hat{\boldsymbol{\theta}}_n - \mathbb{E}[\hat{\boldsymbol{\theta}}_n]]') \\
&= \mathrm{Trace}(\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n] + \mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\,\mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)') \\
&= \|\mathrm{Bias}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0)\|_2^2 + \mathrm{Trace}(\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n]).
\end{aligned}$$

□

Loosely speaking, the bias and the variance of an estimator are linked to the estimator's "complexity". Estimators with higher complexity often fit the data better, resulting in lower bias, but they are more sensitive to data variations, leading to higher variance. Conversely, estimators with lower complexity tend to have lower variance but higher bias, a phenomenon known as the *bias-variance tradeoff*.

Apart from simple cases, computing the finite-sample properties of estimators, such as their MSE or predictive risk, is infeasible or overly complicated. This is because they require computations under the DGP of complex transformations of the data. When direct computation is not possible, we can rely on *concentration inequalities* or *asymptotic approximations*.

Concentration inequalities are inequalities that bound the probability that a random variable deviates from a particular value, typically its expectation. In this chapter, we focus on inequalities that control the MSE or predictive risk of an estimator, such as:

$$\mathbb{P}[d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) \leq h(\boldsymbol{y}, \boldsymbol{X}, n, p)] \geq 1 - \delta,$$

or

$$\mathbb{P}[d(\mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n), \mathrm{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0)) \leq h(\boldsymbol{y}, \boldsymbol{X}, n, p)] \geq 1 - \delta,$$

where $\delta \in (0, 1)$ is the *level of confidence*, $d : \mathbb{R}^p \times \mathbb{R}^p \to [0, +\infty)$ is a distance, and h is a real-valued function of the data, the sample size, and the number of predictors.

Large-sample properties

Large-sample or asymptotic theory provides an alternative approach to study and analyse estimators. Classically, this framework develops approximations of the finite-sample properties of estimators, such as their distribution, MSE or predictive risk, by letting the sample size $n \to \infty$. Consequently, these approximations work well when the sample size n is much larger than the number of predictors p. More recently, asymptotic approximations are also developed by letting $p \to \infty$, or having both $n, p \to \infty$ at some rate. Note that, given a sample of size n and number of variables p, there is no general indication on how to choose the appropriate asymptotic regime for n and p, as the goodness of fit of the corresponding asymptotic approximations should be assessed on a case by case basis. In this chapter, we work with two notions from large-sample theory: consistency and asymptotic distribution.

Definition 12 (Consistency). Estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is *consistent*, written $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$ as $n \to \infty$, if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| > \varepsilon] = 0.$$

Definition 13 (Asymptotic distribution). Given a deterministic real-valued sequence $r_{n,p} \to \infty$, let $F_{n,p}$ be the probability distribution of $r_{n,p}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and F a non-degenerate probability distribution. Estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ has *asymptotic distribution* F with rate of convergence $r_{n,p}$ if $F_{n,p}(\boldsymbol{z}) \to F(\boldsymbol{z})$ as $r_{n,p} \to \infty$ for all $\boldsymbol{z}$ at which $F(\boldsymbol{z})$ is continuous. Equivalent short-hand notations are $r_{n,p}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} F$ and $r_{n,p}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \boldsymbol{\eta} \sim F$, as $r_{n,p} \to \infty$.

1.1 Least Squares and Penalized Least Squares

In this chapter, we study the most widely used methods in linear regression analysis: the method of least squares and some of its penalized variants. The method of least squares was first introduced by Legendre [1805] and Gauss [1809], and it consists in minimizing the squared l_2-distance between the target values $\boldsymbol{y}$ and the linear prediction $\text{lm}(\boldsymbol{X}, \boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\theta}$ in the coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^p$.

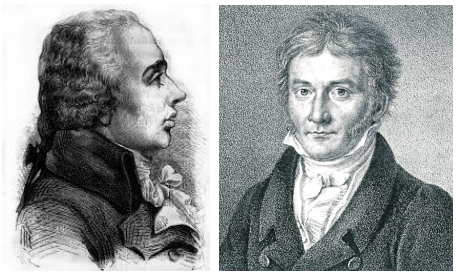

Figure 1.2: Adrien-Marie Legendre (1752–1833) and Johann Carl Friedrich Gauss (1977–1855).

Definition 14 (Least squares estimator). The *Least Squares Estimator* (LSE) is defined as:

$$\hat{\boldsymbol{\theta}}_n^{ls} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}'\boldsymbol{x}_i)^2 = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2. \tag{1.2}$$

In addition to the LSE, we consider the following variants.

Definition 15 (Ridgeless estimator). The *ridgeless* estimator is defined as:

$$\hat{\boldsymbol{\theta}}_n^{rl} = \operatorname*{argmin}_{\hat{\boldsymbol{\theta}}\in\mathbb{R}^p} \left\{ \|\hat{\boldsymbol{\theta}}\|_2^2 \;:\; \hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 \right\}. \tag{1.3}$$

Definition 16 (Ridge estimator). The *ridge estimator* is defined for $\lambda > 0$ as:

$$\hat{\boldsymbol{\theta}}_n^{r}(\lambda) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|_2^2. \tag{1.4}$$

Definition 17 (Lasso estimator). The *lasso estimator* is defined for $\lambda > 0$ as:

$$\hat{\boldsymbol{\theta}}_n^{l}(\lambda) \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 + \lambda\left\|\boldsymbol{\theta}\right\|_1. \tag{1.5}$$

Here is a brief overview of the results that are discussed in detail in the rest of this chapter. A solution to the least squares problem (1.2) always exists. However, when the predictors (i.e., the columns of $\boldsymbol{X}$) are linearly dependent, there are infinitely many solutions.[4] In such cases, the LSE typically considered is the ridgeless estimator, which is always unique.

The ridge and lasso estimators are penalized or regularized versions of the LSE, with penalty term $\lambda\left\|\boldsymbol{\theta}\right\|_2^2$ and $\lambda\left\|\boldsymbol{\theta}\right\|_1$, respectively. The penalty parameter $\lambda > 0$ controls the strength of the penalty. The ridge estimator,

[4]This situation always arises when $p > n$, and it may arise even when $n \leq p$.

introduced by Hoerl and Kennard [1970], was developed to address certain shortcomings of the LSE, particularly in scenarios involving collinear or multicollinear designs – where the predictors in $\boldsymbol{X}$ are linearly dependent or nearly-linearly dependent. The ridge estimator is uniquely defined and often exhibits better statistical properties compared to those of the LSE in settings with multicollinear or many predictors. On the other hand, the lasso estimator, popularized by Tibshirani [1996], offers an approximation of the l_0 estimator, which is defined for some $R > 0$:

$$\hat{\boldsymbol{\theta}}_n^{l_0}(\lambda) \in \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \ : \ \|\boldsymbol{\theta}\|_0 \leq R \right\}, \tag{1.6}$$

where $\|\boldsymbol{\theta}\|_0$ is the number of nonzero elements in $\boldsymbol{\theta}$. A key feature of this estimator is its ability to produce sparse solutions, i.e., to set some coefficients exactly to zero. Consequently, the l_0 estimator can be used to perform parameter estimation and variable selection simultaneously. However, it is the solution of a non-convex problem, and, in general, computing it can be an "NP-hard" problem. The lasso instead shares the ability to produce sparse solutions and it can be easily computed even for large datasets.

Remark 1 (Data standardization). For computational stability, it is recommended to compute linear regression estimators with a least squares loss after having standardized the predictors $\boldsymbol{X}$ so that $\bar{\boldsymbol{x}} := \boldsymbol{X}'\mathbf{1}/n = \mathbf{0}$ and $\boldsymbol{X}_j'\boldsymbol{X}_j = 1$ for each $j = 1\ldots, p$. Without standardization, the solutions would depend on the units used to measure the predictors. Moreover, we may also center the target variable $\boldsymbol{y}$, meaning $\bar{y} := \boldsymbol{y}'\mathbf{1}/n = 0$. These centering conditions are convenient, since they mean that we can omit the intercept term. Given an optimal solution $\hat{\boldsymbol{\theta}}$ on the centered data, we can recover the optimal solutions for the uncentered data: $\hat{\boldsymbol{\theta}}$ is the same and the intercept is given by $\bar{y} - \bar{\boldsymbol{x}}'\hat{\boldsymbol{\theta}}$.

1.1.1 Existence and uniqueness

From here on, we make extensive use of the spectral decomposition of $\boldsymbol{X}$.

Definition 18 (Spectral decomposition of $\boldsymbol{X}$). The spectral decomposition of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}',$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices and, for $r := \operatorname{Rank}(\boldsymbol{X}) \leq \min\{n, p\}$,

$$\boldsymbol{S} = \begin{bmatrix} \operatorname{diag}(s_1, \ldots, s_r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n\times p}.$$

We establish the following key results: the existence of the LSE, the ridgeless, the ridge and the lasso estimators; the closed-form expression of the LSE, ridgeless, and ridge; the uniqueness of the ridgeless and ridge; the uniqueness of the LSE when $\text{Rank}(\boldsymbol{X}) = p$, i.e., when the predictors in $\boldsymbol{X}$ are linearly independent. Notice that this rank condition cannot hold if $n < p$.

Theorem 1.1.1 (Existence and uniqueness of LSE, ridgeless, ridge and lasso)**.** *The following statements hold:*

(i) The set of solutions to the least squares problem (1.2) is non-empty and given by

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}} \, \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 / 2 = \boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X}).$$

(ii) The ridgeless estimator exists, is an element of $\text{Range}(\boldsymbol{X}')$*, and is uniquely given in closed form by*

$$\hat{\boldsymbol{\theta}}_n^{rl} = \boldsymbol{X}^+\boldsymbol{y}. \tag{1.7}$$

(iii) If $\text{Rank}(\boldsymbol{X}) = p$*, then the LSE and the ridgeless estimator are uniquely given in closed form by:*

$$\hat{\boldsymbol{\theta}}_n^{ls} = \hat{\boldsymbol{\theta}}_n^{rl} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1.8}$$

(iv) The ridge estimator with penalty parameter $\lambda > 0$ *exists, is an element of* $\text{Range}(\boldsymbol{X}')$*, and is uniquely given in closed form by:*

$$\hat{\boldsymbol{\theta}}_n^{r}(\lambda) = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1.9}$$

(v) The lasso estimator exists and, in general, it is not unique.

Proof. (i) The least squares problem (1.2) is an unconstrained optimization problem with the objective function

$$f : \mathbb{R}^p \to [0, +\infty); \ \boldsymbol{\theta} \mapsto \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 / 2.$$

By Theorem 2.2.1, the set of least squares minimizers is given by

$$S := \{\hat{\boldsymbol{\theta}} \in \mathbb{R}^p : \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{y}\}.$$

Since $\boldsymbol{X}'\boldsymbol{y} \in \text{Range}(\boldsymbol{X}')$ and $\text{Range}(\boldsymbol{X}') = \text{Range}(\boldsymbol{X}'\boldsymbol{X})$, set S is not empty. Consider a vector $\hat{\boldsymbol{\theta}} \in \boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X})$. Using that $\boldsymbol{X} = \mathbf{P}_{\text{Range}(\boldsymbol{X})}\boldsymbol{X}$ which implies $\boldsymbol{X}' = \boldsymbol{X}'\mathbf{P}_{\text{Range}(\boldsymbol{X})}$, we obtain

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y} = \boldsymbol{X}'\mathbf{P}_{\text{Range}(\boldsymbol{X})}\,\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y}.$$

Therefore, $\boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X}) \subset S$. Now consider a vector $\boldsymbol{v} \in \mathbb{R}^p$ not in set $\boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X})$. That is, $\boldsymbol{v} = \hat{\boldsymbol{\theta}} + \boldsymbol{u}$ with $\hat{\boldsymbol{\theta}} \in \boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X})$ and $\boldsymbol{u} \in \text{Range}(\boldsymbol{X}')$. Since $\boldsymbol{X}\boldsymbol{u} \neq \boldsymbol{0}$,

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{v} = \boldsymbol{X}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{u} \neq \boldsymbol{X}\boldsymbol{y}.$$

We conclude that $\boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X}) = S$.

(ii) The minimum norm least squares problem in (1.3) has a strictly convex and coercive objective function

$$f : \mathbb{R}^p \to \mathbb{R}; \boldsymbol{\theta} \mapsto \|\boldsymbol{\theta}\|_2^2,$$

and a closed convex feasible set $\boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X}) \subset \mathbb{R}^p$. It follows that a solution exists and it is unique; see Propositions 2.2.3 and 2.2.4. Since, for any $\boldsymbol{v} \in \text{Ker}(\boldsymbol{X})$,

$$\left\|\boldsymbol{X}^+\boldsymbol{y}\right\|_2 \leq \left\|\boldsymbol{X}^+\boldsymbol{y}\right\|_2 + \|\boldsymbol{v}\|_2 = \left\|\boldsymbol{X}^+\boldsymbol{y} + \boldsymbol{v}\right\|_2,$$

the ridgeless estimator can be expressed in closed form as $\hat{\boldsymbol{\theta}}_n^{rl} = \boldsymbol{X}^+\boldsymbol{y}$, which is an element of $\text{Range}(\boldsymbol{X}')$ since $\boldsymbol{X}^+ = \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\boldsymbol{X}^+$.

(iii) If $\text{Rank}(\boldsymbol{X}) = p$, then $\text{Ker}(\boldsymbol{X}) = \{\boldsymbol{0}\}$. Moreover, $\boldsymbol{X}'\boldsymbol{X}$ is invertible and we can use the identity $\boldsymbol{X}^+ = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ to conclude that the LSE and the ridgeless estimator are uniquely given by (1.8).

(iv) The ridge problem in (1.4) is an unconstrained optimization problem with the strictly convex, coercive and continuously differentiable objective function

$$f : \mathbb{R}^p \to [0, +\infty);\ \boldsymbol{\theta} \mapsto \|\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\theta}\|_2^2/2 + \lambda/2\,\|\boldsymbol{\theta}\|_2^2.$$

It follows that a solution $\hat{\boldsymbol{\theta}}^r \in \mathbb{R}^n$ exists and it is unique; see, again, Propositions 2.2.3 and 2.2.4. Theorem 2.2.1 implies

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\hat{\boldsymbol{\theta}}^r(\lambda) = \boldsymbol{X}'\boldsymbol{y}.$$

Consider the spectral decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$ in Definition 18. Then,

$$\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I} = \boldsymbol{V}\boldsymbol{S}'\boldsymbol{S}\boldsymbol{V}' + \lambda\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{V}\begin{bmatrix} \text{diag}(s_1^2, \ldots, s_r^2) + \lambda & \boldsymbol{0} \\ \boldsymbol{0} & \lambda\boldsymbol{I} \end{bmatrix}\boldsymbol{V}',$$

which is positive definite, and thus $\hat{\boldsymbol{\theta}}_n^r(\lambda) = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the solution to the FOCs. Finally, to prove that $\hat{\boldsymbol{\theta}}_n^r(\lambda) \in \text{Range}(\boldsymbol{X}')$, notice that $\mathbf{P}_{\text{Range}(\boldsymbol{X}')} = \boldsymbol{V}\boldsymbol{S}^+\boldsymbol{S}\boldsymbol{V}'$, where

$$\boldsymbol{S}^+\boldsymbol{S} = \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Thus,

$$\begin{aligned} \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\hat{\boldsymbol{\theta}}_n^r(\lambda) =& \boldsymbol{V}\boldsymbol{S}^+\boldsymbol{S}\boldsymbol{V}'\boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{V}'\boldsymbol{S}'\boldsymbol{U}'\boldsymbol{y} \\ =& \boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{U}'\boldsymbol{y} = \hat{\boldsymbol{\theta}}_n^r(\lambda). \end{aligned}$$

We conclude that $\mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\hat{\boldsymbol{\theta}}_n^r(\lambda) = \boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S}+\lambda \boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{U}'\boldsymbol{y}$, i.e., $\hat{\boldsymbol{\theta}}_n^r(\lambda) \in \text{Range}(\boldsymbol{X}')$.

(v) The lasso problem in (1.5) is an unconstrained optimization problem with the convex and coercive objective function

$$f : \mathbb{R}^p \to [0, +\infty);\ \boldsymbol{\theta} \mapsto \|\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\theta}\|_2^2/2 + \lambda \|\boldsymbol{\theta}\|_1 .$$

It follows that a solution $\hat{\boldsymbol{\theta}}^l(\lambda) \in \mathbb{R}^p$ exists; see Proposition 2.2.3. However, we demonstrate by counterexample that this solution is, in general, not unique. Consider a sample $(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^n \times \mathbb{R}^{n\times 2}$ where the two predictors are identical, i.e., $\boldsymbol{x}_1 = \boldsymbol{x}_2 \in \mathbb{R}^n$, and assume that there exists a corresponding lasso solution $\hat{\boldsymbol{\theta}}_n^l(\lambda) = [\hat{\theta}_{n1}^l(\lambda), \hat{\theta}_{n2}^l(\lambda)]' \in \mathbb{R}^2$ that is non-zero. Then,

$$\hat{\boldsymbol{\theta}}_1 = [\hat{\theta}_{n1}^l(\lambda) + \hat{\theta}_{n2}^l(\lambda), 0]$$

and

$$\hat{\boldsymbol{\theta}}_2 = [0, \hat{\theta}_{n1}^l(\lambda) + \hat{\theta}_{n2}^l(\lambda)]$$

are two distinct coefficient vectors that produce the same fit, $\boldsymbol{X}\hat{\boldsymbol{\theta}}_1 = \boldsymbol{X}\hat{\boldsymbol{\theta}}_2$, and have identical l_1-norms, $\|\hat{\boldsymbol{\theta}}_1\|_1 = \|\hat{\boldsymbol{\theta}}_2\|_1$. Consequently, in this example there exist at least two distinct lasso solutions.

□

Remark 2 (Computation ridgeless and ridge). The closed form expressions of the LSE, ridgeless and ridge estimators are useful analytical result. However, for numerical stability, it is recommended to compute these estimators by solving their corresponding *normal equations*, which are $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}}_n = \boldsymbol{X}'\boldsymbol{y}$ for the LSE or ridgless, and $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})\hat{\boldsymbol{\theta}}_n^r = \boldsymbol{X}'\boldsymbol{y}$ for the ridge.

Remark 3 (Collinearity). Using the notation in Definition 18, the minimum nonzero eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$ is s_r^2. If $r < p$, then $\boldsymbol{X}'\boldsymbol{X}$ has $p - r$ zero eigenvalues and the predictors are said to be collinear, that is, they are linearly dependent. In this case $\text{Ker}(\boldsymbol{X})$ is not trivial (it contains nonzero elements), hence the LSE is not unique. Moreover, if $s_r \approx 0$, then the computation of

$$\boldsymbol{X}^+ = \boldsymbol{V} \begin{bmatrix} \text{diag}(1/s_1, \dots, 1/s_r) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{U}',$$

and hence of the ridgeless estimator, is unstable. The ridge estimator instead may not display these computational hurdles, provided that the penalty parameter λ is large enough. That is because the minimum eigenvalue of $(\boldsymbol{X}'\boldsymbol{X} + \lambda I)$ is $s_r^2 + \lambda$. In Section 1.1.5 we show that, if $s_r \approx 0$, the ridgeless (ridge) estimator's MSE and MPR satisfy loose (sharp) concentration inequalities.

Remark 4 (Uniqueness of the lasso solution). Tibshirani [2013] shows that, under some conditions, the lasso estimator is unique. For instance, if the predictors in $\boldsymbol{X}$ are in *general position*, then the lasso solution is unique. Specifically, a set $(\boldsymbol{a}_j)_{j=1}^p$ where $\boldsymbol{a}_j \in \mathbb{R}^n$ for all j is in general position if any affine subspace of $\mathbb{R}^n$ of dimension $k < n$ contains at most $k + 1$ elements of the set $\{\pm\boldsymbol{x}_1, \pm\boldsymbol{x}_2, \dots \pm \boldsymbol{x}_p\}$, excluding antipodal pairs of points (that is, points differing only by a sign flip). If the predictors are (non redundant) continuous random variables, they are almost surely in general position, and hence the lasso solution is unique. As a result, non-uniqueness of the lasso solution typically occurs with discrete-valued data, such as those comprising dummy or categorical variables.

Since the LSE, ridgeless, ridge and lasso estimators exist, their linear predictions exist too. Moreover, the linear predictions of the uniquely defined estimators, like ridgeless and ridge, are trivially unique. Remarkably, some estimators that may not be unique entail unique linear predictions. The next lemma implies that the LSE and lasso are among these estimators.

Lemma 1.1.2. *Let $h : \mathbb{R}^p \to (-\infty, +\infty]$ be a proper convex function. Then $\boldsymbol{X}\boldsymbol{\theta}_1 = \boldsymbol{X}\boldsymbol{\theta}_2$ and $h(\boldsymbol{\theta}_1) = h(\boldsymbol{\theta}_2)$ for every minimizers $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$ of*

$$f : \mathbb{R}^p \to (-\infty, +\infty]; \boldsymbol{\theta} \mapsto \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + h(\boldsymbol{\theta}).$$

Proof. Assume that $\boldsymbol{X}\boldsymbol{\theta}_1 \neq \boldsymbol{X}\boldsymbol{\theta}_2$, and let $\delta := \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$. By Proposition

2.2.2, the set of minimizers of f is convex. Thus, for any $\alpha \in (0,1)$:

$$\begin{aligned}
\delta =& f(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) \\
=& \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}[\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2]\right\|_2^2 + h(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) \\
<& \alpha\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_1\right\|_2^2 + (1-\alpha)\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_2\right\|_2^2 + h(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) \\
\leq& \alpha\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_1\right\|_2^2 + (1-\alpha)\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_2\right\|_2^2 + \alpha h(\boldsymbol{\theta}_1) + (1-\alpha)h(\boldsymbol{\theta}_2) \\
=& \alpha f(\boldsymbol{\theta}_1) + (1-\alpha)f(\boldsymbol{\theta}_2) = \delta,
\end{aligned}$$

where the strict inequality follows from the strict convexity of $g : \boldsymbol{v} \mapsto \|\boldsymbol{y} - \boldsymbol{v}\|_2^2$. Since the conclusion $\delta < \delta$ is absurd, we must have $\boldsymbol{X}\boldsymbol{\theta}_1 = \boldsymbol{X}\boldsymbol{\theta}_2$, and since $f(\boldsymbol{\theta}_1) = f(\boldsymbol{\theta}_2)$, it follows that $h(\boldsymbol{\theta}_1) = h(\boldsymbol{\theta}_2)$ as well. □

While we make use of this lemma for proving uniqueness of the predictions of lasso, we can use a more direct approach for the other estimators' predictions, which directly provides their closed form expressions. We further show that the LSE's prediction has the geometric interpretation of being the unique vector in the range of $\boldsymbol{X}$ that is closest to $\boldsymbol{y}$ in l_2 distance, and that the residual vector is orthogonal to the range of $\boldsymbol{X}$.

Theorem 1.1.3 (Uniqueness of linear predictions)**.** *The following statements hold:*

(i) The linear predictions of the LSE and the ridgeless estimator are uniquely given by:

$$\mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{ls}) = \mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{rl}) = \mathbf{P}_{\mathrm{Range}(\boldsymbol{X})}\,\boldsymbol{y}, \tag{1.10}$$

which is the unique vector $\boldsymbol{v} \in \mathrm{Range}(\boldsymbol{X})$ such that

$$\|\boldsymbol{y} - \boldsymbol{v}\|_2 = \inf\{\|\boldsymbol{y} - \boldsymbol{z}\|_2 \ : \ \boldsymbol{z} \in \mathrm{Range}(\boldsymbol{X})\}.$$

Moreover, the residual vector $\boldsymbol{y} - \mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{ls}) = \boldsymbol{y} - \mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{rl})$ is orthogonal to $\mathrm{Range}(\boldsymbol{X})$.

(ii) The linear prediction of the ridge estimator is uniquely given, for $\lambda > 0$, by

$$\mathrm{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{r}(\lambda)) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

(iii) The linear prediction of the lasso estimator is unique.

Proof. (i) The linear predictions $\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{ls})$ and $\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{rl})$ are uniquely given by (1.10) because all solutions to the least squares problem $\hat{\boldsymbol{\theta}} \in \boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X})$ yield the same prediction

$$\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}) = \boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y} = \mathbf{P}_{\text{Range}(\boldsymbol{X})}\,\boldsymbol{y}.$$

By the definition of $\hat{\boldsymbol{\theta}}_n^{ls}$ and the fact that $\text{Range}(\boldsymbol{X})$ is a closed vector subspace of $\mathbb{R}^n$, the remaining claims follow as a direct application of the Hilbert projection theorem (Theorem 2.2.2).

(ii) This result follows directly from the closed form expression (1.9) of the ridge estimator.

(iii) Since the l_1-norm is convex, the result follows by Lemma 1.1.2.

□

1.1.2 Equivalent expressions and relations

The ridgeless and the ridge, together with their corresponding linear predictions, admit the following simple expressions.

Proposition 1.1.1 (Spectral expression of ridgeless and ridge)**.** *Given the spectral decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$ in Definition 18:*

(i) The ridgeless estimator is given by

$$\hat{\boldsymbol{\theta}}_n^{rl} = \left(\sum_{j=1}^{r} \frac{1}{s_j}\boldsymbol{v}_j\boldsymbol{u}_j'\right)\boldsymbol{y}.$$

The corresponding linear prediction is

$$\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{rl}) = \left(\sum_{j=1}^{r} \boldsymbol{u}_j\boldsymbol{u}_j'\right)\boldsymbol{y}. \tag{1.11}$$

(ii) The ridge estimator with $\lambda > 0$ is given by

$$\hat{\boldsymbol{\theta}}_n^{r}(\lambda) = \left(\sum_{j=1}^{r} \frac{s_j}{s_j^2 + \lambda}\boldsymbol{v}_j\boldsymbol{u}_j'\right)\boldsymbol{y}.$$

The corresponding linear prediction is

$$\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{r}) = \left(\sum_{j=1}^{r} \frac{s_j^2}{s_j^2 + \lambda}\boldsymbol{u}_j\boldsymbol{u}_j'\right)\boldsymbol{y}. \tag{1.12}$$

Proof. (i) From the closed-form expression of the ridgeless estimator,

$$\hat{\boldsymbol{\theta}}_n^{rl} = \boldsymbol{X}^+\boldsymbol{y} = \boldsymbol{V}\boldsymbol{S}^+\boldsymbol{U}'\boldsymbol{y} = \left(\sum_{j=1}^{r} \frac{1}{s_j}\boldsymbol{v}_j\boldsymbol{u}_j'\right)\boldsymbol{y}.$$

Therefore,

$$\boldsymbol{X}\hat{\boldsymbol{\theta}}_n^{rl} = \boldsymbol{U}\boldsymbol{S}'\boldsymbol{S}^+\boldsymbol{U}'\boldsymbol{y} = \left(\sum_{j=1}^{r} \boldsymbol{u}_j\boldsymbol{u}_j'\right)\boldsymbol{y}.$$

(ii) From the closed-form expression of the ridge estimator,

$$\begin{aligned}\hat{\boldsymbol{\theta}}_n^{r} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} &= \boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{U}'\boldsymbol{y} \\ &= \left(\sum_{j=1}^{r} \frac{s_j}{s_j^2 + \lambda}\boldsymbol{v}_j\boldsymbol{u}_j'\right)\boldsymbol{y}.\end{aligned}$$

Therefore,

$$\boldsymbol{X}\hat{\boldsymbol{\theta}}_n^{r} = \boldsymbol{U}\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{U}'\boldsymbol{y} = \left(\sum_{j=1}^{r} \frac{s_j^2}{s_j^2 + \lambda}\boldsymbol{u}_j\boldsymbol{u}_j'\right)\boldsymbol{y}.$$

□

Using Definition 18, matrix $\mathbf{P}_{\text{Range}(\boldsymbol{X})} = \boldsymbol{X}\boldsymbol{X}^+ = \sum_{j=1}^{r} \boldsymbol{u}_j\boldsymbol{u}_j'$, where $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ is an orthogonal basis of Range($\boldsymbol{X}$). From expression (1.11), it follows that the prediction of the ridgeless estimator is the orthogonal projection of $\boldsymbol{y}$ onto the range of $\boldsymbol{X}$. Expression (1.12) instead shows that the ridge estimator shrinks this projection, shrinking less the directions $\boldsymbol{u}_j$ associated to high variance (high s_j), and more the directions $\boldsymbol{u}_j$ associated to low variance (low s_j); see Figure 1.3. Indeed, for fixed $\lambda > 0$, the weight $s_j^2/(s_j^2 + \lambda) \to 0$ as $s_j \to 0$, and $s_j^2/(s_j^2 + \lambda) \to 1$ as $s_j \to \infty$.

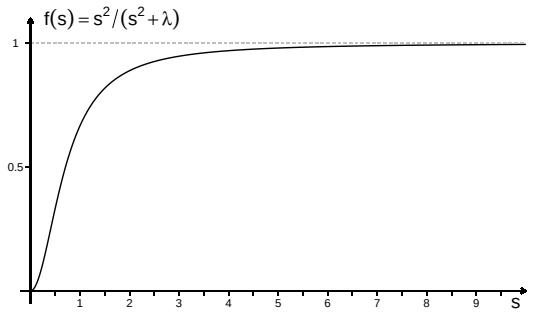

Figure 1.3: Shrinkage of principal components in the linear prediction of ridge when $\lambda = 1/2$.

The ridgeless estimator can also be expressed as a penalized LSE.

Proposition 1.1.2 (Penalized expression of ridgeless)**.** *The ridgeless estimator is the only solution to the least squares problem (1.2) that is in* Range($\boldsymbol{X}'$)*, and it can be expressed as:*

$$\hat{\boldsymbol{\theta}}_n^{rl} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \frac{1}{2}\boldsymbol{\theta}'\,\mathbf{P}_{\mathrm{Ker}(\boldsymbol{X})}\,\boldsymbol{\theta}. \tag{1.13}$$

Proof. From Theorem 1.1.1, the solution set of the least squares problem (1.2) is $\hat{\boldsymbol{\theta}}_n^{rl} + \mathrm{Ker}(\boldsymbol{X})$, where $\hat{\boldsymbol{\theta}}_n^{rl} = \boldsymbol{X}^+\boldsymbol{y}$ is in Range($\boldsymbol{X}'$). Since $\mathrm{Ker}(\boldsymbol{X}) \perp \mathrm{Range}(\boldsymbol{X}')$, $\hat{\boldsymbol{\theta}}_n^{rl}$ is the only solution in Range($\boldsymbol{X}'$). Moreover, penalty $h : \boldsymbol{\theta} \mapsto \boldsymbol{\theta}'\,\mathbf{P}_{\mathrm{Ker}(\boldsymbol{X})}\,\boldsymbol{\theta}$ is zero in $\hat{\boldsymbol{\theta}}_n^{rl}$, and strictly positive at any other least squares solution. We conclude that $\hat{\boldsymbol{\theta}}_n^{rl}$ minimizes (1.13). □

Following linear transformations relate the ridgeless and the ridge estimators.

Proposition 1.1.3 (Links between ridgeless and ridge)**.** *Following relations between the ridge and the minimum norm least squares estimators hold:*

$$\hat{\boldsymbol{\theta}}_n^{r}(\lambda) = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}}_n^{rl}, \tag{1.14}$$

$$\hat{\boldsymbol{\theta}}_n^{rl} = (\boldsymbol{X}'\boldsymbol{X})^{+}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})\hat{\boldsymbol{\theta}}_n^{r}(\lambda), \tag{1.15}$$

$$\lim_{\lambda\to 0} \hat{\boldsymbol{\theta}}_n^{r}(\lambda) = \hat{\boldsymbol{\theta}}_n^{rl}. \tag{1.16}$$

Proof. Using $\boldsymbol{X} = \mathbf{P}_{\mathrm{Range}(\boldsymbol{X})}\,\boldsymbol{X}$ which implies $\boldsymbol{X}' = \boldsymbol{X}'\,\mathbf{P}_{\mathrm{Range}(\boldsymbol{X})}$, we have

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{y},$$

and thus

$$\hat{\boldsymbol{\theta}}_n^r(\lambda) = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}}_n^{rl}.$$

Using $\boldsymbol{X}^+ = \boldsymbol{X}^+(\boldsymbol{X}^+)'\boldsymbol{X}'$, we have

$$\boldsymbol{X}^+\boldsymbol{y} = \boldsymbol{X}^+(\boldsymbol{X}^+)'(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Moreover, $\boldsymbol{X}^+(\boldsymbol{X}^+)' = (\boldsymbol{X}'\boldsymbol{X})^+$ implies

$$\hat{\boldsymbol{\theta}}_n^{rl} = (\boldsymbol{X}'\boldsymbol{X})^+(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})\hat{\boldsymbol{\theta}}_n^r(\lambda).$$

Finally, since $\boldsymbol{X}^+ = \lim_{\lambda \to 0}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'$, we have $\lim_{\lambda \to 0} \hat{\boldsymbol{\theta}}_n^r(\lambda) = \hat{\boldsymbol{\theta}}_n^{rl}$. □

Expression (1.16) explains why estimator (1.3) is called the ridgeless estimator. The ridge and lasso estimators can be expressed as constrained least squares problems.

Proposition 1.1.4 (Equivalence between penalized and constrained least squares)**.** *For $c \geq 0$, $\lambda \geq 0$, and some norm $\|\cdot\| : \mathbb{R}^p \to \mathbb{R}$, define:*

$$\mathcal{C}(c) := \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 / 2 \ : \ \|\boldsymbol{\theta}\| \leq c \right\};$$

$$\mathcal{P}(\lambda) := \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 / 2 + \lambda \|\boldsymbol{\theta}\| \right\}.$$

Then, for a given $c > 0$, there exists $\lambda_0 \geq 0$ such that $\mathcal{C}(c) \subset \mathcal{P}(\lambda_0)$. Conversely, for a given $\lambda > 0$, there exists $c_0 \geq 0$ such that $\mathcal{P}(\lambda) \subset \mathcal{C}(c_0)$.

Proof. The objective function $h : \boldsymbol{\theta} \mapsto \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 / 2$ is convex and continuous, and the constraint set $\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{\theta}\| \leq c\}$ is not empty, closed, bounded and convex. By the KKT theorem for convex problems, $\hat{\boldsymbol{\theta}} \in \mathcal{C}(c)$ for any $c > 0$ if and only if $\hat{\boldsymbol{\theta}}$ satisfies the KKT conditions, for some corresponding $\lambda_0 \geq 0$:

$$\begin{aligned}
&\mathbf{0} \in \lambda_0 \partial \|\hat{\boldsymbol{\theta}}\| + \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}'\boldsymbol{y}, \\
&\|\hat{\boldsymbol{\theta}}\| \leq c, \\
&\lambda_0(\|\hat{\boldsymbol{\theta}}\| - c) = 0.
\end{aligned}$$

By Theorem 2.2.1, the first of these conditions implies that $\hat{\boldsymbol{\theta}} \in \mathcal{P}(\lambda_0)$. Now fix a $\lambda > 0$ and notice that $\mathcal{P}(\lambda)$ is not empty, given that its objective function is convex, continuous and coercive; see Proposition 2.2.3. We can thus take some $\hat{\boldsymbol{\theta}} \in \mathcal{P}(\lambda)$. Then, $\hat{\boldsymbol{\theta}}$ satisfies the KKT conditions for $c_0 = \|\hat{\boldsymbol{\theta}}\|$, which implies $\hat{\boldsymbol{\theta}} \in \mathcal{C}(c_0)$. □

Note that the link between the penalty parameter λ and the constraint parameter c is not explicit.

1.1.3 Geometric interpretation

We illustrate the geometry of the least squares, ridge, and lasso solutions through a simple example. Consider the linear model (1.1), with $p = 2$, $\varepsilon_{0i} \sim iiN(0, 1)$, $\boldsymbol{\theta}_0 = [1.5, 0.5]'$, $\mathbb{E}[\boldsymbol{x}_i \varepsilon_0] = \mathbf{0}$, and

$$\boldsymbol{x}_i \sim iiN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

Figure 1.4 shows the level curves of the least squares loss function $f(\boldsymbol{\theta}) := \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2/2$, corresponding to values $f_1 < f_2 < f_3 < f_4$. Its minimizer, or least squares solution $\hat{\boldsymbol{\theta}}_n^{\text{ls}}$, which coincides with the ridgeless solution $\hat{\boldsymbol{\theta}}_n^{rl}$, is highlighted in the figure.

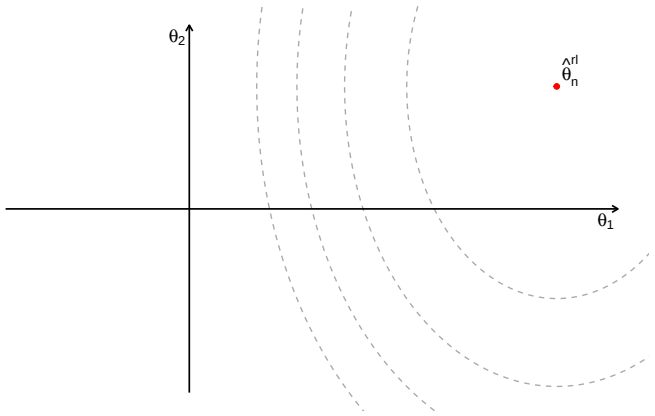

Figure 1.4: Geometry of the least squares solution.

To illustrate the geometry of the ridge solution, we consider the constrained formulation of the ridge problem; see Proposition 1.1.4. Figure 1.5 demonstrates the impact of imposing the ridge constraint, represented by the sphere $\{\boldsymbol{\theta} \in \mathbb{R}^2 \ : \ \|\boldsymbol{\theta}\|_2 \leq c\}$ with $c = 0.5$, on the least squares problem. The ridge solution $\hat{\boldsymbol{\theta}}_n^r$ is located at the intersection between the ridge constraint and the lower level set of the least squares loss at the lowest height (see Appendix 2.2 Definition 71) for which the intersection is non-empty. If the ridgeless solution $\hat{\boldsymbol{\theta}}_n^{rl}$ lies within the constraint boundary, then $\hat{\boldsymbol{\theta}}_n^r$ coincides with $\hat{\boldsymbol{\theta}}_n^{rl}$. Otherwise, the ridge solution $\hat{\boldsymbol{\theta}}_n^r$, by construction, is closer to the origin than $\hat{\boldsymbol{\theta}}_n^{rl}$, demonstrating the shrinkage effect of the ridge penalty. In general, $\hat{\boldsymbol{\theta}}_n^r$ is dense (i.e., contains no zero elements) with probability one.

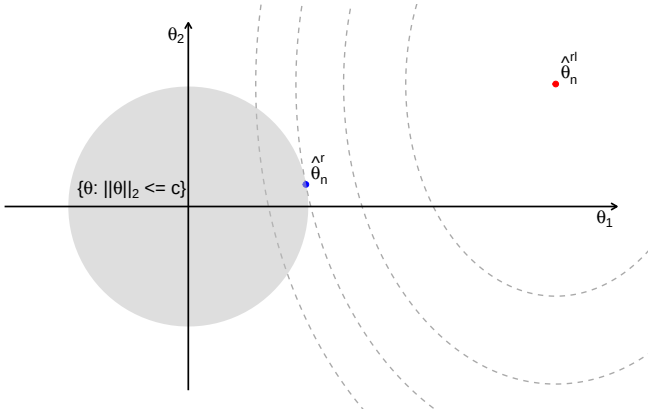

Figure 1.5: Geometry of the ridge solution.

Figure 1.6 illustrates the effect of the lasso constraint, represented by the rotated square $\{\boldsymbol{\theta} \in \mathbb{R}^2 \ : \ \|\boldsymbol{\theta}\|_1 \leq c\}$ with $c = 0.5$, on the least squares solution. Like the ridge solution, the lasso solution $\hat{\boldsymbol{\theta}}_n^l$ is located at the intersection of the lasso constraint and the lower level set of the least squares loss at the lowest height for which the intersection is non-empty. is located at the intersection between the lasso constraint and the lower level set of the least squares loss at the lowest height for which the intersection is not empty. For small values of c, this intersection is more likely to occur along one of the coordinate axes. As a result, the lasso solution tends to be sparse, meaning that some components of $\hat{\boldsymbol{\theta}}_n^l$ are exactly zero.

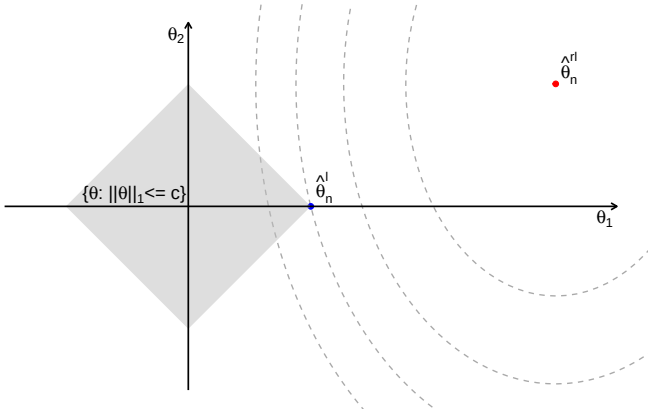

Figure 1.6: Geometry of the lasso solution.

As discussed in Section 1.1, the lasso estimator serves as an approximation to the l_0-estimator (1.6). This relationship becomes evident through visual comparison of Figure 1.6 and Figure 1.7. The lasso constraint set $\{\boldsymbol{\theta} \ : \ \|\boldsymbol{\theta}\|_1 \leq c\}$ is the convex hull (i.e., the smallest convex superset) of the constraint set underlying the l_0-estimator, which is given by: $\{\boldsymbol{\theta} \ : \ \|\boldsymbol{\theta}\|_0 \leq c, \ \|\boldsymbol{\theta}\|_\infty \leq 1\}$. Further details on this approximation can be found in Argyriou et al. [2012].

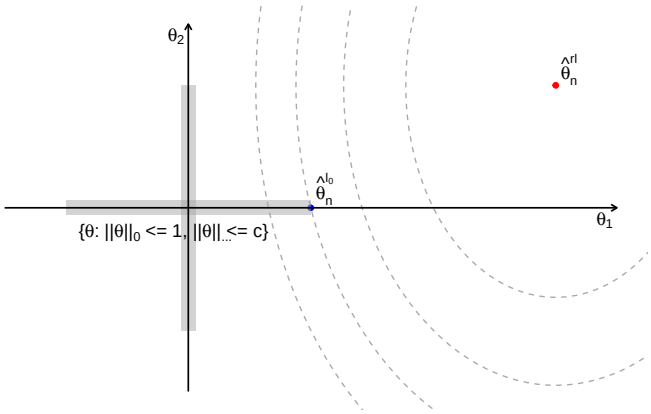

Figure 1.7: Geometry of the l_0 solution.

To illustrate the geometry of the ridgeless solution, consider the linear model (1.1) with $p = 2$, $\varepsilon_{0i} \sim \text{iid}N(0,1)$, $x_{i1} \sim \text{iid}N(0,1)$, $\mathbb{E}[x_{i1}\varepsilon_{0i}] = 0$, and $x_{i2} = 2x_{i1}$. In this case, the predictors are linearly dependent. As a result, the second-moment matrix of the predictors is reduced-rank:

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'] = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix},$$

with $\text{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']) = 1$. Let $\mathbb{E}[x_{i1}y_i] = 1$. The identifying condition $\mathbb{E}[\boldsymbol{x}_i\varepsilon_{0i}] = \mathbf{0}$ holds if and only if the population coefficient $\boldsymbol{\theta}_0$ satisfies

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']\boldsymbol{\theta}_0 = \mathbb{E}[\boldsymbol{x}_i y_i] \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \boldsymbol{\theta}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Thus, any coefficient in the set $\boldsymbol{\theta}_0^{\text{rl}} + \text{Kernel}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$ satisfies this condition, where

$$\boldsymbol{\theta}_0^{rl} := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^{+}\mathbb{E}[\boldsymbol{x}_i y_i] = [0.2, 0.4]'.$$

If the sample size $n >$ Rank$(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, then Ker$(\boldsymbol{X}) \supset$ Ker$(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, and the same issue arises in the finite-sample least squares problem, where the objective function $f(\boldsymbol{\theta})$ is minimized at any point on the affine set

$$\hat{\boldsymbol{\theta}}_n^{rl} + \text{Ker}(\boldsymbol{X}).$$

Figure 1.8 depicts the level curves of $f(\boldsymbol{\theta})$ at $f(\hat{\boldsymbol{\theta}}_n^{\text{rl}}) = f_1 < f_2 < f_3$. These curves are parallel lines, unlike the typical ellipses seen in full-rank cases. The ridgeless estimator is the minimum l_2-norm solution to the least squares problem, as expected from its construction.

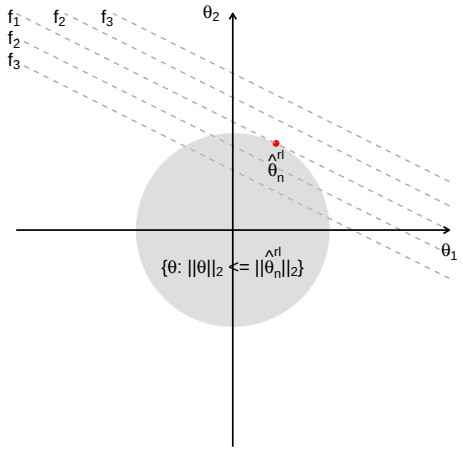

Figure 1.8: Geometry of the ridgeless solution.

1.1.4 Computation of lasso

By Theorem 2.2.1, a function $f : \mathbb{R}^p \to \mathbb{R}$ is minimized at $\boldsymbol{\theta}^* \in \mathbb{R}^p$ if and only if $\mathbf{0} \in \partial f(\boldsymbol{\theta}^*)$. In the lasso problem, the objective function

$$f^l : \mathbb{R}^p \to \mathbb{R}; \boldsymbol{\theta} \mapsto \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \tag{1.17}$$

contains the l_1-norm, which makes f^l non-smooth.[5] As a result, the subdifferential of f^l at a minimizer $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ is not a singleton, implying that the lasso estimator may not be unique. Moreover, due to the complexity of $\partial f^l(\hat{\boldsymbol{\theta}}_n^l(\lambda))$, no closed-form solution exists in general for the lasso estimator.

Fortunately, Proposition 1.1.4 implies that the lasso problem is a quadratic program with a convex constraint, which allows for the computation of the

[5]Notice that f^l is the objective function in (1.5), multiplied by $1/n$. This term does not show up in the penalization term as it is absorbed by λ.

lasso estimator using various quadratic programming algorithms. One particularly simple and effective method is the *cyclical coordinate descent algorithm*, which minimizes the convex objective function by iterating through each coordinate independently. This approach provides insight into how the lasso solution is obtained.

Consider the *soft-thresholding operator* for a given $\lambda > 0$, which is defined as the function

$$\mathcal{S}_\lambda : \mathbb{R} \to \mathbb{R}; \eta \mapsto \begin{cases} \eta - \lambda, & \eta > \lambda \\ 0, & \eta \in [-\lambda, \lambda] \\ \eta + \lambda, & \eta < -\lambda \end{cases}.$$

This operator is illustrated in Figure 1.9.

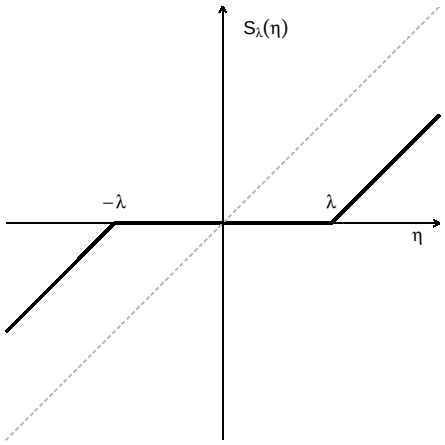

Figure 1.9: Soft-thresholding operator.

The soft-thresholding operator provides a direct way to compute the lasso estimator in a univariate regression model, i.e., when there is only one predictor.

Proposition 1.1.5 (Lasso solution for univariate regression)**.** *Given $\lambda > 0$ and $\boldsymbol{X}_1 \in \mathbb{R}^n$ such that $\boldsymbol{X}_1' \boldsymbol{X}_1 > 0$, we have*

$$\hat{\theta}_n^l(\lambda) := \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}_1 \theta\|_2^2 + \lambda |\theta| \right\} = \frac{\mathcal{S}_\lambda(\boldsymbol{X}_1' \boldsymbol{y}/n)}{\boldsymbol{X}_1' \boldsymbol{X}_1 / n}.$$

Proof. The subdifferential of $f : \theta \mapsto \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}_1 \theta\|_2^2 + \lambda |\theta|$ at $\hat{\theta} \in \mathbb{R}$ reads

$$\partial f(\hat{\theta}) = b\hat{\theta} - a + \lambda \partial |\hat{\theta}|,$$

where $a := \boldsymbol{X}_1'\boldsymbol{y}/n$ and $b := \boldsymbol{X}_1'\boldsymbol{X}_1/n$. From Theorem 2.2.1, and the subdifferential of the absolute value function (Appendix 2.2, Example 7), $\hat{\theta}$ is a minimizer of f if and only if

$$0 \in \partial f(\hat{\theta}) \quad \Longleftrightarrow \quad a \in b\hat{\theta} + \begin{cases} \{\lambda\}, & \hat{\theta} > 0 \\ [-\lambda, \lambda], & \hat{\theta} = 0 \\ \{-\lambda\}, & \hat{\theta} < 0 \end{cases}.$$

This condition reads: (i) if $\hat{\theta} > 0$, then $\hat{\theta} = (a - \lambda)/b$, implying $a > \lambda$; (ii) if $\hat{\theta} = 0$, then $-\lambda \leq a \leq \lambda$; and (iii) if $\hat{\theta} < 0$, then $\hat{\theta} = (a + \lambda)/b$, implying $a < -\lambda$. These cases are summarized by $\hat{\theta} = \mathcal{S}_\lambda(a)/b$. □

Proposition 1.1.5 can be used to show that the $j-$th coordinate of the lasso solution in a multivariate regression model, i.e., when there is more than just one predictor, satisfies an expression based on the soft-thresholding operator applied to the residual of a lasso regression of $\boldsymbol{y}$ onto the predictors $\boldsymbol{X}_k$ at position $k \neq j$.

Theorem 1.1.4 (Lasso solution)**.** *Let $\boldsymbol{X}_j$ denote the $j-$th column of $\boldsymbol{X}$ and $\boldsymbol{X}_{(-j)}$ denote $\boldsymbol{X}$ without the $j-$th column. Assume that $\boldsymbol{X}_j'\boldsymbol{X}_j > 0$ for all $j = 1, \ldots, p$. Then, given $\lambda > 0$, any lasso solution $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ is such that for all $j = 1, \ldots, p$:*

$$\hat{\theta}_{n,j}^l(\lambda) = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{e}_j - \boldsymbol{X}_j\theta\|_2^2 + \lambda|\theta| \right\} = \frac{\mathcal{S}_\lambda(\boldsymbol{X}_j'\boldsymbol{e}_j/n)}{\boldsymbol{X}_j'\boldsymbol{X}_j/n}, \tag{1.18}$$

where $\hat{\theta}_{n,j}(\lambda)$ is the $j-$th element of $\hat{\boldsymbol{\theta}}_n^l(\lambda)$, $\hat{\boldsymbol{\theta}}_{n,(-j)}^l(\lambda)$ is $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ without the $j-$th element, and $\boldsymbol{e}_j := \boldsymbol{y} - \boldsymbol{X}_{(-j)}\hat{\boldsymbol{\theta}}_{n,(-j)}^l(\lambda)$.

Proof. The subdifferential of the lasso objective function f^l defined in (1.17) at $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ is

$$\partial f^l(\hat{\boldsymbol{\theta}}) = (\boldsymbol{X}'\boldsymbol{X}/n)\hat{\boldsymbol{\theta}} - \boldsymbol{X}'\boldsymbol{y}/n + \lambda\partial\|\hat{\boldsymbol{\theta}}\|_1,$$

where

$$\partial\|\hat{\boldsymbol{\theta}}\|_1 = \{\boldsymbol{v} \in \mathbb{R}^p \ : \ v_j \in \partial|\hat{\theta}_j| \text{ for all } j = 1, \ldots, p\}.$$

and the subdifferential of the absolute value function $|\cdot|$ is given in Appendix 2.2, Example 7. From Theorem 2.2.1, a minimizer $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ of f satisfies

$$\boldsymbol{0} \in \partial f(\hat{\boldsymbol{\theta}}_n^l(\lambda)) \quad \Longleftrightarrow \quad \boldsymbol{X}'\boldsymbol{y}/n \in \boldsymbol{X}'\boldsymbol{X}/n\hat{\boldsymbol{\theta}}_n^l(\lambda) + \lambda\partial\|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1.$$

This condition holds if and only if for all $j = 1, \ldots, p$:

$$\begin{aligned}
\boldsymbol{X}_j'\boldsymbol{y}/n &\in \boldsymbol{X}_j'\boldsymbol{X}/n\hat{\boldsymbol{\theta}}_n^l(\lambda) + \lambda\partial|\hat{\theta}_{n,j}(\lambda)| \quad \Longleftrightarrow \\
\boldsymbol{X}_j'\boldsymbol{e}_j/n &\in \boldsymbol{X}_j'\boldsymbol{X}_j/n\hat{\theta}_{n,j}(\lambda) + \lambda\partial|\hat{\theta}_{n,j}(\lambda)| \quad \Longleftrightarrow \\
\hat{\theta}_{n,j}(\lambda) &= \frac{\mathcal{S}_\lambda(\boldsymbol{X}_j'\boldsymbol{e}_j/n)}{\boldsymbol{X}_j'\boldsymbol{X}_j/n},
\end{aligned}$$

where the first double implication follows from

$$\boldsymbol{X}_j'\boldsymbol{X}\hat{\boldsymbol{\theta}}_n^l(\lambda) = \sum_{k=1}^{p} \boldsymbol{X}_j'\boldsymbol{X}_k\hat{\theta}_{n,k}^l(\lambda) = \boldsymbol{X}_j'\boldsymbol{X}_{(-j)}\hat{\boldsymbol{\theta}}_{n,(-j)}^l(\lambda) + \boldsymbol{X}_j'\boldsymbol{X}_j\hat{\theta}_{n,j}(\lambda),$$

and the last double implication follows from Proposition 1.1.5 since, by Theorem 2.2.1,

$$\begin{aligned}
\boldsymbol{X}_j'\boldsymbol{e}_j/n &\in \boldsymbol{X}_j'\boldsymbol{X}_j/n\hat{\theta}_{n,j}(\lambda) + \lambda\partial|\hat{\theta}_{n,j}(\lambda)| \quad \Longleftrightarrow \\
\hat{\theta}_{n,j}(\lambda) &= \underset{\theta\in\mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{e}_j - \boldsymbol{X}_j\theta\|_2^2 + \lambda|\theta| \right\}.
\end{aligned}$$

□

Theorem 1.1.4 suggests that the lasso solution can be computed by a cyclical coordinate minimization algorithm. This method is an iterative algorithm that, given a candidate solution $\hat{\boldsymbol{\theta}}^{(k)}$ at iteration $t+1$, it chooses to update a coordinate j as

$$\hat{\theta}_j^{(t+1)} = \underset{\theta\in\mathbb{R}}{\operatorname{argmin}}\, f(\hat{\theta}_1^{(t)}, \ldots, \hat{\theta}_{j-1}^{(t)}, \theta, \hat{\theta}_{j+1}^{(t)}, \ldots, \hat{\theta}_p^{(t)}),$$

and sets $\hat{\theta}_k^{(t+1)} = \hat{\theta}_k^{(t)}$ for $k \neq j$. A typical choice for the lasso solution would be to cycle through the coordinates in their natural order: from 1 to p. The coordinate descent algorithm is guaranteed to converge to a global minimizer of any convex cost function $f : \mathbb{R}^p \to \mathbb{R}$ satisfying the additive decomposition:

$$f : \boldsymbol{\theta} \mapsto g(\boldsymbol{\theta}) + \sum_{j=1}^{p} h_j(\theta_j),$$

where $g : \mathbb{R}^p \to \mathbb{R}$ is differentiable and convex, and the univariate functions $h_j : \mathbb{R}^p \to \mathbb{R}$ are convex (but not necessarily differentiable); see Tseng [2001]. What makes this algorithm work for the lasso problem is the fact that objective function (1.17) satisfies this separable structure.

Remark 5. If the predictors are measured in different units, it is recommended to standardize them so that $\boldsymbol{X}_j'\boldsymbol{X}_j = 1$ for all j. In this case, the lasso update (1.18) has the simpler form:

$$\hat{\theta}^l_{n,j}(\lambda) = \mathcal{S}_\lambda(\boldsymbol{X}_j'\boldsymbol{e}_j/n).$$

Algorithm 1 summarizes the pseudo-code of the cyclical coordinate descent algorithm for computing the lasso estimator. This algorithm proceeds by cyclically applying the soft-thresholding update in (1.18) for each coordinate, simultaneously updating the residuals $\boldsymbol{e}_j := \boldsymbol{y} - \boldsymbol{X}_{(-j)}\hat{\boldsymbol{\theta}}^l_{n,(-j)}(\lambda)$. The ridgeless or the ridge estimators can be used to initialize the procedure.

Algorithm 1 Cyclical coordinate descent method for the lasso estimator.

Require: $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ such that $\boldsymbol{X}_j'\boldsymbol{X}_j > 0$ for all $j = 1, \ldots, p$
Require: Penalty parameter $\lambda > 0$
Require: Initial estimator $\hat{\boldsymbol{\theta}}^i_n$ (e.g., ridgeless or ridge)
Require: Maximum number of iterations T
  Standardize $\boldsymbol{y}$ and $\boldsymbol{X}$ so that $\boldsymbol{y}'\boldsymbol{1} = 0$, $\boldsymbol{X}'\boldsymbol{1} = \boldsymbol{0}$ and $\text{diag}(\boldsymbol{X}'\boldsymbol{X}/n) = \boldsymbol{I}$
  $\hat{\boldsymbol{\theta}}^{(1)} \leftarrow \hat{\boldsymbol{\theta}}^i_n$
  for $t = 2, \ldots, T$ **do**
    for $j = 1, \ldots, p$ **do**
      $\boldsymbol{e}_j \leftarrow \boldsymbol{y} - \boldsymbol{X}_{(-j)}\hat{\boldsymbol{\theta}}^{(t-1)}_{(-j)}(\lambda)$
      $\hat{\theta}^{(t-1)}_j \leftarrow \mathcal{S}_\lambda\left(\boldsymbol{X}_j'\boldsymbol{e}_j/n\right)$
    end for
    $\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}^{(t-1)}$
    if a suitable stopping rule is satisfied **then**
      Stop and output $\hat{\boldsymbol{\theta}}^{(t)}$
    end if
    $\hat{\boldsymbol{\theta}}^{(t)}$
  end for
  Output $\hat{\boldsymbol{\theta}}^{(T)}$

In practice, it is often desirable to compute the lasso solution not for a single fixed value of λ, for the entire solution path over a range of λ values. A common approach begins by selecting a value of λ just large enough that the only optimal solution is the zero vector. This value is denoted as $\lambda_{\max} = \max_j\{\boldsymbol{X}_j'\boldsymbol{y}/n\}$. From there, we gradually decrease λ by a small amount and run coordinate descent until convergence using the previous solution as a "warm start". By further decreasing the previous solution as a "warm start," we then run coordinate descent until convergence. In this way we can

efficiently compute the solutions over a grid of λ values. This approach is known as *pathwise coordinate descent*.

Coordinate descent is particularly efficient for the lasso because the update rule (1.18) is available in closed form, eliminating the need for iterative searches along each coordinate. Additionally, the algorithm exploits the inherent sparsity of the problem: for sufficiently large values of λ, most coefficients will be zero and will remain unchanged. There are also computational strategies that can predict the active set of variables, significantly speeding up the algorithm. More details on the pathwise coordinate descent algorithm for lasso can be found in Friedman et al. [2007].

Homotopy methods are another class of techniques for solving the lasso estimator. They produce the entire path of solutions in a sequential fashion, starting at zero. An homotopy method that is particularly efficient at computing the entire lasso path is the *least angle regression* (LARS) algorithm; see Efron et al. [2004].

1.1.5 Finite-sample properties of ridgeless and ridge

This section presents finite-sample expressions and bounds for the bias, variance, MSE and MPR of the LSE, ridgeless and ridge estimators. The main underlying assumption is that linear model (1.1) satisfies the typical regression condition $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \mathbf{0}$. Furthermore, we work with a fixed design matrix $\boldsymbol{X}$ (or equivalently we work conditionally on $\boldsymbol{X}$).

The next proposition derives the bias, MSE and MPR of the LSE when it is well-defined, that is, when $\mathrm{Rank}(\boldsymbol{X}) = p$, which implies $p \leq n$.

Proposition 1.1.6 (Finite-sample properties of LSE (fixed design))**.** *Assume that the linear model (1.1) holds with* $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \mathbf{0}$*. Then, for a fixed design matrix such that* $\mathrm{Rank}(\boldsymbol{X}) = p$*:*

(i) The LSE is unbiased: $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{ls}] = \boldsymbol{\theta}_0$.

(ii) The variance of the LSE is given by

$$\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{V}\mathrm{ar}[\boldsymbol{\varepsilon}_0]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Further let $\mathbb{V}\mathrm{ar}[\boldsymbol{\varepsilon}_0] = \sigma^2\boldsymbol{I}$ *with* $\sigma > 0$*. Then:*

(iii) $\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

(iv) The LSE is the best linear unbiased estimator, in the sense that $\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\theta}}_n] - \mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}]$ *is positive semi-definite for any other unbiased linear estimator* $\tilde{\boldsymbol{\theta}}_n$*, i.e.,* $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{A}\boldsymbol{y}$ *for some* $\boldsymbol{A} \in \mathbb{R}^{p\times n}$.

(v) *The MSE of the LSE is given by:* $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_n^{ls}, \boldsymbol{\theta}_0) = \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{1}{\lambda_j}$, *where* $\lambda_1 \geq \ldots \geq \lambda_p > 0$ *are the eigenvalues of* $\boldsymbol{X}'\boldsymbol{X}/n$. *Therefore,*

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_n^{ls}, \boldsymbol{\theta}_0) \leq \frac{\sigma^2 p}{\lambda_p n}.$$

(vi) *The mean predictive risk of the LSE is given by:*

$$\mathrm{MPR}(\hat{\boldsymbol{\theta}}_n^{ls}, \boldsymbol{\theta}_0) = p\sigma^2/n.$$

Proof. (i) From the closed form expression (1.8):

$$\hat{\boldsymbol{\theta}}_n^{ls} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}_0 + \boldsymbol{\theta}_0.$$

Thus, unbiasedness follows directly since:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{ls}] - \boldsymbol{\theta}_0 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbb{E}[\boldsymbol{X}'\boldsymbol{\varepsilon}_0] = \mathbf{0}.$$

(ii) The closed form expression of $\hat{\boldsymbol{\theta}}_n^{ls}$ immediately implies the expression

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}] =& \mathbb{V}\mathrm{ar}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}_0] \\ =& (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{V}\mathrm{ar}[\boldsymbol{\varepsilon}_0]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

(iii) Das ist trivial.

(iv) A linear estimator $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{A}\boldsymbol{y}$ is unbiased if and only if $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$. Let $\boldsymbol{M} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, and notice that $(\boldsymbol{I}-\boldsymbol{M})(\boldsymbol{I}-\boldsymbol{M}) = (\boldsymbol{I}-\boldsymbol{M})$, i.e., $(\boldsymbol{I}-\boldsymbol{M})$ is idempotent. It follows that:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\theta}}_n] - \mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}] =& \sigma^2(\boldsymbol{A}\boldsymbol{A}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ =& \sigma^2(\boldsymbol{A}\boldsymbol{A}' - \boldsymbol{A}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{A}') \\ =& \sigma^2\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{A}' \\ =& \sigma^2[\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{M})][\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{M})]',\end{aligned}$$

which is positive semi definite.

(v) Using the linearity of the Trace operator and the SVD decomposition of $\boldsymbol{X}$ in Definition 18:

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0\|_2^2] =& \mathbb{E}[\mathrm{Trace}((\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0)')] \\ =& \mathrm{Trace}(\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}]) \\ =& \frac{\sigma^2}{n}\mathrm{Trace}((\boldsymbol{X}'\boldsymbol{X}/n)^{-1}) \\ =& \frac{\sigma^2}{n}\mathrm{Trace}(\boldsymbol{V}'(\boldsymbol{S}'\boldsymbol{S}/n)^{-1}\boldsymbol{V}) = \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{1}{\lambda_j}.\end{aligned}$$

(vi) Simple computations give

$$\begin{aligned}\mathbb{E}[\|\mathrm{lm}(\boldsymbol{X},\hat{\boldsymbol{\theta}}_n^{ls}) - \mathrm{lm}(\boldsymbol{X},\boldsymbol{\theta}_0)\|_2^2/n] \\ =&\mathbb{E}[\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0)\|_2^2/n] \\ =&\mathbb{E}[\mathrm{Trace}((\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0)'\boldsymbol{X}'\boldsymbol{X}/n(\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0))] \\ =&\mathbb{E}[\mathrm{Trace}((\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n^{ls} - \boldsymbol{\theta}_0)'\boldsymbol{X}'\boldsymbol{X}/n)] \\ =&\mathrm{Trace}(\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\theta}}_n^{ls}]\boldsymbol{X}'\boldsymbol{X}/n) = \sigma^2 p/n.\end{aligned}$$

□

This proposition shows that the LSE's accuracy decreases:

- as the variance σ^2 of the error term increases;
- as the number of predictors per observation p/n increases;
- as the "degree of singularity" of the design matrix $1/\lambda_p$ increases.

We now aim to drop the requirement that $\mathrm{Rank}(\boldsymbol{X}) = p$, to allow for high-dimensional settings where:

- $n < p$, or even
- $\mathrm{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']) = r_0$, not necessarily equals to p.

This last case We next show that, when $\mathrm{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']) < p$, the typical linear regression condition $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \boldsymbol{0}$ no longer identifies a unique estimand $\boldsymbol{\theta}_0$.

Proposition 1.1.7. *Given random variables $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$, let $\varepsilon(\boldsymbol{\theta}) := y - \boldsymbol{x}'\boldsymbol{\theta}$ for every $\boldsymbol{\theta} \in \mathbb{R}^p$. The set*

$$S := \{\boldsymbol{\theta}_0 \in \mathbb{R}^p \ : \ y = \boldsymbol{x}'\boldsymbol{\theta}_0 + \varepsilon(\boldsymbol{\theta}_0),\ \mathbb{E}[\boldsymbol{x}\varepsilon(\boldsymbol{\theta}_0)] = \boldsymbol{0}\}$$

is either empty or $S = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^+\mathbb{E}[\boldsymbol{x}y] + \mathrm{Ker}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$.

Proof. The fact that, in some cases, set S can be empty is obvious. Moreover, since

$$S = \{\boldsymbol{\theta}_0 \in \mathbb{R}^p \ : \ \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']\boldsymbol{\theta}_0 = \boldsymbol{E}[\boldsymbol{x}y]\},$$

then $\mathbb{E}[\boldsymbol{x}y] \in \mathrm{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$ when S is not empty. In this case, it follows from Theorem 2.1.5 that $S = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^+\mathbb{E}[\boldsymbol{x}y] + \mathrm{Ker}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$. □

However, when S is not-empty, there is an element of $S \cap \mathrm{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$ that is well-defined when $\mathrm{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']) < p$, and it is equal to $\boldsymbol{\theta}_0$ by construction when $\mathrm{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']) = p$. We define it as follows.

Definition 19 (Ridgeless estimand). The *ridgeless estimand* is defined as the vector $\boldsymbol{\theta}_0^{rl} \in \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$ given by $\boldsymbol{\theta}_0^{rl} := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^+\mathbb{E}[\boldsymbol{x}y]$.[6]

We can now extend Proposition 1.1.6 to the fixed design setting where $\text{Rank}(\boldsymbol{X}) \leq p$.

Proposition 1.1.8 (Finite-sample properties of ridgeless (fixed design))**.** *Assume that the linear model (1.1) holds with* $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \boldsymbol{0}$ *and denote* $r_0 := \text{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$*. Then, for a fixed design matrix:*

(i) $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\boldsymbol{\theta}_0^{rl}$*. If* $\text{Range}(\boldsymbol{X}') = \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$*, which implies* $n \geq r_0$*, then the ridgeless estimator is unbiased:*

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \boldsymbol{\theta}_0^{rl}.$$

(ii) The variance of the ridgeless estimator is given by

$$\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] = \boldsymbol{X}^+\,\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl})](\boldsymbol{X}^+)',$$

where $\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl}) := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_0^{rl}$*.*

Further let $\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl})] = \sigma^2\boldsymbol{I}$ *with* $\sigma > 0$*, and define* $r := \text{Rank}(\boldsymbol{X}) \leq \min\{n, p\}$*. Then:*

(iii) $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^+$*.*

(iv) The MSE of the ridgeless estimator is given by:

$$\text{MSE}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = \frac{\sigma^2}{n}\sum_{j=1}^{r}\frac{1}{\lambda_j} + \left\|\mathbf{P}_{\text{Ker}(\boldsymbol{X})}\,\boldsymbol{\theta}_0^{rl}\right\|_2^2,$$

where $\lambda_1 \geq \ldots \geq \lambda_r > 0$ *are the positive eigenvalues of* $\boldsymbol{X}'\boldsymbol{X}/n$*.*

(v) The mean predictive risk of the ridgeless estimator is given by:

$$\text{MPR}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = r\sigma^2/n.$$

(vi) If $\text{Range}(\boldsymbol{X}') = \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$*, we have*

$$\text{MSE}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = \frac{\sigma^2}{n}\sum_{j=1}^{r_0}\frac{1}{\lambda_j} \leq \frac{\sigma^2 r_0}{\lambda_{r_0} n}.$$

and

$$\text{MPR}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = r_0\sigma^2/n.$$

[6]The result that $\boldsymbol{\theta}_0^{rl} \in \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$ follows from the identity $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^+ = \mathrm{P}_{\text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])}\,\mathbb{E}[\boldsymbol{x}\boldsymbol{x}']^+$. Notice that we used the ridgeless estimand in Section 1.1.3.

Proof. (i) Using Proposition 1.1.7, we have $\mathbb{E}[\boldsymbol{X}\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl})] = \mathbf{0}$, which implies $\mathbb{E}[\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl})] = \mathbf{0}$ under a (non-trivial) fixed design. Simple computations then give

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \boldsymbol{X}^+\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X}^+\boldsymbol{X}\boldsymbol{\theta}_0^{rl} + \boldsymbol{X}^+\mathbb{E}[\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl})] = \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\boldsymbol{\theta}_0^{rl}.$$

If $\text{Range}(\boldsymbol{X}') = \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, we conclude that $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \boldsymbol{\theta}_0^{rl}$ since $\boldsymbol{\theta}_0^{rl} \in \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$.

(ii) The closed-form expression of $\hat{\boldsymbol{\theta}}^{rl}$ immediately implies

$$\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] = \boldsymbol{X}^+\,\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0^{rl})](\boldsymbol{X}^+)'.$$

(iii) It follows since $\boldsymbol{X}^+(\boldsymbol{X}^+)' = (\boldsymbol{X}'\boldsymbol{X})^+$.

(iv) Using the fact that $\text{Rank}(\boldsymbol{X}) = r$:

$$\text{Trace}(\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}]) = \frac{\sigma^2}{n}\,\text{Trace}((\boldsymbol{X}'\boldsymbol{X}/n)^+) = \frac{\sigma^2}{n}\sum_{j=1}^{r}\frac{1}{\lambda_j}.$$

Moreover,

$$\text{Bias}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = (\mathbf{P}_{\text{Range}(\boldsymbol{X}')} - \boldsymbol{I})\boldsymbol{\theta}_0^{rl} = -\,\mathbf{P}_{\text{Ker}(\boldsymbol{X})}\,\boldsymbol{\theta}_0^{rl}.$$

The result then follows using Proposition 1.0.3.

(v) Proposition 1.0.1 and $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\boldsymbol{\theta}_0^{rl}$ imply $\text{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0^{rl}) = \boldsymbol{X}\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}]$. Therefore:

$$\begin{aligned}
&\mathbb{E}[\|\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^{rl}) - \text{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0^{rl})\|_2^2/n] \\
&\quad= \mathbb{E}[\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])\|_2^2/n] \\
&\quad= \mathbb{E}[\text{Trace}\{(\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])'\boldsymbol{X}'\boldsymbol{X}/n(\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])\}] \\
&\quad= \mathbb{E}[\text{Trace}\{(\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])(\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])'\boldsymbol{X}'\boldsymbol{X}/n\}] \\
&\quad= \text{Trace}(\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}]\boldsymbol{X}'\boldsymbol{X}/n) \\
&\quad= \sigma^2/n\,\text{Trace}[(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}'\boldsymbol{X}] = \sigma^2/n\,\text{Trace}(\boldsymbol{X}^+\boldsymbol{X}),
\end{aligned}$$

where the last equality follows from the identity $(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}' = \boldsymbol{X}^+$. Finally, considering the spectral decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$ in Definition 18, we obtain:

$$\begin{aligned}
\sigma^2/n\,\text{Trace}(\boldsymbol{X}^+\boldsymbol{X}) &= \sigma^2/n\,\text{Trace}(\boldsymbol{V}(\boldsymbol{S}')^+\boldsymbol{S}\boldsymbol{V}') \\
&= \sigma^2/n\,\text{Trace}\begin{bmatrix} \boldsymbol{I}_r & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix} \\
&= \sigma^2 r/n.
\end{aligned}$$

(vi) If $\text{Range}(\boldsymbol{X}') = \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, then $\text{Rank}(\boldsymbol{X}) = r_0$ and $\text{Ker}(\boldsymbol{X}) = \text{Ker}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$. Therefore $\text{Bias}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = \mathbf{0}$ as $\boldsymbol{\theta}_0^{rl} \in \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, and we have

$$\text{MSE}(\boldsymbol{\theta}_n^{rl}, \boldsymbol{\theta}_0^{rl}) = \frac{\sigma^2}{n} \sum_{j=1}^{r_0} \frac{1}{\lambda_j}.$$

□

Proposition 1.1.9 (Finite-sample properties of ridge (fixed design))**.** *Assume that the linear model (1.1) holds with* $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \mathbf{0}$*. Denote* $r_0 := \text{Rank}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$*,* $\lambda > 0$ *and* $\boldsymbol{Q}(\lambda) := (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}$*. Then, for a fixed design matrix:*

(i) The ridge estimator is biased: $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^r(\lambda)] = \boldsymbol{Q}(\lambda)\boldsymbol{\theta}_0^{rl}$*.*

(ii) The variance of the ridge estimator is given by

$$\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)] = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\,\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}_0]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}.$$

Further let $\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}_0] = \sigma^2 \boldsymbol{I}$ *with* $\sigma > 0$*. Then:*

(iii) $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)] = \sigma^2(\boldsymbol{X}'\boldsymbol{X}+\lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\lambda \boldsymbol{I})^{-1}$*. Moreover,* $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] - \mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)]$ *is positive definite.*

(iv) The MSE of the ridge estimator is given by:

$$\text{MSE}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) = \frac{\sigma^2}{n} \sum_{j=1}^{r} \frac{\lambda_j}{(\lambda_j + \lambda/n)^2} + \left\| [\boldsymbol{I} - \boldsymbol{Q}(\lambda)]\boldsymbol{\theta}_0^{rl} \right\|_2^2,$$

where $\lambda_1 \geq \ldots \geq \lambda_r > 0$ *are the positive eigenvalues of* $\boldsymbol{X}'\boldsymbol{X}/n$*.*

(v) The mean predictive risk of the ridge estimator is given by:

$$\text{MPR}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) = \frac{\sigma^2}{n} \sum_{j=1}^{r} \frac{\lambda_j^2}{(\lambda_j + \lambda/n)^2}.$$

(vi) If $\text{Range}(\boldsymbol{X}') = \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$*, we have*

$$\lim_{\lambda \to 0} \text{MSE}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) = \frac{\sigma^2}{n} \sum_{j=1}^{r_0} \frac{1}{\lambda_j} \leq \frac{\sigma^2 r_0}{\lambda_{r_0} n},$$

and

$$\lim_{\lambda \to 0} \text{MPR}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) = r_0 \sigma^2 / n.$$

Proof. (i) Using the link between ridge and ridgeless estimators in identity (1.14) we have, with $\boldsymbol{\theta}_0^{rl}$ as defined in (19):

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^r(\lambda)] = \boldsymbol{Q}(\lambda)\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \boldsymbol{Q}(\lambda)\,\mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\boldsymbol{\theta}_0^{rl}.$$

The result then follows from $\boldsymbol{Q}(\lambda)\,\mathbf{P}_{\text{Range}(\boldsymbol{X}')} = \boldsymbol{Q}(\lambda)$.

(ii) The closed-form expression of $\hat{\boldsymbol{\theta}}^r$ immediately implies

$$\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)] = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\,\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}_0]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}.$$

(iii) The expression follows trivially from the previous item. To show that $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] - \mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)]$ is positive definite, consider the spectral decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$ in Definition 18. Since $\text{Rank}(\boldsymbol{X}) = r$, we have $\boldsymbol{X}'\boldsymbol{X}/n = \sum_{j=1}^r \lambda_j \boldsymbol{v}_j\boldsymbol{v}_j'$, where $\lambda_j = s_j^2/n$ for $j = 1, \ldots, r$. It follows $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] = \frac{\sigma^2}{n}\sum_{j=1}^r \frac{1}{\lambda_j}\boldsymbol{v}_j\boldsymbol{v}_j'$. Instead,

$$\begin{aligned}\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)] =& \sigma^2\boldsymbol{V}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}' \\ =& \frac{\sigma^2}{n}\sum_{j=1}^r \frac{\lambda_j}{(\lambda_j + \lambda/n)^2}\boldsymbol{v}_j\boldsymbol{v}_j'.\end{aligned}$$

The result then follows using that, for $j = 1, \ldots, r$,

$$1/\lambda_j > \frac{\lambda_j}{(\lambda_j + \lambda/n)^2}.$$

(iv) Using the linearity of the Trace and the fact that $\boldsymbol{V}$ is orthogonal:

$$\begin{aligned}\text{Trace}(\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^r(\lambda)]) =& \text{Trace}\left(\frac{\sigma^2}{n}\sum_{j=1}^r \frac{\lambda_j}{(\lambda_j + \lambda/n)^2}\boldsymbol{v}_j\boldsymbol{v}_j'\right) \\ =& \frac{\sigma^2}{n}\sum_{j=1}^r \frac{\lambda_j}{(\lambda_j + \lambda/n)^2}.\end{aligned}$$

Moreover, $\text{Bias}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) = [\boldsymbol{Q}(\lambda) - \boldsymbol{I}]\boldsymbol{\theta}_0^{rl}$. The result then follows using Proposition 1.0.3.

(v) In Proposition 1.1.8 we obtained $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}] = \mathbf{P}_{\text{Range}(\boldsymbol{X}')}\,\boldsymbol{\theta}_0^{rl}$. Proposition 1.0.1 thus implies $\text{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0^{rl}) = \boldsymbol{X}\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}]$, and so we can write

$$\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^r(\lambda)) - \text{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0^{rl}) = \boldsymbol{X}(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}]).$$

Therefore:

$$\begin{aligned}
\text{MPR}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) &= \mathbb{E}[\|\text{lm}(\boldsymbol{X}, \hat{\boldsymbol{\theta}}_n^r(\lambda)) - \text{lm}(\boldsymbol{X}, \boldsymbol{\theta}_0^{rl})\|_2^2/n] \\
&= \mathbb{E}[\text{Trace}\{(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])'\boldsymbol{X}'\boldsymbol{X}/n(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])\}] \\
&= \mathbb{E}[\text{Trace}\{\boldsymbol{X}'\boldsymbol{X}/n(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])'\}] \\
&= \text{Trace}\left\{\boldsymbol{X}'\boldsymbol{X}/n\mathbb{E}\left[(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])(\boldsymbol{Q}(\lambda)\hat{\boldsymbol{\theta}}_n^{rl} - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}])'\right]\right\}.
\end{aligned}$$

Let $\boldsymbol{Q} := \boldsymbol{Q}(\lambda)$, $\boldsymbol{E}^{rl} := \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{rl}]$ and $\boldsymbol{V}^{rl} := \mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}]$. Then, the expected value inside the Trace reads:

$$\begin{aligned}
&\mathbb{E}\left\{\left[\hat{\boldsymbol{\theta}}_n^{rl} - \boldsymbol{E}^{rl} - (\boldsymbol{I} - \boldsymbol{Q})\hat{\boldsymbol{\theta}}_n^{rl}\right]\left[\hat{\boldsymbol{\theta}}_n^{rl} - \boldsymbol{E}^{rl} - (\boldsymbol{I} - \boldsymbol{Q})\hat{\boldsymbol{\theta}}_n^{rl}\right]'\right\} \\
&= \boldsymbol{V}^{rl} + (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{V}^{rl}(\boldsymbol{I} - \boldsymbol{Q})' - (\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{V}^{rl} - (\boldsymbol{I} - \boldsymbol{Q})(\boldsymbol{I} - \boldsymbol{Q})' \\
&= \boldsymbol{Q}\boldsymbol{V}^{rl}\boldsymbol{Q}'.
\end{aligned}$$

Using $\mathbb{V}\text{ar}[\hat{\boldsymbol{\theta}}_n^{rl}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^+$ and the SVD decomposition of $\boldsymbol{X}$ in Definition 18, we conclude:

$$\begin{aligned}
&\text{MPR}(\hat{\boldsymbol{\theta}}_n^r(\lambda), \boldsymbol{\theta}_0^{rl}) \\
&= \sigma^2 \,\text{Trace}\left[\boldsymbol{X}'\boldsymbol{X}/n\boldsymbol{Q}(\lambda)(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{Q}(\lambda)'\right] \\
&= \sigma^2 \,\text{Trace}\left[\boldsymbol{X}'\boldsymbol{X}/n(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^+\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\right] \\
&= \sigma^2/n \,\text{Trace}\left[\boldsymbol{S}'\boldsymbol{S}/n(\boldsymbol{S}'\boldsymbol{S}/n + \lambda/n\boldsymbol{I})^{-1}\boldsymbol{S}'\boldsymbol{S}/n(\boldsymbol{S}'\boldsymbol{S})^+\boldsymbol{S}'\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S}/n + \lambda/n\boldsymbol{I})^{-1}\right] \\
&= \frac{\sigma^2}{n}\sum_{j=1}^{r}\frac{\lambda_j^2}{(\lambda_j + \lambda/n)^2}.
\end{aligned}$$

(vi) If $\text{Range}(\boldsymbol{X}') = \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, then $\text{Rank}(\boldsymbol{X}) = r_0$ and $\text{Ker}(\boldsymbol{X}) = \text{Ker}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$. Therefore, since

$$\lim_{\lambda\to 0}\boldsymbol{Q}(\lambda) = \lim_{\lambda\to 0}[(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}']\boldsymbol{X} = \mathbf{P}_{\text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])},$$

and $\boldsymbol{\theta}_0^{rl} \in \text{Range}(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])$, we obtain

$$\lim_{\lambda\to\infty}\frac{\sigma^2}{n}\sum_{j=1}^{r_0}\left\{\frac{\lambda_j}{(\lambda_j + \lambda/n)^2} + \left\|[\boldsymbol{I} - \boldsymbol{Q}(\lambda)]\boldsymbol{\theta}_0^{rl}\right\|_2^2\right\} = \frac{\sigma^2}{n}\sum_{j=1}^{r_0}\frac{1}{\lambda_j}.$$

□

The next proposition shows that there are penalty parameter values for which the MSE of ridge is lower than the MSE of ridgeless.

Proposition 1.1.10. *Assume that the linear model (1.1) holds with* $\mathbb{E}[\boldsymbol{x}\varepsilon_0] = \mathbf{0}$ *and* $\mathbb{V}\text{ar}[\boldsymbol{\varepsilon}_0] = \sigma^2\boldsymbol{I}$ *for* $\sigma > 0$*. Then, for a fixed design matrix* $\boldsymbol{X}$*, there exists* $\lambda^* > 0$ *such that*

$$\text{MSE}(\hat{\boldsymbol{\theta}}_n^r(\lambda^*), \boldsymbol{\theta}_0^{rl}) < \text{MSE}(\hat{\boldsymbol{\theta}}_n^{rl}, \boldsymbol{\theta}_0^{rl}).$$

Proof. See Farebrother [1976]. □

1.1.6 Finite sample properties of lasso

In this section, we study the finite sample properties of the lasso estimator under a fixed design matrix $\boldsymbol{X}$. Given the lack of a closed form expression for the lasso estimator, we do not have access to closed form expressions for its bias and variance. Therefore, instead of deriving its MSE and MPR, we find concentration inequalities for its estimation risk $\|\hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0\|_2^2$ and predictive risk $\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0)\|_2^2/n$.

Before doing that, we first show some auxiliary properties satisfied by any lasso solution. Given a set C, let $|C|$ denote its cardinality, i.e., the number of elements in C, and consider the index set $S \subset \{1, \ldots, p\}$ with complementary index set $S^c = \{1, \ldots, p\} \setminus S$. We use the notation $\boldsymbol{v}_S = [v_i]_{i \in S} \in \mathbb{R}^{|S|}$ for the subvector of $\boldsymbol{v} \in \mathbb{R}^p$ with entries indexed by S. Further define, for some $\alpha \geq 1$, the set

$$\mathcal{C}_\alpha(S) := \{\boldsymbol{v} \in \mathbb{R}^p \ : \ \|\boldsymbol{v}_{S^c}\|_1 \leq \alpha \|\boldsymbol{v}_S\|_1\}.$$

In words, $\mathcal{C}_\alpha(S)$ is the set of vectors in $\mathbb{R}^p$ whose subvector in S^c has size smaller or equal to α times the size of the subvector in S, where the size of vectors is measured using the l_1-norm. Finally, consider the following definition.

Definition 20 (Support of a vector). The *support* of vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is defined as

$$\text{Supp}(\boldsymbol{\theta}) := \{j \in \{1, \ldots, p\} \ : \ \theta_j \neq 0\}.$$

The next lemma shows that, for an appropriate choice of the penalty parameter, the lasso estimator satisfies some basic inequalities and has an estimation error contained in $\mathcal{C}_\alpha(S)$ for some $\alpha \geq 1$ and some index set S.

Lemma 1.1.5 (Auxiliary properties of lasso)**.** *Suppose that the linear model (1.1) holds. If* $\lambda \geq 2\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty > 0$*, then any lasso solution* $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ *satisfies:*

(i) The predictive risk bound

$$\text{PR}(\hat{\boldsymbol{\theta}}_n^l(\lambda), \boldsymbol{\theta}_0) \leq 12\lambda \|\boldsymbol{\theta}_0\|_1 . \tag{1.19}$$

(ii) An estimation error $\hat{\boldsymbol{\eta}} := \hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0 \in \mathcal{C}_3(\mathrm{Supp}(\boldsymbol{\theta}_0))$ *such that*

$$\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq 3\sqrt{s_0}\lambda\|\hat{\boldsymbol{\eta}}\|_2. \tag{1.20}$$

Proof. Under the linear model (1.1), we have for any $\boldsymbol{\theta} \in \mathbb{R}^p$:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 = \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\theta}'\boldsymbol{X}\boldsymbol{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}_0'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} - 2\boldsymbol{\varepsilon}_0'\boldsymbol{X}\boldsymbol{\theta}.$$

Since $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ is a lasso solution,

$$0 \leq \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_2^2 + \lambda\|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1 \leq \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2 + \lambda\|\boldsymbol{\theta}_0\|_1,$$

which holds if and only if

$$\frac{1}{2n}\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2 \leq \boldsymbol{\varepsilon}_0'\boldsymbol{X}\hat{\boldsymbol{\eta}}/n + \lambda(\|\boldsymbol{\theta}_0\|_1 - \|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1). \tag{1.21}$$

(i) By Hölder inequality,

$$\boldsymbol{\varepsilon}_0'\boldsymbol{X}\hat{\boldsymbol{\eta}}/n \leq |\boldsymbol{\varepsilon}_0'\boldsymbol{X}\hat{\boldsymbol{\eta}}/n| \leq \|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty \|\hat{\boldsymbol{\eta}}\|_1.$$

Thus, using the choice $\lambda \geq 2\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty$ in (1.21) yields

$$0 \leq \frac{1}{2n}\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2 \leq \lambda/2\|\hat{\boldsymbol{\eta}}\|_1 + \lambda(\|\boldsymbol{\theta}_0\|_1 - \|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1). \tag{1.22}$$

Using the triangle inequality

$$\|\hat{\boldsymbol{\eta}}\|_1 \leq \|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1 + \|\boldsymbol{\theta}_0\|_1, \tag{1.23}$$

we further obtain

$$0 \leq \lambda/2(\|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1 + \|\boldsymbol{\theta}_0\|_1) + \lambda(\|\boldsymbol{\theta}_0\|_1 - \|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1),$$

which, using that $\lambda > 0$, implies $\|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1 \leq 3\|\boldsymbol{\theta}_0\|_1$. Substituting this result into the triangle inequality (1.23) yields

$$\|\hat{\boldsymbol{\eta}}\|_1 \leq 4\|\boldsymbol{\theta}_0\|_1. \tag{1.24}$$

Moreover, again by the triangle inequality,

$$\|\boldsymbol{\theta}_0\|_1 = \|(\boldsymbol{\theta}_0 + \hat{\boldsymbol{\eta}}) - \hat{\boldsymbol{\eta}}\|_1 \leq \|\boldsymbol{\theta}_0 + \hat{\boldsymbol{\eta}}\|_1 + \|\hat{\boldsymbol{\eta}}\|_1,$$

which implies:

$$\|\boldsymbol{\theta}_0 + \hat{\boldsymbol{\eta}}\|_1 \geq \|\boldsymbol{\theta}_0\|_1 - \|\hat{\boldsymbol{\eta}}\|_1. \tag{1.25}$$

Using (1.24) and (1.25) in the basic inequality (1.22), we obtain

$$\begin{aligned}\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq& \lambda\|\hat{\boldsymbol{\eta}}\|_1 + 2\lambda(\|\boldsymbol{\theta}_0\|_1 - \|\boldsymbol{\theta}_0 + \hat{\boldsymbol{\eta}}\|_1)\\ \leq& 3\lambda\|\hat{\boldsymbol{\eta}}\|_1 \leq 12\lambda\|\boldsymbol{\theta}_0\|_1.\end{aligned}$$

(ii) Let $S_0 := \text{Supp}(\boldsymbol{\theta}_0)$. Using that $\boldsymbol{\theta}_{0S_0^c} = \mathbf{0}$, we have

$$\|\boldsymbol{\theta}_0\|_1 - \|\hat{\boldsymbol{\theta}}_n^l(\lambda)\|_1 = \|\boldsymbol{\theta}_{0S_0}\|_1 - \|\boldsymbol{\theta}_{0S_0} + \hat{\boldsymbol{\eta}}_{S_0}\|_1 - \|\hat{\boldsymbol{\eta}}_{S_0^c}\|_1. \tag{1.26}$$

Substituting (1.26) into the basic inequality (1.22) yields:

$$0 \leq \|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq \lambda\|\hat{\boldsymbol{\eta}}\|_1 + 2\lambda(\|\boldsymbol{\theta}_{0S_0}\|_1 - \|\boldsymbol{\theta}_{0S_0} + \hat{\boldsymbol{\eta}}_{S_0}\|_1 - \|\hat{\boldsymbol{\eta}}_{S_0^c}\|_1). \tag{1.27}$$

By the triangle inequality,

$$\|\boldsymbol{\theta}_{0S_0}\|_1 = \|\boldsymbol{\theta}_{0S_0} + \hat{\boldsymbol{\eta}}_{S_0} - \hat{\boldsymbol{\eta}}_{S_0}\|_1 \leq \|\boldsymbol{\theta}_{0S_0} + \hat{\boldsymbol{\eta}}_{S_0}\|_1 + \|\hat{\boldsymbol{\eta}}_{S_0}\|_1.$$

Therefore, using the decomposition $\|\hat{\boldsymbol{\eta}}\|_1 = \|\hat{\boldsymbol{\eta}}_{S_0}\|_1 + \|\hat{\boldsymbol{\eta}}_{S_0^c}\|_1$, (1.27) reads

$$\begin{aligned} 0 \leq \|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq & \lambda\|\hat{\boldsymbol{\eta}}\|_1 + 2\lambda(\|\hat{\boldsymbol{\eta}}_{S_0}\|_1 - \|\hat{\boldsymbol{\eta}}_{S_0^c}\|_1) \\ = & \lambda(3\|\hat{\boldsymbol{\eta}}_{S_0}\|_1 - \|\hat{\boldsymbol{\eta}}_{S_0^c}\|_1), \end{aligned}$$

which implies that $\hat{\boldsymbol{\eta}} \in \mathcal{C}_3(\text{Supp}(\boldsymbol{\theta}_0))$. Finally, using the relation between the l_1- and the l_2-norm ($\|\boldsymbol{v}\|_1 \leq \sqrt{s}\,\|\boldsymbol{v}\|_2$ for every $\boldsymbol{v} \in \mathbb{R}^s$), we conclude that

$$\begin{aligned} \|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq & \lambda(3\|\hat{\boldsymbol{\eta}}_{S_0}\|_1 - \|\hat{\boldsymbol{\eta}}_{S_0^c}\|_1) \leq 3\lambda\|\hat{\boldsymbol{\eta}}_{S_0}\|_1 \leq 3\sqrt{s_0}\lambda\|\hat{\boldsymbol{\eta}}_{S_0}\|_2 \\ \leq & 3\sqrt{s_0}\lambda\|\hat{\boldsymbol{\eta}}\|_2. \end{aligned}$$

□

We derive the main properties of lasso under the following *restricted eigenvalue* condition on the design matrix, which leverages the result that for $\lambda \geq 2\,\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty$, the estimation error of lasso $\hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0 \in \mathcal{C}_3(\text{Supp}(\boldsymbol{\theta}_0))$.

Assumption 1 (Restricted eigenvalue condition). The design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is such that for all $\boldsymbol{\eta} \in \mathcal{C}_3(\text{Supp}(\boldsymbol{\theta}_0))$ there exists $\kappa > 0$ for which:

$$\|\boldsymbol{X}\boldsymbol{\eta}\|_2^2/n \geq \kappa\,\|\boldsymbol{\eta}\|_2^2,$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is the coefficient of linear model (1.1).

In the next proposition we derive bounds on the squared l_2 estimation risk and predictive risk. We then provide intuition on why the restricted eigenvalue condition is required.

Theorem 1.1.6. *Suppose that the linear model (1.1) holds and that Assumption 1 holds. Let* $s_0 := |\,\text{Supp}(\boldsymbol{\theta}_0)| \leq p$. *Then, any lasso solution* $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ *with* $\lambda \geq 2\,\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty > 0$ *satisfies:*

(i) The estimation risk bound

$$\|\hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0\|_2^2 \leq \frac{9}{\kappa^2} s_0 \lambda^2. \tag{1.28}$$

(ii) The predictive risk bound

$$\mathrm{PR}(\hat{\boldsymbol{\theta}}_n^l(\lambda), \boldsymbol{\theta}_0) \leq \frac{9}{\kappa} s_0 \lambda^2. \tag{1.29}$$

Proof. In Lemma (1.1.5), we obtained Inequality (1.20), which reads

$$\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq 3\sqrt{s_0}\lambda\|\hat{\boldsymbol{\eta}}\|_2,$$

where $\hat{\boldsymbol{\eta}} := \hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0 \in \mathcal{C}_3(\mathrm{Supp}(\boldsymbol{\theta}_0))$.

(i) Using Assumption 1 on the left hand side of Inequality (1.20) yields

$$\kappa\|\hat{\boldsymbol{\eta}}\|_2^2 \leq 3\sqrt{s_0}\lambda\|\hat{\boldsymbol{\eta}}\|_2.$$

If $\|\hat{\boldsymbol{\eta}}\|_2 > 0$, the result follows by dividing both sides of the inequality by $\|\hat{\boldsymbol{\eta}}\|_2$. If instead $\|\hat{\boldsymbol{\eta}}\|_2 = 0$, the result is trivially obtained.

(ii) Using Assumption 1 on the right hand side of Inequality (1.20) yields,

$$\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2^2/n \leq 3\sqrt{s_0}\lambda\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2/\sqrt{n\kappa}.$$

If $\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2 > 0$, the result follows by dividing both sides of the inequality by $\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2/\sqrt{n}$. If instead $\|\boldsymbol{X}\hat{\boldsymbol{\eta}}\|_2 = 0$, the result is trivially obtained.

□

The sparsity parameter $s_0 = |\,\mathrm{Supp}(\boldsymbol{\theta}_0)|$ plays a major role in the bounds of Theorem 1.1.6. We say that $\boldsymbol{\theta}_0$ is hard sparse if it has some zero entries. More formally:

Definition 21 (Hard sparsity). Coefficient $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is *hard sparse* if $s_0 := |\,\mathrm{Supp}(\boldsymbol{\theta}_0)| < p$.

In high-dimensional regimes, hard sparsity is typically imposed as an identifying condition for $\boldsymbol{\theta}_0$. Consider an asymptotic regime where

$$\lim_{n\to\infty} p/n = K > 0, \quad \text{and} \quad \lim_{p\to\infty} s_0/p = s_\infty \in (0, 1].$$

Then, $s_0 \to \infty$ as $n \to \infty$. In this setting, the lasso converges to $\boldsymbol{\theta}_0$ only if κ and λ compensate for the divergence of s_0, i.e., $\lim_{n\to\infty} s_0\lambda/\kappa^2 = 0$. Notice however that $2\,\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty$, the lower bound for λ in Theorem 1.1.6, is monotonically non decreasing as we add columns to $\boldsymbol{X}$. Moreover, intuition suggests that Assumption 1 with a large κ is an increasingly more restrictive assumption as $p \to \infty$.

Remark 6. It is possible to extend the results in Lemma 1.1.5 and Theorem 1.1.6 using a milder restriction than hard sparsity, called weak sparsity. This restriction formalizes the notion that $\boldsymbol{\theta}_0$ can be well approximated by means of a hard sparse vector.

Definition 22 (Weak sparsity). Coefficient $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is *weak sparse* if $\boldsymbol{\theta}_0 \in B_q(r)$ where, for $q \in [0, 1]$ and radius $r > 0$,

$$B_q(r) := \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{\theta}\|_q^q \leq r\}.$$

Setting $q = 0$ in Definition 22 recovers Definition 21 with $s = r$. For $q \in (0, 1]$, we restrict the way the ordered coefficients

$$\max_{j=1,\ldots,p} |\theta_{0j}| = \theta_{0(1)} \geq \theta_{0(2)} \geq \ldots, \geq \theta_{0(p-1)} \geq \theta_{0(p)} = \min_{j=1,\ldots,p} |\theta_{0j}|$$

decay. More precisely, if the ordered coefficients satisfy the bound $|\theta_{0j}| \leq Cj^{-\alpha}$ for some suitable $C \geq 0$ and $\alpha > 0$, then $\boldsymbol{\theta}_0 \in B_q(r)$ for a radius r that depends on C and α.

Restricted eigenvalue condition

Inequality (1.20) of Lemma (1.1.5) establishes an upper bound for the prediction risk of the lasso solution in terms of its estimation risk. Conversely, for an appropriate choice of the penalty parameter, the restricted eigenvalue condition in Assumption 1 provides an upper bound for the estimation risk of the lasso solution based on its prediction risk. These two bounds are combined to obtain the estimation and predictive risk bounds in Theorem 1.1.6.

To provide more intuition on why the restricted eigenvalue condition is needed, consider the constrained version of the lasso estimator:

$$\hat{\boldsymbol{\theta}}_n^{lc}(R) := \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \ : \ \|\boldsymbol{\theta}\|_1 \leq R \right\},$$

where the radius $R := \|\boldsymbol{\theta}_0\|_1$. With this choice, the true parameter $\boldsymbol{\theta}_0$ is feasible for the problem. Additionally, we have $L_n(\hat{\boldsymbol{\theta}}_n^{lc}(R)) \leq L_n(\boldsymbol{\theta}_0)$ where

$$L_n : \mathbb{R}^p \to \mathbb{R}; \boldsymbol{\theta} \mapsto \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 / (2n)$$

is the least squares loss function. Under mild regularity conditions, it can be shown that the loss difference $L_n(\boldsymbol{\theta}_0) - L_n(\hat{\boldsymbol{\theta}}_n^{lc}(R))$ decreases as the sample size n increases. Under what conditions does this imply that the estimation risk, $\|\hat{\boldsymbol{\eta}}\|_2^2$ with $\hat{\boldsymbol{\eta}} := \hat{\boldsymbol{\theta}}_n^{lc}(R) - \boldsymbol{\theta}_0$, also decreases? Since L_n is a quadratic function,

the estimation risk will decrease if the function has positive curvature in every direction (i.e., if there are no flat regions). This occurs when the Hessian, $\nabla^2 L_n(\hat{\boldsymbol{\theta}}_n^{lc}(R)) = \boldsymbol{X}'\boldsymbol{X}/n$, has eigenvalues that are uniformly lower-bounded by a positive constant κ. This condition is equivalently expressed as

$$\|\boldsymbol{X}\boldsymbol{\eta}\|_2^2/n \geq \kappa\|\boldsymbol{\eta}\|_2^2 > 0$$

for all nonzero $\boldsymbol{\eta} \in \mathbb{R}^p$.

In the high-dimensional setting, where $p > n$, the Hessian has rank at most n, meaning that the least squares loss is flat in at least $p-n$ directions. As a result, the uniform curvature condition must be relaxed. By Lemma 1.1.5, the estimation error of lasso lies in the subset $\mathcal{C}_3(\text{Supp}(\boldsymbol{\theta}_0)) \subset \mathbb{R}^p$ for an appropriate choice of the penalty parameter (equivalently, of the constrained radius R). For this reason, we require the condition to hold only in the directions $\boldsymbol{\eta}$ that lie in $\mathcal{C}_3(\text{Supp}(\boldsymbol{\theta}_0))$, hoping that $|\text{Supp}(\boldsymbol{\theta}_0)| \leq \text{Rank}(\boldsymbol{X})$. With this adjustment, even in high-dimensional settings, a small difference in the loss function still leads to an upper bound on the difference between the lasso estimate and the true parameter.

Verifying that a given design matrix $\boldsymbol{X}$ satisfies the restricted eigenvalue condition is challenging. Developing methods to discover random design matrices that satisfy this condition with high probability remains an active area of research.

Slow rates and fast rates

Consider assuming that the error term in linear model (1.1) is sub-Gaussian with mean zero and variance proxy σ^2. It is then possible to find a choice of λ that only depends on the unknown σ and that ensures that the estimation and prediction risks are upper-bounded with high probability.

Theorem 1.1.7. *Suppose that the linear model (1.1) holds and that $\boldsymbol{\varepsilon}_0$ is a vector if independent random variables with $\varepsilon_{0i} \sim \text{sub-G}(\sigma^2)$ where variance proxy $\sigma > 0$. Further suppose that the columns of $\boldsymbol{X}$ are standardized so that $\max_{j=1,\dots,p} \|\boldsymbol{X}_j\|_2 / \sqrt{n} \leq C$ for some constant $C > 0$. Then, for all $\delta > 0$, any lasso solution $\hat{\boldsymbol{\theta}}_n^l(\lambda)$ with regularization parameter*

$$\lambda = 2C\sigma\left(\sqrt{2\ln(p)/n} + \delta\right) \tag{1.30}$$

satisfies with probability $1 - 2e^{-n\delta^2/2}$:

$$\text{PR}(\hat{\boldsymbol{\theta}}_n^l(\lambda), \boldsymbol{\theta}_0) \leq 24C \|\boldsymbol{\theta}_0\|_1 \sigma(\sqrt{2\ln(p)/n} + \delta). \tag{1.31}$$

Further suppose that Assumption 1 holds and let $s_0 := |\text{Supp}(\boldsymbol{\theta}_0)| \leq p$. Then, with probability $1 - 2e^{-n\delta^2/2}$:

(i) The estimation risk bound

$$\|\hat{\boldsymbol{\theta}}_n^l(\lambda) - \boldsymbol{\theta}_0\|_2^2 \leq \frac{72C^2\sigma^2 s_0}{\kappa^2}(2\ln(p)/n + \delta^2). \tag{1.32}$$

(ii) The predictive risk bound

$$\mathrm{PR}(\hat{\boldsymbol{\theta}}_n^l(\lambda), \boldsymbol{\theta}_0) \leq \frac{72C^2\sigma^2 s_0}{\kappa}(2\ln(p)/n + \delta^2). \tag{1.33}$$

Proof. From the union bound:

$$\begin{aligned}
\mathbb{P}\left[\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty \geq t\right] &= \mathbb{P}\left[\max_{j=1,\ldots,p} |\boldsymbol{X}_j'\boldsymbol{\varepsilon}_0/n| \geq t\right] \\
&= \mathbb{P}\left[\cup_{j=1,\ldots,p}\{|\boldsymbol{X}_j'\boldsymbol{\varepsilon}_0/n| \geq t\}\right] \\
&\leq \sum_{j=1}^{p} \mathbb{P}\left[|\boldsymbol{X}_j'\boldsymbol{\varepsilon}_0/n| \geq t\right].
\end{aligned}$$

Since $\varepsilon_{01}, \ldots, \varepsilon_{on}$ are independent random variables with sub-G(σ^2) distribution, from Proposition 2.3.8 we have that for any $t \in \mathbb{R}$:

$$\begin{aligned}
\sum_{j=1}^{p} \mathbb{P}\left[|\boldsymbol{X}_j'\boldsymbol{\varepsilon}_0/n| \geq t\right] &\leq \sum_{j=1}^{p} 2\exp\left(-\frac{t^2}{2\sigma^2\|\boldsymbol{X}_j/n\|_2^2}\right) \\
&\leq 2p\exp\left(-\frac{t^2 n}{2\sigma^2 C^2}\right).
\end{aligned}$$

Substituting $t = C\sigma\left(\sqrt{2\ln(p)/n} + \delta\right)$ we get

$$\begin{aligned}
2p\exp\left(-\frac{t^2 n}{2\sigma^2 C^2}\right) &= 2\exp(\ln(p))\exp\left(-n\delta^2/2 - \ln(p) - \delta\sqrt{2n\ln(p)}\right) \\
&= 2\exp\left(-n\delta^2/2\right)\exp\left(-\delta\sqrt{2n\ln(p)}\right) \\
&\leq 2\exp\left(-n\delta^2/2\right),
\end{aligned}$$

since $-\delta\sqrt{2n\ln(p)} < 0$. We conclude that, for all $\delta > 0$:

$$\mathbb{P}[2\|\boldsymbol{X}'\boldsymbol{\varepsilon}_0/n\|_\infty \leq 2C\sigma(\sqrt{2\ln(p)/n} + \delta)] \geq 1 - 2e^{-n\delta^2/2}.$$

Consequently, if we set $\lambda = 2C\sigma(\sqrt{2\ln(p)/n} + \delta)$, we obtain from (1.19) of Lemma 1.1.5 that (1.31) holds with probability at least $1 - 2e^{-n\delta^2/2}$. Moreover, under Assumption 1, we obtain from (1.28) and (1.29) of Theorem 1.1.6

that (1.32) and (1.33) hold with probability at least $1 - 2e^{-n\delta^2/2}$, by using the inequality:[7]

$$2\ln(p)/n + \delta^2 + 2\sqrt{2\ln(p)/n}\delta \leq 2(2\ln(p)/n + \delta^2).$$

□

As long as $n \geq 2\ln(p)$, the ratio $2\ln(p)/n$ can be significantly smaller than $\sqrt{2\ln(p)/n}$. For this reason, the bounds (1.31) and (1.33) are often referred to as the *slow rates* and *fast rates* for the prediction risk of lasso, respectively.

[7] This inequality follows from $2ab \leq a^2 + b^2$ for any two real numbers a and b.

Bibliography

Arthur Albert. Regression and the moore-penrose pseudoinverse. 1972.

Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems*, 25, 2012.

Sheldon Axler. *Linear algebra done right*. Springer Nature, 2024.

Heinz H Bauschke, Patrick L Combettes, Heinz H Bauschke, and Patrick L Combettes. *Correction to: convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Richard William Farebrother. Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–250, 1976.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2): 302–332, 2007.

Carl F Gauss. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. *sumtibus Frid. Perthes et I. H. Besser*, 1809.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143 (143):8, 2015.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Adrien-Marie Legendre. Nouvelles méthodes pour la détermination des orbites des comètes. *F. Didot*, 1805.

Eliakim H Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the american mathematical society*, 26:294–295, 1920.

Marc Nerlove et al. *Returns to scale in electricity supply*. Institute for mathematical studies in the social sciences, 1961.

Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 267–288, 1996.

Ryan J Tibshirani. The lasso problem and uniqueness. *The Electronic Journal of Statistics*, 7:1456–1490, 2013.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109:475–494, 2001.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Chapter 2

Appendix

2.1 Linear algebra

This section introduces a selection of definitions and results from linear algebra that are used in these lecture notes. A book-length exposition of linear algebra can be found in Axler [2024], among others.

Vector space

We introduce useful definitions and results for real vector spaces.

Definition 23 (Vector space). A (real) *vector space* is a set V along with an addition on V and a scalar multiplication on V with the following properties:

1. (commutativity) $u + v = v + u$ for all $u, v \in V$;
2. (associativity) $(u + v) + w = u + (v + w)$ and $(ab)v = a(bv)$ for all $u, v, w \in V$ and all $a, b \in \mathbb{R}$;
3. (additive identity) there exists an element $0 \in V$ such that $v + 0 = v$ for all $v \in V$;
4. (additive inverse) for every $v \in V$, there exists $w \in V$ such that $v + w = 0$;
5. (multiplicative identity) $1v = v$ for all $v \in V$;
6. (distributive properties) $a(u + v) = au + av$ and $(a + b)v = av + bv$ for all $a, b \in \mathbb{R}$ and all $u, v \in V$.

Definition 24 (Subspace). A subset U of a vector space V is a *subspace* of V if U is a vector space, (using the same addition and scalar multiplication as on V).

Proposition 2.1.1. *A subset U of a vector space V is a subspace of V if and only if it satisfies these conditions:*

(i) (additive identity) $0 \in U$*;*

(ii) (closed under addition) $u, v \in U$ *implies* $u + v \in U$*;*

(iii) (closed under scalar multiplication) $a \in \mathbb{R}$ *and* $u \in U$ *implies* $au \in U$*.*

Definition 25 (Linear combination). A *linear combination* of vectors $v_1, \ldots, v_n$ in vector space V with coefficients $a_1, \ldots, a_n \in \mathbb{R}$ is:

$$a_1 v_1 + \ldots, + a_n v_n.$$

Definition 26 (Span). The *span* of vectors $v_1, \ldots, v_n$ in vector space V is defined as

$$\text{span}(v_1, \ldots, v_n) := \{a_1 v_1 + \ldots + a_n v_n \ : \ a_1, \ldots, a_n \in \mathbb{R}\}.$$

Definition 27 (Linear independence). The vectors $v_1, \ldots, v_n$ in vector space V are *linearly independent* if

$$\{a_1, \ldots, a_n \in \mathbb{R} \ : \ a_1 v_1 + \ldots + a_n v_n = 0\} = \{a_1 = \ldots = a_n = 0\}.$$

Definition 28 (Linear dependence). The vectors $v_1, \ldots, v_n$ in vector space V are *linearly dependent* if they are not linearly independent.

Definition 29 (Basis). A *basis* of a vector space V is a set of vectors in V that are linearly independent and span V.

Inner products and norms

Definition 30 (Inner product). An *inner product* on a vector space V is a function that takes each ordered pair (v, u) of elements of V to a number $\langle v, u \rangle \in \mathbb{R}$ and satisfies:

1. (positivity) $\langle v, v \rangle \geq 0$ for all $v \in V$;
2. (definiteness) $\langle v, v \rangle = 0$ if and only if $v = 0$;
3. (additivity in first slot) $\langle v + u, w \rangle = \langle v, w \rangle + \langle u, w \rangle$ for all $v, u, w \in V$;
4. (homogeneity in first slot) $\langle av, u \rangle = a\langle v, u \rangle$ for all $a \in \mathbb{R}$ and all $v, u \in V$;
5. (conjugate symmetry) $\langle v, u \rangle = \langle u, v \rangle$ for all $v, u \in V$.

Proposition 2.1.2 (Basic properties of an inner product)**.** *An inner product $\langle \cdot, \cdot \rangle$ on vector space V satisfies:*

1. $\langle 0, v \rangle = \langle v, 0 \rangle$ *for every* $v \in V$.
2. $\langle v, u + w \rangle = \langle v, u \rangle + \langle v, w \rangle$ *for every* $v, u, w \in V$.
3. $\langle v, au \rangle = a\langle v, u \rangle$ *for all* $a \in \mathbb{R}$ *and all* $v, u \in V$.

Definition 31 (Orthogonal vectors). Two vectors v and u in vector space V are *orthogonal* if $\langle v, u \rangle = 0$.

Definition 32 (Orthogonal subspace). U and W are *orthogonal subspaces* of vector space V if $\langle u, w \rangle = 0$ for all $u \in U$ and all $w \in W$.

Definition 33 (Orthonormal basis). The set of vectors $\{v_1, \ldots, v_n\}$ in vector space V is an *orthonormal basis* of V if it is a basis of V such that $\langle v_i, v_j \rangle = 0$ and $\|v_i\| = 1$ for all $i, j = 1, \ldots, n$ with $i \neq j$.

Definition 34 (Norms). Given inner product $\langle \cdot, \cdot \rangle$ on vector space V, the *norm* of $v \in V$ is defined by $\|v\| := \sqrt{\langle v, v \rangle}$.

Proposition 2.1.3 (Properties of norms)**.** *For v in vector space V:*

1. $\|v\| = 0$ *if and only if* $v = 0$.
2. $\|av\| = a \|v\|$ *for all* $a \in \mathbb{R}$.

Definition 35 (Linear function). $L : V \to W$ from a vector space V to another vector space W is a *linear function* if:

(i) $T(v + u) = T(v) + T(u)$ for all $v, u \in V$;

(ii) $T(av) = aT(v)$ for all $a \in \mathbb{R}$ and $v \in V$.

Theorem 2.1.1 (Cauchy–Schwarz inequality)**.** *Suppose v and u are two vectors in vector space V. Then,*

$$|\langle v, u \rangle| \leq \|v\| \|u\| .$$

This inequality is an equality if and only if there is $a \in \mathbb{R}$ such that $v = au$.

Theorem 2.1.2 (Triangle inequality)**.** *Suppose v and u are two vectors in vector space V. Then,*

$$\|v + u\| \leq \|v\| + \|u\| .$$

This inequality is an equality if and only if there is $a \geq 0$ such that $v = au$.

Theorem 2.1.3 (Parallelogram equality)**.** *Suppose v and u are two vectors in vector space V. Then,*

$$\|v + u\|^2 + \|v - u\|^2 = 2(\|v\|^2 + \|u\|^2).$$

The Euclidean space

Definition 36 ((Real) $n-$tuple). A *(real) $n-$tuple* is a ordered list of n real numbers.

With a slight abuse of terminology, we sometimes we use the term *vector* to mean a (real) $n-$tuple.

Definition 37 (Real Euclidean space). The *real Euclidean space* of dimension n, denoted $\mathbb{R}^n$, is the set of all $n-$tuples.

Elements of a real Euclidean space are written in bold. For example, $\boldsymbol{a} \in \mathbb{R}^n$, which means $a = (a_1, \ldots, a_n)$ with $a_1, \ldots, a_n \in \mathbb{R}$.

Definition 38 (Euclidean inner product). The *Euclidean inner product* of $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^n$ is defined as $\langle \boldsymbol{v}, \boldsymbol{u} \rangle_e := \sum_{i=1}^n v_i u_i$.

Definition 39 (l_p-norm). The *l_p-norm* $\|\cdot\|_p$ on $\mathbb{R}^n$ is defined for all $\boldsymbol{v} \in \mathbb{R}^n$ as $\|\boldsymbol{v}\|_p := (\sum_{i=1}^n |v_i|^p)^{1/p}$ when $p \in [1, +\infty)$, and $\|\boldsymbol{v}\|_p := \max_{i=1}^n |v_i|$ when $p = +\infty$.

Theorem 2.1.4 (Hölder inequality)**.** *Suppose $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^p$. Then,*

$$|\langle \boldsymbol{v}, \boldsymbol{u} \rangle| \leq \|\boldsymbol{v}\|_p \|\boldsymbol{u}\|_q .$$

Matrices

Definition 40 (Matrix). An $n \times p$ *matrix* is a collection of p $n-$tuples.

The collection of all $n \times p$ matrices is denoted $\mathbb{R}^{n \times p}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, we write $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p]$ where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p \in \mathbb{R}^n$ are p $n-$tuples. Written more explicitly,

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & \ldots & A_{1,p} \\ \vdots & \ddots & \vdots \\ A_{n,1} & \ldots & A_{n,p} \end{bmatrix},$$

that is, the elements of $\boldsymbol{A}$, the $n-$tuples, are organized in columns. We denote:

- the $i,j-$th element of $\boldsymbol{A}$ by $A_{i,j}$;
- the $j-$th column $\boldsymbol{A}$ by $\boldsymbol{A}_j$;
- the $i-$th row $\boldsymbol{A}$ by $\boldsymbol{A}_{(i)}$.

Notice that a matrix in $\mathbb{R}^{n\times p}$ can be equivalently seen as a collection of n $p-$tuples, where the $p-$tuples represent the rows of the matrix.

Definition 41 (Column and row vector). A $n-$*column vector* is a $n-$tuple seen as a matrix in $\mathbb{R}^{n\times 1}$. A $n-$*row vector* is a $n-$tuple seen as a matrix in $\mathbb{R}^{1\times n}$.

Throughout these lecture notes, we denote $n-$tuples as column vectors, and use the simple notation $\boldsymbol{v}\in\mathbb{R}^n$ instead of $\boldsymbol{v}\in\mathbb{R}^{n\times 1}$.

Definition 42 (Matrix addition). The sum of two matrices of the same size is the matrix obtained by adding corresponding entries in the matrices. That is, for $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{n\times p}$, we define $\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{C}$ where $\boldsymbol{C}\in\mathbb{R}^{n\times p}$ and $C_{i,j}=A_{i,j}+B_{i,j}$ for $i=1,\ldots,n$ and $j=1,\ldots,p$.

Definition 43 (Matrix-scalar multiplication). The product of a scalar and a matrix is the matrix obtained by multiplying each entry in the matrix by the scalar That is, for $\boldsymbol{A}\in\mathbb{R}^{n\times p}$ and $a\in\mathbb{R}$, we define $a\boldsymbol{A}=\boldsymbol{B}$ where $\boldsymbol{B}\in\mathbb{R}^{n\times p}$ and $B_{i,j}=aA_{i,j}$ for $i=1,\ldots,n$ and $j=1,\ldots,p$.

Definition 44 (Matrix multiplication). Given $\boldsymbol{A}\in\mathbb{R}^{n\times p}$ and $\boldsymbol{B}\in\mathbb{R}^{p\times m}$, the product $\boldsymbol{AB}=\boldsymbol{C}$ where $\boldsymbol{C}\in\mathbb{R}^{n\times m}$ and $C_{i,j}=\sum_{r=1}^{p}A_{i,r}B_{r,j}$ for $i=1,\ldots,n$ and $j=1,\ldots,m$.

Note that we define the product of two matrices only when the number of columns of the first matrix equals the number of rows of the second matrix.

Definition 45 (Transpose of a matrix). The *transpose* of a matrix $\boldsymbol{A}\in\mathbb{R}^{n\times p}$ is the matrix $\boldsymbol{B}\in\mathbb{R}^{p\times n}$ with $j,i-$entry given by $B_{j,i}=A_{i,j}$ for $i=1,\ldots,n$ and $j=1,\ldots,p$. We denote it by $\boldsymbol{A}'$.

It follows that the Euclidean inner product between $\boldsymbol{v},\boldsymbol{u}\in\mathbb{R}^n$ is

$$\langle\boldsymbol{v},\boldsymbol{u}\rangle_e=\boldsymbol{v}'\boldsymbol{u}.$$

Definition 46 (Range of a matrix). The *range* of a matrix $\boldsymbol{A}\in\mathbb{R}^{n\times p}$ is defined as

$$\text{Range}(\boldsymbol{A}):=\{\boldsymbol{u}\in\mathbb{R}^n\ :\ \boldsymbol{u}=\boldsymbol{A}\boldsymbol{v}\text{ for some }\boldsymbol{v}\in\mathbb{R}^p\}.$$

The range of a matrix is also called the *column space*, i.e., the space spanned by the matrix's columns, since:

Proposition 2.1.4. *Let* $\boldsymbol{A}=[\boldsymbol{A}_1,\ldots,\boldsymbol{A}_p]\in\mathbb{R}^{n\times p}$. *Then,* $\text{Range}(\boldsymbol{A})=\text{span}(\boldsymbol{A}_1,\ldots,\boldsymbol{A}_p)$.

Definition 47 (Kernel of a matrix). The *kernel*, or *null space*, of a matrix $\boldsymbol{A}\in\mathbb{R}^{n\times p}$ is defined as

$$\text{Ker}(\boldsymbol{A}):=\{\boldsymbol{v}\in\mathbb{R}^p\ :\ \boldsymbol{A}\boldsymbol{v}=\boldsymbol{0}\}.$$

Proposition 2.1.5. *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$*. Then,* $\text{Range}(\boldsymbol{A})$ *and* $\text{Ker}(\boldsymbol{A}')$ *are orthogonal subspaces of* $\mathbb{R}^n$ *such that* $\mathbb{R}^n = \text{Range}(\boldsymbol{A}) + \text{Ker}(\boldsymbol{A}')$*.*

Definition 48 (Rank of a matrix)*.* The *rank* of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, denoted $\text{Rank}(\boldsymbol{A})$, is the maximum number of linearly independent columns of $\boldsymbol{A}$.

Proposition 2.1.6. *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$*. Then,* $\text{Rank}(\boldsymbol{A}) \leq \min\{n, p\}$*.*

Definition 49 (Eigenvalue)*.* $\lambda \in \mathbb{R}$ is an *eigenvalue* of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ if there exists $\boldsymbol{v} \in \mathbb{R}^p$ such that $\boldsymbol{v} \neq 0$ and

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}.$$

Definition 50 (Eigenvector)*.* Given matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ with eigenvalue $\lambda \in \mathbb{R}$, $\boldsymbol{v} \in \mathbb{R}^p$ is an *eigenvector* of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ corresponding to λ if $\boldsymbol{v} \neq 0$ and

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}.$$

Proposition 2.1.7. *Every matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ *has an eigenvalue.*

Proposition 2.1.8. *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$*. Then,* $\boldsymbol{A}$ *has at most* $\text{Rank}(\boldsymbol{A})$ *distinct eigenvalues.*

Proposition 2.1.9. *Suppose* $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ *are distinct eigenvalues of* $A \in \mathbb{R}^{n\times p}$ *and* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r \in \mathbb{R}^p$ *are corresponding eigenvectors. Then,* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ *are linearly independent.*

Definition 51 (Singular values)*.* The *singular values* of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ are the nonnegative square roots of the eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$.

Definition 52 (Symmetric matrix)*.* A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *symmetric* if $\boldsymbol{A}' = \boldsymbol{A}$.

Definition 53 (Positive definite matrix)*.* A square symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive definite* if $\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} > 0$ for all $\boldsymbol{v} \in \mathbb{R}^n$ such that $\boldsymbol{v} \neq 0$.

Definition 54 (Positive semi-definite matrix)*.* A square symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive semi-definite* if $\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^n$.

Proposition 2.1.10. *A square symmetric matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is positive definite (positive semi-definite) if and only if all of its eigenvalues are positive (nonnegative).*

Definition 55 (Identity matrix)*.* The *identity* matrix on $\mathbb{R}^n$ is defined as

$$\boldsymbol{I} := \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Definition 56 (Diagonal of a matrix). The *diagonal* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ indicates the elements "on the diagonal": $A_{1,1}, \ldots, A_{n,n}$.

Definition 57 (Diagonal matrix). A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a *diagonal* matrix if all its elements outside of the diagonal are zero. We can write $\boldsymbol{A} = \text{diag}(A_{1,1}, \ldots, A_{n,n})$.

Definition 58 (Invertible matrix, matrix inverse). A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *invertible* if there is a matrix $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$. We call $\boldsymbol{B}$ the *inverse* of $\boldsymbol{A}$ and denote it by $\boldsymbol{A}^{-1}$.

Proposition 2.1.11. *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is invertible if and only if* $\text{Rank}(\boldsymbol{A}) = n$*, or equivalently, if and only if* $\text{Ker}(\boldsymbol{A}) = \emptyset$.

Proposition 2.1.12. *A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is invertible if and only if it is positive definite.*

Definition 59 (Orthogonal matrix). A square matrix $\boldsymbol{P} \in \mathbb{R}^{p\times p}$ is *orthogonal*, or orthonormal, if $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}$.

Definition 60 (Projection matrix). A square matrix $\boldsymbol{P} \in \mathbb{R}^{p\times p}$ is a *projection* matrix if $\boldsymbol{P} = \boldsymbol{P}^2$.

Definition 61 (Orthogonal projection matrix). A square matrix $\boldsymbol{P} \in \mathbb{R}^{p\times p}$ is an *orthogonal projection* matrix if it is a projection matrix and $\boldsymbol{P} = \boldsymbol{P}'$.

Projections and orthogonal projections have the following properties.

Proposition 2.1.13. *For any projection matrix $\boldsymbol{P} \in \mathbb{R}^{p\times p}$ and vector $\boldsymbol{b} \in \mathbb{R}^p$, we have*

$$\boldsymbol{b} = \boldsymbol{P}\boldsymbol{b} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b}.$$

If $\boldsymbol{P}$ is an orthogonal projection matrix, then

$$(\boldsymbol{P}\boldsymbol{b})'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b} = 0.$$

Definition 62 (Trace). The *trace* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, denoted $\text{Trace}(\boldsymbol{A})$, is the sum of its diagonal elements:

$$\text{Trace}(\boldsymbol{A}) = A_{11} + \ldots, A_{n,n}.$$

Proposition 2.1.14. *The* Trace *is a linear function.*

Proposition 2.1.15 (Properties of the trace)**.**

1. $\text{Trace}(\boldsymbol{A}) = \lambda_1 + \ldots + \lambda_n$ *for all $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with $\lambda_1, \ldots, \lambda_n$ denoting the (not necessarily distinct) eigenvalues of $\boldsymbol{A}$.*
2. $\text{Trace}(\boldsymbol{A}) = \text{Trace}(\boldsymbol{A}')$ *for all $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.*
3. $\text{Trace}(\boldsymbol{AB}) = \text{Trace}(\boldsymbol{BA})$ *for all for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$.*
4. $\text{Trace}(\boldsymbol{A}'\boldsymbol{B}) = \text{Trace}(\boldsymbol{A}\boldsymbol{B}') = \text{Trace}(\boldsymbol{B}'\boldsymbol{A}) = \text{Trace}(\boldsymbol{B}\boldsymbol{A}')$ *for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times p}$.*

2.1.1 Moore-Penrose inverse

The Moore-Penrose inverse, or matrix pseudoinverse, is a generalization of the inverse of a matrix that was independently introduced by Moore [1920] and Penrose [1955].

Definition 63 (Moore-Penrose inverse). The matrix $\boldsymbol{A}^+ \in \mathbb{R}^{p\times n}$ is a *Moore-Penrose inverse* of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ if

(i) $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$;

(ii) $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$;

(iii) $(\boldsymbol{A}\boldsymbol{A}^+)' = \boldsymbol{A}\boldsymbol{A}^+$;

(iv) $(\boldsymbol{A}^+\boldsymbol{A})' = \boldsymbol{A}^+\boldsymbol{A}$.

Properties and examples of the Moore-Penrose inverse

We now list the main properties of the Moore-Penrose inverse.

Proposition 2.1.16. *For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, the Moore-Penrose inverse $\boldsymbol{A}^+$ exists and is unique.*

Proposition 2.1.17. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ have $\mathrm{Rank}(\boldsymbol{A}) = p$. Then, $\boldsymbol{A}^+ = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$.*

Proposition 2.1.18. *Let the square matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ have $\mathrm{Rank}(\boldsymbol{A}) = p$. Then, $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$.*

Proposition 2.1.19. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$. Then,*

$$\boldsymbol{A}^+ = \lim_{\lambda\to 0}(\boldsymbol{A}'\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}' = \lim_{\lambda\to 0}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}' + \lambda\boldsymbol{I})^{-1}.$$

Proof. See Albert [1972]. □

Proposition 2.1.20. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times p}$. Then:*

1. $\boldsymbol{A} = (\boldsymbol{A}^+)^+$.
2. $\boldsymbol{A}^+ = (\boldsymbol{A}'\boldsymbol{A})^+\boldsymbol{A}' = \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^+$.
3. $(\boldsymbol{A}')^+ = (\boldsymbol{A}^+)'$.
4. $(\boldsymbol{A}'\boldsymbol{A})^+ = \boldsymbol{A}^+(\boldsymbol{A}')^+$.
5. $(\boldsymbol{A}\boldsymbol{A}')^+ = (\boldsymbol{A}')^+\boldsymbol{A}^+$.

6. $\text{Range}(\boldsymbol{A}^+) = \text{Range}(\boldsymbol{A}') = \text{Range}(\boldsymbol{A}^+\boldsymbol{A}) = \text{Range}(\boldsymbol{A}'\boldsymbol{A})$.

7. $\text{Ker}(\boldsymbol{A}^+) = \text{Ker}(\boldsymbol{A}\boldsymbol{A}^+) = \text{Ker}((\boldsymbol{A}\boldsymbol{A}')^+) = \text{Ker}(\boldsymbol{A}\boldsymbol{A}') = \text{Ker}(\boldsymbol{A}')$.

Proposition 2.1.21. *For any matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$*:*

1. $\boldsymbol{A}\boldsymbol{A}^+$ *is an orthogonal projection onto* $\text{Range}(\boldsymbol{A})$.

2. $\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^+$ *is an orthogonal projection onto* $\text{Ker}(\boldsymbol{A}')$.

3. $\boldsymbol{A}^+\boldsymbol{A}$ *is an orthogonal projection onto* $\text{Range}(\boldsymbol{A}')$.

4. $\boldsymbol{I} - \boldsymbol{A}^+\boldsymbol{A}$ *is an orthogonal projection onto* $\text{Ker}(\boldsymbol{A})$.

We also collect some examples of the Moore-Penrose inverse.

Example 2. If $a \in \mathbb{R}$, then $a^+ = \begin{cases} a^{-1} & a \neq 0 \\ 0 & a = 0 \end{cases}$.

Example 3. If $\boldsymbol{A} = \text{diag}(A_1, \ldots, A_{p-k}, 0, \ldots, 0) \in \mathbb{R}^{p\times p}$, then

$$\boldsymbol{A}^+ = \text{diag}(1/A_1, \ldots, 1/A_{p-k}, 0, \ldots, 0).$$

Example 4. If $\boldsymbol{A} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, then $\boldsymbol{A}^+ = [1/5.2/5]$.

Example 5. If $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, then $\boldsymbol{A}^+ = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

Example 6. If $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, then $\boldsymbol{A}^+ = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}$.

Systems of linear equations and least squares

The Moore-Penrose inverse plays a central role in the study of solutions to systems of linear equations and linear least squares problems.

Theorem 2.1.5 (Solutions of systems of linear equations)**.** *For* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ *and* $\boldsymbol{b} \in \mathbb{R}^n$*, let* $L := \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{b}\}$*. The following statements hold:*

(i) If $\boldsymbol{b} \notin \text{Range}(\boldsymbol{A})$*, then* L *is empty.*

(ii) If $\boldsymbol{b} \in \text{Range}(\boldsymbol{A})$*, then* $L = \boldsymbol{A}^+\boldsymbol{b} + \text{Ker}(\boldsymbol{A})$.

Corollary 2.1.6. *Given a square matrix* $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{b} \in \mathbb{R}^p$*, let* $L := \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \boldsymbol{A\theta} = \boldsymbol{b}\}$*. Then,* $\boldsymbol{A}^+\boldsymbol{b}$ *is the unique element of* L *if and only if* $\text{Rank}(\boldsymbol{A}) = p$*. In this case,* $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$*.*

Corollary 2.1.7. *For* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *and* $\boldsymbol{y} \in \mathbb{R}^n$*:*

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}} \|\boldsymbol{y} - \boldsymbol{X\theta}\|_2^2 = \boldsymbol{X}^+\boldsymbol{y} + \text{Ker}(\boldsymbol{X}).$$

2.1.2 Eigenvalue and Singular value decomposition

This section introduces the eigenvalue and the singular value decompositions, which are matrix factorizations with many applications to statistics and machine learning.

Definition 64 (Singular value decomposition). The *Singular Value Decomposition* (SVD) of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ with rank $r := \text{Rank}(\boldsymbol{A})$ is given by

$$\boldsymbol{A} = \boldsymbol{USV}' = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i',$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrix, and

$$\boldsymbol{S} = \begin{bmatrix} \text{diag}(s_1, \ldots, s_r) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{n\times p},$$

where $s_1, \ldots, s_r$ are the positive *singular values* of $\boldsymbol{A}$.

Proposition 2.1.22 (Existence)**.** *Any matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ *admits a singular value decomposition.*

The next proposition demonstrates the relation of the SVD to the four fundamental subspaces of a matrix.

Proposition 2.1.23. *Consider Definition 64. Then,*

(i) $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$ *is an orthonormal basis of* $\text{Range}(\boldsymbol{A})$*.*

(ii) $\{\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_n\}$ *is an orthonormal basis of* $\text{Ker}(\boldsymbol{A}')$*.*

(iii) $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$ *is an orthonormal basis of* $\text{Range}(\boldsymbol{A}')$*.*

(iv) $\{\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_p\}$ *is an orthonormal basis of* $\text{Ker}(\boldsymbol{A})$*.*

We thus have $\text{Range}(\boldsymbol{A}) = \sum_{j=1}^{r} \boldsymbol{u}_j\boldsymbol{u}_j'$ and $\text{Range}(\boldsymbol{A}') = \sum_{j=1}^{r} \boldsymbol{v}_j\boldsymbol{v}_j'$. Moreover, if $r < p$, we have $\text{Ker}(\boldsymbol{A}') = \sum_{j=r+1}^{p} \boldsymbol{u}_j\boldsymbol{u}_j'$, and if $r < n$, we have $\text{Ker}(\boldsymbol{A}) = \sum_{j=r+1}^{n} \boldsymbol{v}_j\boldsymbol{v}_j'$.

Proposition 2.1.24. *The Moore-Penrose inverse of a matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ *admitting SVD decomposition* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$ *is given by* $\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{S}^+\boldsymbol{U}'$.

Definition 65 (Eigenvalue decomposition). The *eigenvalue decomposition* of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ with n linearly independent eigenvectors $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$ is given by

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1},$$

where $\boldsymbol{Q} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n]$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Proposition 2.1.25 (Relation between the singular value and the eigenvalue decompositions)**.** *Given a matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ *with SVD* $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}'$*:*

1. $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{V}\boldsymbol{S}'\boldsymbol{S}\boldsymbol{V}'$*;*
2. $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{U}\boldsymbol{S}\boldsymbol{S}'\boldsymbol{U}'$.

2.2 Convex analysis

This section introduces a selection of definitions and results from convex analysis that are used in these lecture notes. A book-length exposition of convex analysis can be found in Bauschke et al. [2017], among others.

Basic definitions

Definition 66 (Closed set). A set $C \subset \mathbb{R}^p$ is *closed* if it contains all of its limit points.

Definition 67 (Bounded set). A set $C \subset \mathbb{R}^p$ is *bounded* if there exists $r > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\beta} \in \mathbb{R}^p$, we have $\|\boldsymbol{\theta} - \boldsymbol{\beta}\| < r$.

Definition 68 (Convex set). A set $C \subset \mathbb{R}^p$ is *convex* if for all $0 < \alpha < 1$ and all $\boldsymbol{\theta}, \boldsymbol{\beta} \in C$:

$$\alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\beta} \in C.$$

In particular, $\mathbb{R}^p$ and $\emptyset$ are convex.

Definition 69 (Epigraph of a function). The *epigraph* of function $f : \mathbb{R}^p \to (-\infty, +\infty]$ is

$$\operatorname{epi}(f) := \{(\boldsymbol{\theta}, \xi) \in \mathbb{R}^p \times \mathbb{R} \ : \ f(\boldsymbol{\theta}) \leq \xi\}.$$

Definition 70 (Domain of a function). The *domain* of function $f : \mathbb{R}^p \to (-\infty, +\infty]$ is

$$\operatorname{dom}(f) := \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ f(\boldsymbol{\theta}) < +\infty\}.$$

Definition 71 (Lower level set of a function). The *lower level set* of function $f : \mathbb{R}^p \to (-\infty, +\infty]$ at height $\xi \in \mathbb{R}$ is

$$\text{lev}_{\leq \xi}(f) := \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ f(\boldsymbol{\theta}) \leq \xi\}.$$

Definition 72 (Proper function). function $f : \mathbb{R}^p \to (-\infty, +\infty]$ is *proper* if $\text{dom}(f) \neq \emptyset$.

Definition 73 (Convex function). Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be a proper function. Then f is *convex* if its epigraph $\text{epi}(f)$ is convex. Equivalently, f is convex if for all $0 < \alpha < 1$ and all $\boldsymbol{\theta}, \boldsymbol{\beta} \in \mathbb{R}^p$ such that $\boldsymbol{\theta} \neq \boldsymbol{\beta}$:

$$f(\alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\beta}) \leq \alpha f(\boldsymbol{\theta}) + (1-\alpha) f(\boldsymbol{\beta}).$$

Definition 74 (Strictly convex function). Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be a proper function. Then f is *strictly convex* if for all $0 < \alpha < 1$ and all $\boldsymbol{\theta}, \boldsymbol{\beta} \in \mathbb{R}^p$ such that $\boldsymbol{\theta} \neq \boldsymbol{\beta}$:

$$f(\alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\beta}) < \alpha f(\boldsymbol{\theta}) + (1-\alpha) f(\boldsymbol{\beta}).$$

Definition 75 (Limit inferior). The *limit inferior* of $f : \mathbb{R}^p \to (-\infty, +\infty]$ at a point $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is

$$\liminf_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} f(\boldsymbol{\theta}) = \lim_{\varepsilon \to 0} \left(\inf\{f(\boldsymbol{\theta}) \ : \ \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \varepsilon\}\right).$$

Definition 76 (Lower semicontinuous function). Function $f : \mathbb{R}^p \to (-\infty, +\infty]$ is *lower semicontinuous* at $\boldsymbol{\theta}^* \in \mathbb{R}^p$ if

$$\liminf_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} f(\boldsymbol{\theta}) \geq f(\boldsymbol{\theta}^*).$$

Definition 77 (Coercive function). Function $f : \mathbb{R}^p \to (-\infty, +\infty]$ is *coercive* if

$$\lim_{\|\boldsymbol{\theta}\| \to +\infty} f(\boldsymbol{\theta}) = +\infty.$$

Definition 78 (Subdifferential). Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be a proper function. The *subdifferential* of f is the set-valued operator:[1]

$$\partial f : \mathbb{R}^p \to 2^{\mathbb{R}^p}; \boldsymbol{\theta} \mapsto \{\boldsymbol{\beta} \in \mathbb{R}^p \ : \ \langle \boldsymbol{v} - \boldsymbol{\theta}, \boldsymbol{\beta} \rangle + f(\boldsymbol{\theta}) \leq f(\boldsymbol{v}) \ \forall \boldsymbol{v} \in \mathbb{R}^p\}.$$

Let $\boldsymbol{\theta} \in \mathbb{R}^p$. Then f is *subdifferentiable* at $\boldsymbol{\theta}$ if $\partial f(\boldsymbol{\theta}) \neq \emptyset$; the elements of $\partial f(\boldsymbol{\theta})$ are the *subgradients* of f at $\boldsymbol{\theta}$.

[1] Given a set C, the set of all subsets of C, including the empty set and C itself, is denoted 2^C. This set is called the power set of C.

Graphically, a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is a subgradient of a proper function $f : \mathbb{R}^p \to (-\infty, +\infty]$ at $\boldsymbol{\theta} \in \mathrm{dom}(f)$ if

$$f_{\boldsymbol{\beta},\boldsymbol{\theta}} : \boldsymbol{v} \mapsto \langle \boldsymbol{v} - \boldsymbol{\theta}, \boldsymbol{\beta} \rangle + f(\boldsymbol{\theta}),$$

which coincides with f at $\boldsymbol{\theta}$, lies below f.

Example 7. The subdifferential of the absolute value function $|\cdot|$ at $\theta \in \mathbb{R}$ is given by

$$\partial|\theta| = \begin{cases} \{1\}, & \theta > 0 \\ [-1, 1], & \theta = 0 \\ \{-1\}, & \theta < 0 \end{cases}.$$

See Bauschke et al. [2017, Example 16.15].

Minimizers of convex optimization problems

Definition 79 (Global minimizer). $\boldsymbol{\theta}^*$ is a *(global) minimizer* of a proper function $f : \mathbb{R}^p \to (-\infty, +\infty]$ over $C \subset \mathbb{R}^p$ if $f(\boldsymbol{\theta}^*) = \inf_{\boldsymbol{\theta} \in C} f(\boldsymbol{\theta})$. The set of minimizers of f over C is denoted by

$$\operatorname*{argmin}_{\boldsymbol{\theta} \in C} f(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in R^p} \{ f(\boldsymbol{\theta}) \ : \ \boldsymbol{\theta} \in C \}.$$

Definition 80 (Local minimizer). $\boldsymbol{\theta}^*$ is a *local minimizer* of a proper function $f : \mathbb{R}^p \to (-\infty, +\infty]$ if there exists $c > 0$ such that

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta} \in R^p} f(\boldsymbol{\theta}) \ : \ \|\boldsymbol{\theta}\| \leq c \}.$$

Proposition 2.2.1 (Convex problems: local minimizers are global minimizers)**.** *Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be proper and convex. Then every local minimizer of f is a minimizer.*

Proposition 2.2.2 (Convex problems: argmin is convex)**.** *Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be proper and convex and $C \subset \mathbb{R}^p$. Then $\operatorname{argmin}_{\boldsymbol{\theta} \in C} f(\boldsymbol{\theta})$ is convex.*

Proposition 2.2.3 (Existence of minimizers)**.** *Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be proper, convex and lower semicontinuous and C be a closed convex subset of $\mathbb{R}^p$ such that $C \cap \mathrm{dom}(f) \neq \emptyset$. Suppose that one of the following holds:*

(i) f is coercive.

(ii) C is bounded.

Then f has a minimizer over C.

Proof. Since $C \cap \mathrm{dom}(f) \neq \emptyset$, there exists $\boldsymbol{\theta} \in \mathrm{dom}(f)$ such that $D = C \cap \mathrm{lev}_{\leq f(\boldsymbol{\theta})}(f)$ is not empty, closed and convex. Moreover, D is bounded since C or $\mathrm{lev}_{\leq f(\boldsymbol{\theta})}(f)$ is. The result therefore follows from Bauschke et al. [2017, Thm. 11.10]. □

Proposition 2.2.4 (Uniqueness of minimizers)**.** *Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be proper and strictly convex. Then f has at most one minimizer.*

Proof. Set $\mu := \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$ and suppose that there exist two distinct points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathrm{dom}(f)$ such that $f(\boldsymbol{\theta}_1) = f(\boldsymbol{\theta}_2) = \mu$. Since $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathrm{lev}_{\leq \mu}(f)$, which is convex, so does $\boldsymbol{\beta} = (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)/2$. Therefore $f(\boldsymbol{\beta}) = \mu$. It follows from the strict convexity of f that

$$\mu = f(\boldsymbol{\beta}) < \max\{f(\boldsymbol{\theta}_1), f(\boldsymbol{\theta}_2)\} = \mu,$$

which is impossible. □

Global minimizers of proper functions can be characterized by a simple rule which extends a seventeenth century result due to Pierre Fermat.

Theorem 2.2.1 (Fermat's rule)**.** *Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be proper. Then*

$$\operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}) = \{\boldsymbol{\theta}^* \in \mathbb{R}^p \ : \ \mathbf{0} \in \partial f(\boldsymbol{\theta}^*)\}.$$

Proof. Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$. Then $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta})$ if and only if, for every $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\langle \boldsymbol{\beta} - \boldsymbol{\theta}^*, \mathbf{0} \rangle + f(\boldsymbol{\theta}^*) \leq f(\boldsymbol{\beta}).$$

By definition of subgradient, this last requirement reads $\mathbf{0} \in \partial f(\boldsymbol{\theta}^*)$. □

Theorem 2.2.2 (Hilbert projection theorem)**.** *For every vector $\boldsymbol{\theta} \in \mathbb{R}^p$ and every nonempty closed convex $C \subset \mathbb{R}^p$, there exists a unique vector $\boldsymbol{\beta} \in \mathbb{R}^p$ for which*

$$\|\boldsymbol{\theta} - \boldsymbol{\beta}\|_2^2 = \inf_{\boldsymbol{\eta} \in C} \|\boldsymbol{\theta} - \boldsymbol{\eta}\|_2^2 .$$

If C is a vector subspace of $\mathbb{R}^p$, then the minimizer $\boldsymbol{\beta}$ is the unique element in C such that $\boldsymbol{\theta} - \boldsymbol{\beta}$ is orthogonal to C.

2.3 Probability theory

This section introduces a selection of definitions and results from probability theory that are used in these lecture notes. A book-length exposition of probability theory can be found in Billingsley [2017] and Vershynin [2018], among others.

All random variables are (real valued and) defined on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 81 (Cumulative Distribution Function). The *Cumulative Distribution Function* (CDF) of random variable X is the function

$$F_X : \mathbb{R} \to [0,1]; x \mapsto \mathbb{P}[X \leq x].$$

Definition 82 (Expected value). The *expected value* of a random variable X is

$$\mathbb{E}[X] := \int_\Omega X d\mathbb{P}.$$

Definition 83 (Moment generating function). the *Moment Generating Function* (MGF) of a random variable X is the function

$$M_X : \mathbb{R} \to [0, +\infty]; t \mapsto \mathbb{E}[\exp(tX)].$$

Definition 84 (Moment of order p). For $p \in \mathbb{R}$, the *moment of order* p of a random variable X is $\mathbb{E}[|X|^p]$.

Definition 85 (L^p-norm). The *L^p-norm* of a random variable X is, for $p > 0$:

$$\|X\|_{L^p} := \mathbb{E}[|X|^p]^{1/p},$$

and for $p = \infty$:

$$\|X\|_{L^p} := \operatorname{ess\,sup} |X| := \sup\{b \in \mathbb{R} \ : \ \mathbb{P}(\{\omega : X(\omega) < b\}) = 0\}.$$

Definition 86 (L^p-space). The space $L^p = L^p(\Omega, \mathcal{F}, \mathbb{P})$ consists of all random variables X with finite L^p norm:

$$L^p := \{X \ : \ \|X\|_{L^p} < +\infty\}.$$

Definition 87 (Conjugate exponents). $p, q \in [1, \infty]$ are *conjugate exponents* if $1/p + 1/q = 1$.

The inner product between L^p and L^q where $p, q \in [1, \infty]$ are conjugate exponents is for all $X \in L^p$ and $Y \in L^q$:

$$\langle X, Y \rangle_{L^p - L^q} := \mathbb{E}[XY].$$

The inner product in L^2 is for all $X, Y' in L^2$:

$$\langle X, Y \rangle_{L^2} := \mathbb{E}[XY].$$

The Variance of $X \in L^2$ is

$$\mathbb{V}\mathrm{ar}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \|X - \mathbb{E}[X]\|_{L^2}^2 ,$$

and the standard deviation is

$$\sigma(X) := \sqrt{\mathbb{V}\mathrm{ar}[X]} = \|X - \mathbb{E}[X]\|_{L^p} .$$

The covariance between $X, Y \in L^2$ is

$$\mathbb{C}\mathrm{ov}[X, Y] := \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle_{L^2} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Classical inequalities

Theorem 2.3.1 (Jensen's inequality)**.** *For any random variable X and a convex function $f : \mathbb{R} \to \mathbb{R}$, we have*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(x)].$$

The following proposition is a consequence of Jensen's inequality.

Proposition 2.3.1. *For any random variable X and any $p, q \in [0, \infty]$ with $p \leq q$:*

$$\|X\|_{L^p} \leq \|Y\|_{L^q} .$$

Therefore, $L^p \subset L^q$ for any $p, q \in [0, \infty]$ with $p \leq q$.

Theorem 2.3.2 (Minkowski's inequality)**.** *For any $p \in [1, \infty]$ and any random variables $X, Y \in L^p$:*

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p} .$$

Theorem 2.3.3 (Cauchy-Schwarz inequality)**.** *For any random variables $X, Y \in L^2$:*

$$|\langle X, Y \rangle| = |\mathbb{E}[XY]| \leq \|X\|_{L^2} \|Y\|_{L^2} .$$

Theorem 2.3.4 (Hölder's inequality)**.** *For any random variables $X \in L^p$ and $Y \in L^q$ with conjugate exponents $p, q \in (1, \infty)$:*

$$|\mathbb{E}[XY]| \leq \|X\|_{L^p} \|Y\|_{L^q} .$$

This inequality also holds for $p = 1$ and $q = \infty$.

The tails and the moments of a random variable are connected.

Proposition 2.3.2 (Integral identity)**.** *For any nonnegative random variable X:*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]dx.$$

The two sides of this identity are either both finite or both infinite.

Theorem 2.3.5 (Markov's inequality)**.** *For any nonnegative random variable X and $x > 0$:*

$$\mathbb{P}[X \geq x] \leq \mathbb{E}[X]/x.$$

A consequence of Markov's inequality is Chebyshev's inequality, which bounds the concentration of a random variable about its mean.

Theorem 2.3.6 (Chebyshev's inequality)**.** *Let X be a random variable with finite mean μ and finite variance σ^2. Then, for any $x > 0$:*

$$\mathbb{P}[|X - \mu| \geq x] \leq \sigma^2/x^2.$$

Proposition 2.3.3 (Generalization of Markov's inequality)**.** *For any random variable X with mean $\mu \in \mathbb{R}$ and finite moment of order $p \geq 1$, and for any $x > 0$:*

$$\mathbb{P}[|X - \mu| \geq x] \leq \mathbb{E}[|X - \mu|^p]/x^p.$$

Concentration of sums of independent random variables

Concentration inequalities quantify how a random variable deviates around its mean.

Definition 88 (Symmetric Bernoulli distribution)*.* A random variable X has a *symmetric Bernoulli* distribution if

$$\mathbb{P}[X = -1] = \mathbb{P}[X = +1] = 1/2.$$

Theorem 2.3.7 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be an independent symmetric Bernoulli random variables, and $\boldsymbol{a} \in \mathbb{R}^n$. Then, for any $x \geq 0$:*

$$\mathbb{P}\left[\sum_{i=1}^n a_i X_i \geq x\right] \leq \exp\left(-\frac{x^2}{2\left\|\boldsymbol{a}\right\|_2^2}\right).$$

Theorem 2.3.8 (Two-sided Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be an independent symmetric Bernoulli random variables, and $\boldsymbol{a} \in \mathbb{R}^n$. Then, for any $x > 0$:*

$$\mathbb{P}\left[\left|\sum_{i=1}^n a_i X_i\right| \geq x\right] \leq 2\exp\left(-\frac{x^2}{2\left\|\boldsymbol{a}\right\|_2^2}\right).$$

Theorem 2.3.9 (Hoeffding's inequality for bounded random variables)**.** *Let $X_1, \ldots, X_n$ be an independent random variables. Assume that $X_i \in [l_i, u_i]$ with $l_i, u_i \in \mathbb{R}$ and $l_i \leq u_i$. Then, for any $x > 0$:*

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq x\right] \leq \exp\left(-\frac{2x^2}{\sum_{i=1}^{n}(u_i - l_i)^2}\right).$$

Theorem 2.3.10 (Chernoff's inequality)**.** *Chernoff's inequality Let X_i be independent Bernoulli random variables with parameter $p_i \in [0, 1]$. Let $S_n := \sum_{i=1}^{n} X_i$ and its mean $\mu := \mathbb{E}[S_n]$. Then, for any $x > 0$:*

$$\mathbb{P}[S_n \geq x] \leq \exp(-\mu)\left(\frac{e\mu}{x}\right)^x.$$

Proposition 2.3.4 (Tails of the standard normal distribution)**.** *Let $Z \sim N(0, 1)$. Then, for all $z > 0$:*

$$\left(\frac{1}{z} - \frac{1}{z^3}\right)\frac{1}{\sqrt{2\pi}}e^{-z^2/2} \leq \mathbb{P}[Z \geq z] \leq \frac{1}{z}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}.$$

In particular, for $z \geq 1$:

$$\mathbb{P}[Z \geq z] \leq \frac{1}{\sqrt{2\pi}}e^{-z^2/2}.$$

Proposition 2.3.5 (Tails of the normal distribution)**.** *Let $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$. Then, for all $x \geq 0$:*

$$\mathbb{P}[X - \mu \geq x] \leq \exp\left(\frac{-x^2}{2\sigma^2}\right).$$

Proposition 2.3.6. *Let $Z \sim N(0, 1)$. Then, for all $z \geq 0$:*

$$\mathbb{P}[|Z| \geq z] \leq 2\exp(-z^2/2).$$

Proposition 2.3.7 (Sub-Gaussian properties)**.** *Let X be a random variable. Then, there are constants $C_1, \ldots, C_5 > 0$ for which the following properties are equivalent.*

(i) The tails of X satisfy for all $x \geq 0$:

$$\mathbb{P}[|X| \geq x] \leq 2\exp(-x^2/C_1^2).$$

(ii) The moments of X satisfy for all $p \geq 1$:

$$\|X\|_{L^p} = \mathbb{E}[|X|^p]^{1/p} \leq C_2\sqrt{p}.$$

(iii) The MGF of X^2 satisfies for all $t \in \mathbb{R}$ such that $|t| \leq 1/C_3$:

$$\mathbb{E}[\exp(t^2 X^2)] \leq \exp(C_3^2 t^2).$$

(iv) The MGF of X^2 is bounded at some point, namely

$$\mathbb{E}[\exp(X^2/C_4^2)] \leq 2.$$

If further $\mathbb{E}[X] = 0$, these properties are equivalent to:

(v) The MGF of X satisfies for all $t \in \mathbb{R}$:

$$\mathbb{E}[\exp(tX)] \leq \exp(C_5^2 t^2).$$

Definition 89 (Sub-Gaussian random variables). A random variable X that satisfies the equivalent conditions of Proposition 2.3.7 is a *sub-Gaussian* random variable, denoted $X \sim$ sub-G.

Gaussian, symmetric Bernoulli, uniform, and bounded random variables are examples of sub-Gaussian random variables. The tails of the distribution of a sub-Gaussian random variable decay at least as fast as the tails of a Gaussian distribution. The Poisson, exponential, Pareto, and Cauchy distribution are examples of distributions that are not sub-Gaussian.

Definition 90 (Variance proxy). For a random variable $X \sim$ sub-G, if there is some $s > 0$ such that for all $t \in \mathbb{R}$:

$$\mathbb{E}[e^{(X-\mathbb{E}[X])t}] \leq \exp(s^2 t^2/2),$$

then s^2 is called *variance proxy*.

Proposition 2.3.8 (Weighted sum of independent sub-Gaussian random variables)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables, all with variance proxy σ^2 where $\sigma > 0$. Then, for any $\boldsymbol{a} \in \mathbb{R}^n$:*

$$\mathbb{P}\left[\sum_{i=1}^n a_i X_i \geq t\right] \leq \exp\left(-\frac{t^2}{2\sigma^2 \|\boldsymbol{a}\|_2^2}\right),$$

and

$$\mathbb{P}\left[\sum_{i=1}^n a_i X_i \leq -t\right] \leq \exp\left(-\frac{t^2}{2\sigma^2 \|\boldsymbol{a}\|_2^2}\right).$$

Definition 91 (Sub-Gaussian norm). The *sub-Gaussian norm* $\|X\|_{\psi_2}$ of random variable X is defined as

$$\|X\|_{\psi_2} := \inf\{t > 0 \ : \ \mathbb{E}[\exp(X^2/t^2)] \leq 2\}.$$

Proposition 2.3.9. *If X is a sub-Gaussian random variable, then $X - \mathbb{E}[X]$ is sub-Gaussian and for a constant $C > 0$:*

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq C \|X\|_{\psi_2}.$$

Proposition 2.3.10 (Sums of independent sub-Gaussian)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with mean zero. Then $\sum_{i=1}^n X_i$ is also a sub-Gaussian random variable, and, for a constant $C > 0$:*

$$\left\| \sum_{i=1}^n X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^n \|X_i\|_{\psi^2}^2.$$

We can now extend the Hoeffding's inequality to sub-Gaussian distributions.

Proposition 2.3.11 (General Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with mean zero and $C > 0$ a constant. Then, for every $t \geq 0$:*

$$\mathbb{P}\left[\left| \sum_{i=1}^n X_i \right| \geq t \right] \leq 2 \exp\left(- \frac{Ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2} \right).$$

Proposition 2.3.12. *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with mean zero, $\boldsymbol{a} \in \mathbb{R}^n$, $K = \max_{i=1}^n \|X_i\|_{\psi_2}$ and $C > 0$ a constant. Then, for every $t \geq 0$:*

$$\mathbb{P}\left[\left| \sum_{i=1}^n a_i X_i \right| \geq t \right] \leq 2 \exp\left(- \frac{Ct^2}{K^2 \|\boldsymbol{a}\|_2^2} \right).$$

Proposition 2.3.13 (Khintchine's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables, all with mean zero and unit variance proxy, $\boldsymbol{a} \in \mathbb{R}^n$, $K = \max_{i=1}^n \|X_i\|_{\psi_2}$ and $C > 0$ a constant. Then, for every $p \in [2, \infty)$:*

$$\left(\sum_{i=1}^n a_i^2 \right)^{1/2} \leq \left\| \sum_{i=1}^n a_i X_i \right\|_{L^p} \leq CK\sqrt{p} \left(\sum_{i=1}^n a_i^2 \right)^{1/2}.$$

The sub-Gaussian distribution does not embed distributions whose tails are heavier than Gaussian.

Proposition 2.3.14 (Sub-exponential properties)**.** *Let X be a random variable. Then, there are constants $K_1, \ldots, K_5 > 0$ for which the following properties are equivalent.*

(i) The tails of X satisfy for all $x \geq 0$:

$$\mathbb{P}[|X| \geq x] \leq 2\exp(-x/K_1).$$

(ii) The moments of X satisfy for all $p \geq 1$:

$$\|X\|_{L^p} = \mathbb{E}[|X|^p]^{1/p} \leq K_2 p.$$

(iii) The MGF of $|X|$ satisfies for all $t \in \mathbb{R}$ such that $0 \leq t \leq 1/K_3$:

$$\mathbb{E}[\exp(t|X|)] \leq \exp(L_3 t).$$

(iv) The MGF of $|X|$ is bounded at some point, namely

$$\mathbb{E}[\exp(|X|/K_4)] \leq 2.$$

If further $\mathbb{E}[X] = 0$, these properties are equivalent to:

(v) The MGF of X satisfies for all $t \in \mathbb{R}$ such that $|t| \leq 1/K_5$:

$$\mathbb{E}[\exp(tX)] \leq \exp(K_5^2 t^2).$$

Definition 92 (Sub-exponential random variables). A random variable X that satisfies the equivalent conditions of Proposition 2.3.14 is a *sub-exponential* random variable.

Sub-Gaussian, Poisson, exponential, Pareto, Levy, Weibull, log-normal, Cauchy, t-distributed random variables are examples of sub-exponential random variables.

Definition 93 (Sub-exponential norm). The *sub-exponential norm* $\|X\|_{\psi_1}$ of random variable X is defined as

$$\|X\|_{\psi_1} := \inf\{t > 0 \ : \ \mathbb{E}[\exp(|X|/t)] \leq 2\}.$$

Proposition 2.3.15. *If X is a sub-exponential random variable, then $X - \mathbb{E}[X]$ is sub-exponential and for a constant $C > 0$:*

$$\|X - \mathbb{E}[X]\|_{\psi_1} \leq C \|X\|_{\psi_1}.$$

Proposition 2.3.16 (Sub-exponential is sub-Gaussian squared)**.** *A random variable X is sub-exponential if and only if X^2 is sub-Gaussian. Moreover,*

$$\left\|X^2\right\|_{\psi_1} := \|X\|_{\psi_2}^2.$$

Proposition 2.3.17 (Product of sub-Gaussians is sub-exponential)**.** *Let X and Y be sub-Gaussian random variables. Then XY is sub-exponential. Moreover,*

$$\|XY\|_{\psi_1} = \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

Theorem 2.3.11 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-exponential random variables with mean zero and let $C > 0$ be a constant. Then, for every $t \geq 0$:*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i\right| \geq t\right] \leq 2\exp\left(-C\min\left\{\frac{t^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right).$$

Theorem 2.3.12 (Bernstein's inequality for weighted sums)**.** *Let $X_1, \ldots, X_n$ be independent sub-exponential random variables with mean zero, $C > 0$ be a constant, $K := \max_{i=1}^{n} \|X_i\|_{\psi_1}$ and $\boldsymbol{a} \in \mathbb{R}^n$. Then:*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} a_i X_i\right| \geq t\right] \leq 2\exp\left(-C\min\left\{\frac{t^2}{K^2\|\boldsymbol{a}\|_2^2}, \frac{t}{K\|\boldsymbol{a}\|_\infty}\right\}\right).$$

Corollary 2.3.13 (Bernstein's inequality for averages)**.** *Let $X_1, \ldots, X_n$ be independent sub-exponential random variables with mean zero, $C > 0$ be a constant and $K := \max_{i=1}^{n} \|X_i\|_{\psi_1}$. Then:*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i/n\right| \geq t\right] \leq 2\exp\left(-Cn\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right).$$

Alphabetical Index