Lecture notes – Statistics B

Modern Statistical Theory

Niklas Pfister

January 30, 2024

Disclaimer: These lecture notes are based on existing work. Chapter 1 on statistical theory is inspired by the lecture notes from Sara van de Geer and Hans-Rudolf Künsch for the lecture 'Fundamentals of Mathematical Statistics' at ETH Zürich. Chapter 2 and 3 on kernel methods and sparsity are heavily based on Rajen Shah's lecture notes for the lecture 'Modern Statistical Methods' at the University of Cambridge.

Acknowledgements

I would like to thank Rajen Shah for providing me both his lecture notes and exercises while developing this course. Furthermore, I would like to thank Sara van de Geer and Hans-Rudolf Künsch for allowing me to use their lecture notes for the course Theoretical Statistics, which has inspired large parts of these notes. Finally, I would like to thank Anton Rask Lundborg and Jeff Adams for proof-reading and helpful feedback.

Contents

Chapter 1

Statistical Theory

1.1 Statistical models

In statistics we mathematically model observable phenomena with stochastic models. The starting point for any statistical analysis is to construct a *statistical model* for the given problem. Such a model consists of the three components; data, a stochastic model and a statistic.

Data We start by assuming some type of data is available (e.g., tabular measurements, images, or time-series observations). In many cases (and in particular throughout this course), we assume the data can be mathematically described as n repeated observations $(x_1, \ldots, x_n) \in \mathcal{X}^n$ lying in the same measurable observation space $\mathcal{X}$. For example, these could be n repeated real-valued measurements of some quantity, in which case the only differences between the x_i's would be the measurement error. convention that $x = (x_1, \ldots, x_n)$ and $X = (X_1, \ldots, X_n)$

Stochastic model We model the data generating process, i.e., the process that produced the data, by a stochastic model. This allows us to use the language and tools of probability theory to model randomness in the data. Furthermore, the stochastic model provides a formal way of encoding prior knowledge that a practitioner might have about the data. Formally, a stochastic model is defined as a subset of probability distributions on $\mathcal{X}^n$, i.e.,

$$\mathcal{P} \subseteq \{P \,|\, P \text{ probability distribution on } \mathcal{X}^n\}.$$

Underlying this modeling philosophy lies the idea that there is a (true) distribution $P_0 \in \mathcal{P}$ such that the data generating process consists of drawing a single random vector $(X_1, \ldots, X_n) \sim P_0$.[1] We use the convention $X = (X_1, \ldots, X_n)$ whenever the sample size n is clear from context. The constraints used to define the model $\mathcal{P}$ and which encode prior knowledge are called *model assumptions*. The following are some prominent examples.

- *Independent and identically distributed assumption:* $X_1, \ldots, X_n$ are assumed to be independent and identically distributed (i.i.d.). In this case it is enough to specify a distribution over X_1 instead of the joint vector.
- *Parametric model assumption:* There exists a parameter set $\Theta \subseteq \mathbb{R}^p$ and a mapping $\theta \mapsto P_\theta$ such that the statistical model can be expressed as $\mathcal{P} = \{P_\theta \,|\, \theta \in \Theta\}$.
- *Density assumption:* There exists a σ-finite measure ν that dominates the statistical model, i.e., for all $P \in \mathcal{P}$ the Radon-Nikodym derivative $p = \frac{dP}{d\nu}$ exists. The Radon-Nikodym derivative p is called a density.

[1] We use the convention that a zero subscript indicates that a quantity is related to the true underlying distribution.

Statistic Our goal in statistics is to infer (part of) the stochastic model. To do so, we construct statistics extracts the aspect of the stochastic model we are interested in from the data. Formally, a statistic is a measurable function $T : \mathcal{X}^n \to \Gamma$ which does not depend on the true data generating distribution. In the statistical literature (and sometimes in these notes) the term statistic can also refer to the random variable resulting from evaluating a statistic T at the random data sample, i.e., $T(X_1, \ldots, X_n)$, instead of the function T itself. Depending on the target of the statistical analysis, we distinguish two main goals.

- *Estimation:* If the goal is to estimate a particular parameter or property of the stochastic model, the statistic is called an *estimator*. This is a function that outputs the best guess for the target of interest based on the observed data.
- *Hypothesis testing:* If the goal is to validate a scientific hypothesis, the statistic is called a *hypothesis test*. This is a function that takes the observations and uses them to output whether or not the hypothesis is supported by the data.

We illustrate these three components with two toy examples.

Example 1.1 (Poisson model)**.** *Consider a small insurance company, which observes a given number of claims each day. Assume we observed the number of claims $X_1, \ldots, X_n$ during $n = 200$ days. A possible model is the Poisson model, in which $X_1, \ldots, X_n$ are i.i.d. and the number of claims on any particular day X_i has a Poisson distribution with parameter $\theta_0 > 0$, that is, for all $k \in \mathbb{N}_0$ it holds that*

$$\mathbb{P}_{\theta_0}(X_i = k) = \frac{\theta_0^k}{k!} e^{-\theta_0}.$$

Imagine we are interested in the probability of at least 4 claims on a particular day and call this γ_0. Then, it holds that

$$\begin{aligned}\gamma_0 &= \mathbb{P}_{\theta_0}(X \geq 4) \\ &= 1 - \mathbb{P}_{\theta_0}(X \leq 3) \\ &= 1 - \left(1 + \theta_0 + \frac{\theta_0^2}{2} + \frac{\theta_0^3}{3!}\right) e^{-\theta_0} \\ &=: g(\theta_0).\end{aligned}$$

We can construct an estimator to estimate γ_0 using the plug-in principle discussed in Section 1.2.1 below. First, observe that the sample average $\hat{\theta}(X) := \frac{1}{n}\sum_{i=1}^n X_i$ is a possible estimator for θ_0. Combining this estimator with the function g defined above

$$\hat{\gamma}(X) := g(\hat{\theta}(X))$$

leads to an estimator for γ_0.

Example 1.2 (Regression model)**.** *Assume we want to find a functional relation between tulip growth and the amount with which it was watered. We have measurements from n different tulips $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $Y_i \in \mathbb{R}$ is the size of the i-th tulip after one month and $X_i \in \mathbb{R}$ is the amount of daily water the i-th tulip was given. A possible model for this is a (conditional expectation) regression model, which is defined as the statistical model $\mathcal{P}$ consisting of all probability distributions over (X_i, Y_i) which satisfy that there exists $\varepsilon_i \sim \mu_0$ such that*

$$Y_i = f_0(X_i) + \varepsilon_i \quad \text{with } \mathbb{E}[\varepsilon_i | X_i] = 0,$$

where f_0 is a function in a pre-specified function space $\mathcal{F}$ and μ_0 is a probability distribution on $\mathbb{R}$ with mean 0. Our goal is to estimate f_0. One way to do this is to use the least-squares estimate given by

$$\hat{f}(X, Y) := \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

The case where the function class $\mathcal{F}$ consists of only linear functions is called linear regression and is discussed in Section 1.3. As explained in Section 4.1.1 the conditional expectation $\mathbb{E}[Y|X]$ minimizes the population least square loss.

1.2 Parametric statistics

In parametric statistics, we make the assumption that there is a parameter space $\Theta \subseteq \mathbb{R}^p$ and a mapping $\theta \mapsto P_\theta$ that describes the stochastic model,

$$\mathcal{P} = \{P_\theta \,|\, \theta \in \Theta\}.$$

Throughout this section, we additionally assume i.i.d. observations and that there is a function $g : \Theta \to \Gamma$ that maps each parameter $\theta \in \Theta$ to a parameter of interest $\gamma = g(\theta)$. We assume that $\Gamma \subseteq \mathbb{R}^d$. Much of statistics deals with constructing and analyzing the performance of an estimator $T : \mathcal{X}^n \to \Gamma$ that estimates the parameter of interest γ.

1.2.1 Constructing estimators

How to construct an estimator generally depends on the application at hand. There are, however, some general principles that work in many settings. A general principle is known as *plug-in estimation*. It is based on the *empirical measure* which is the function $\widehat{P}_n : \mathcal{X}^n \to \{P \,|\, P \text{ probability distribution}\}$ defined for all $x \in \mathcal{X}^n$ by

$$\widehat{P}_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i},$$

where δ_{x_i} is the Dirac measure on $\mathcal{X}$ that puts mass one at x_i and mass zero everywhere else. The empirical measure is a general purpose estimator of the true data generating distribution P_{θ_0}. In fact, for all $\theta \in \Theta$, the strong law of large numbers implies that $\lim_{n\to\infty} \widehat{P}_n(X)(A) = P_\theta(A)$ P_θ-a.s. for all measurable sets $A \subseteq \mathcal{X}$. It is helpful to think of the empirical measure as an extension of the empirical cumulative distribution function $\widehat{F}_n(X)(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}$ from $\mathcal{X} = \mathbb{R}$ to general measurable spaces $\mathcal{X}$.

A plug-in estimator is constructed by combining the empirical measure with a function that maps the true distribution to the parameter of interest. More specifically, assume there is a function

$$Q : \overline{\mathcal{P}} \to \Gamma,$$

where $\overline{\mathcal{P}} = \mathcal{P} \cup (\cup_{n\in\mathbb{N}}\{\widehat{P}_n(x) \,|\, x \in \mathcal{X}^n\})$ and which satisfies for all $\theta \in \Theta$ that $Q(P_\theta) = g(\theta) = \gamma$. Then the estimator defined by

$$T(X) = Q(\widehat{P}_n(X))$$

is called a plug-in estimator. While this construction might seem rather abstract at first, most estimators can be expressed in this form. Under further regularity conditions on the function Q it is possible to prove general results about the asymptotic behavior of these estimators. Developing this theory requires advanced mathematical tools from the field of empirical process theory and is beyond the scope of this course.

We now present three more explicit procedures to construct estimators, each of which are themselves plug-in estimators. For simplicity, we assume that the parameter of interest is the full parameter, i.e., $\Gamma = \Theta$, $\gamma = \theta$ and $p = d$.

Moment estimators This type of estimator is constructed by matching the empirical moments to the population moments. More specifically, assume for simplicity $\mathcal{X} \subseteq \mathbb{R}$. Define for all $j \in \{1, \ldots, d\}$ the j-th population moment for all $\theta \in \Theta$ by

$$\mu_j(P_\theta) \coloneqq \mathbb{E}_\theta[X_i^j] = \int_{\mathcal{X}} z^j P_\theta(dz)$$

and the j-th empirical moment for all $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ by

$$\widehat{\mu}_j(x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} x_i^j = \int_{\mathcal{X}} z^j \widehat{P}_n(x)(dz) = \mu_j(\widehat{P}_n(x)).$$

Furthermore, assume the function $m : \Theta \to \mathbb{R}^d$ defined for all $\theta \in \Theta$ by $m(\theta) \coloneqq (\mu_1(P_\theta), \dots, \mu_d(P_\theta))$ is invertible. Then, the moment estimator $\widehat{\theta}^{\text{moment}} : \mathcal{X}^n \to \Theta$ is defined for all $x \in \mathcal{X}^n$ as

$$\widehat{\theta}^{\text{moment}}(x) \coloneqq m^{-1}(\widehat{\mu}_1(x), \dots, \widehat{\mu}_d(x)).$$

By construction the moment estimator is a plug-in estimator where the function $Q : \overline{P} \to \Theta$ is defined for all $P \in \overline{\mathcal{P}}$ as

$$Q(P) \coloneqq m^{-1}(\mu_1(P), \dots, \mu_d(P)).$$

Maximum likelihood estimators This type of estimator is constructed by choosing the parameter for which the data have the highest likelihood of being observed. To define the likelihood function we assume i.i.d. data and that for all $\theta \in \Theta$ the density p_θ corresponding to P_θ exists. The likelihood function $L : \Theta \times \mathcal{X}^n \to \mathbb{R}$ is defined for all $(\theta, x) \in \Theta \times \mathcal{X}^n$ as

$$L(\theta, x) \coloneqq \prod_{i=1}^{n} p_\theta(x_i). \tag{1.2.1}$$

It quantifies how likely it is to observe $(X_1, \dots, X_n) = (x_1, \dots, x_n)$ if the true data generating mechanism satisfies $X_i \overset{iid}{\sim} P_\theta$. The maximum likelihood estimator (MLE) is defined for all $x \in \mathcal{X}^n$ by the optimization

$$\widehat{\theta}^{\text{MLE}}(x) \coloneqq \text{argmax}_{\theta \in \Theta} L(\theta, x).$$

Additional regularity conditions, which we do not discuss here, are required for the MLE to exist and be unique. It is common to optimize over the log-likelihood function $\ell : (\theta, x) \mapsto \log(L(\theta, x))$ instead as this does not effect the location of the optimum but turns the product in (1.2.1) into a sum. In the exercises you will show that this is also a plug-in estimator, where the Q function is defined for all $P \in \overline{\mathcal{P}}$ by $Q(P) \coloneqq \arg\max_{\theta \in \Theta} \mathbb{E}_P[\log(p_\theta(X_i))]$.

Minimal risk estimators This type of estimator is constructed by defining an (empirical) risk function $\mathcal{R} : \Theta \times \mathcal{X}^n \to \mathbb{R}$ which for all $x \in \mathcal{X}^n$ maps each potential parameter $\theta \in \Theta$ to a value $\mathcal{R}(\theta, x)$ quantifying how good the data x fits the distribution P_θ. We require that the risk function does not depend on the true data generating parameter θ_0. We further denote by $\theta \mapsto \mathbb{E}_{\theta_0}[\mathcal{R}(\theta, X)]$ the population risk. One can then construct an estimator by minimizing the risk over all possible parameters

$$\widehat{\theta}^{\text{risk}}(x) \coloneqq \arg\min_{\theta \in \Theta} \mathcal{R}(\theta, x).$$

Existence and uniqueness of this optimization problem requires additional assumption. Often the risk has the form $\mathcal{R}(\theta, x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta, x_i)$ for some (observation-wise) loss $\mathcal{L} : \Theta \times \mathcal{X} \to \mathbb{R}$. In that case, the population risk corresponds to the expected loss $\theta \mapsto \mathbb{E}_{\theta_0}[\mathcal{L}(\theta, X_1)]$.

The MLE is just a special case of a minimal risk estimator, which can be seen by taking the negative log-likelihood as loss function. Minimal risk estimators often appear in the context of regression models. A common risk in those cases is the mean squared prediction error given by $\mathcal{L}(f, (x, y)) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2$. See Section 1.3.1 for an example.

1.2.2 Classical optimality theory

In this section, we consider several ways of assessing the quality of an estimator. Let T be an estimator for the parameter of interest $\gamma = g(\theta) \in \mathbb{R}^d$. Given that $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ is the true parameter, three fundamental criteria for quantify the quality of an estimator are the following.

- *Bias*

$$\text{bias}_{\theta_0}(T) \coloneqq \mathbb{E}_{\theta_0}[T(X)] - g(\theta_0).$$

- *Variance*

$$\text{Var}_{\theta_0}(T) \coloneqq \text{Cov}_{\theta_0}(T(X)) = \mathbb{E}_{\theta_0}\big[T(X) \cdot T(X)^\top\big] - \mathbb{E}_{\theta_0}\big[T(X)\big] \cdot \mathbb{E}_{\theta_0}\big[T(X)\big]^\top.$$

- *Mean squared error*

$$\mathrm{MSE}_{\theta_0}(T) \coloneqq \mathbb{E}_{\theta_0}\Big[\big(T(X) - g(\theta_0)\big)^\top \big(T(X) - g(\theta_0)\big)\Big].$$

An estimator T satisfying $\mathrm{bias}_\theta(T) = 0$ for all $\theta \in \Theta$ is called *unbiased*. Since the bias and mean squared error explicitly depend on the true parameter, they cannot be estimated by simply replacing P_{θ_0} with $\widehat{P}_n(x)$. All three quantities are connected by the *bias-variance decomposition* which states for all $\theta \in \Theta$ that

$$\mathrm{MSE}_\theta(T) = \mathrm{bias}_\theta(T)^\top \mathrm{bias}_\theta(T) + \mathrm{trace}\big(\mathrm{var}_\theta(T)\big).$$

Ideally we want all three quantities to be small. The bias-variance decomposition can be useful in this respect as it describes how the quantities are related to each other. Often, minimizing either the bias or the variance leads to an increase in the other quantity, which is known as the *bias-variance trade-off*. The three quantities provide a useful method for comparing two estimators and can be used to define notions of optimality.

- *Minimax optimal MSE:* An estimator T of $g(\theta)$ is said to have minimax optimal MSE if

$$\sup_{\theta\in\Theta} \mathrm{MSE}_\theta(T) = \inf_{\tilde{T}} \sup_{\theta\in\Theta} \mathrm{MSE}_\theta(\tilde{T}),$$

  where the infimum is taken over all estimators $\tilde{T}$ of $g(\theta)$.

- *Uniform minimum variance unbiased (UMVU):* An unbiased estimator T of $g(\theta)$ is called UMVU if among all other unbiased estimators it has the smallest variance uniformly across all parameters. Formally, for all other unbiased estimators $\tilde{T}$ of $g(\theta)$ it holds that

$$\forall \theta \in \Theta: \quad \mathrm{Var}_\theta(\tilde{T}) - \mathrm{Var}_\theta(T) \text{ is positive semi-definite.}$$

A lot of work has gone into analyzing different notions of optimality and how to compare estimators. To showcase this, we present the Cramér-Rao lower bound, which is a famous result related to UMVU estimators. For simplicity, we assume sufficient regularity of the statistical model (e.g., that densities p_θ exist and $\theta \mapsto p_\theta$ is "smooth" with uniformly bounded derivatives). A key ingredient for the Cramér-Rao lower bound and many other statistical analyses (in particular involving MLEs) is *the score function* $S_\theta : \mathcal{X}^n \to \mathbb{R}^p$ which is defined for all $x \in \mathcal{X}^n$ by

$$S_\theta(x) \coloneqq \nabla_\theta \log(L(\theta, x)),$$

where L is the likelihood function defined in (1.2.1). It is sometimes also customary to define the score function using only a single sample, i.e., $s_\theta(x_i) \coloneqq \nabla_\theta \log(p_\theta(x_i))$.[2] Given a data realization $x \in \mathcal{X}^n$, the value of the score function $S_\theta(x)$ quantifies how sensitive the likelihood is at each parameter value θ – large score values imply that small changes in the parameter lead to large changes in the likelihood. Based on the score function, the *Fisher information* is defined as the function $\mathcal{I} : \Theta \to \mathbb{R}^{p\times p}$ satisfying for all $\theta \in \Theta$ that

$$\mathcal{I}(\theta) \coloneqq \mathrm{Cov}_\theta(S_\theta(X)).$$

We denote by $i : \theta \mapsto \frac{1}{n}\mathcal{I}(\theta)$ the single sample Fisher information (here we implicitly assume the samples are i.i.d.). The Fisher information quantifies the information the data contain about the parameter θ. It can be shown, given sufficient regularity, that the following identities hold

$$\mathbb{E}_\theta[S_\theta(X)] = 0 \quad \text{and} \quad \mathcal{I}(\theta) = -\mathbb{E}_\theta\big[(D_\theta S_\theta)(X)\big], \tag{1.2.2}$$

[2] For independent samples the two definitions differ by a factor of n, that is, $S_\theta(x) = \sum_{i=1}^n s_\theta(x_i) = n s_\theta(x_1)$. Therefore you should always remember to check which definition is used.

where $D_\theta S_\theta$ corresponds to the Jacobian matrix of S_θ with respect to θ and evaluated at θ. To get a better intuition about the score function and Fisher information it helps to think about maximum likelihood estimation. In order to compute the MLE one maximizes the log-likelihood. The MLE therefore lies at a point at which the derivative of the log-likelihood (i.e., the score function) is zero and for which the Hessian matrix (second derivative matrix) of the log-likelihood (i.e., the derivative of the score function) is negative definite, which indicates a maximum. By (1.2.2) the Fisher information is equal to the expected curvature of the maximum of the log-likelihood. This means that a larger Fisher information implies a more peaked maximum which makes it easier to estimate the parameter.

Using the Fisher information, we can state the Cramér-Rao lower bound which gives a lower bound on the variance of an unbiased estimator.

Theorem 1.3 *(Cramér-Rao lower bound).*
Let T be an unbiased estimator of $g(\theta)$ and assume sufficient regularity on the statistical model. Then it holds for all $\theta \in \Theta$ that

$$\mathrm{Var}_\theta(T) - \big(D_\theta g(\theta)\big)\mathcal{I}(\theta)^{-1}\big(D_\theta g(\theta)\big)^\top$$

is positive semi-definite.

Under the i.i.d. assumption and for $\Theta = \mathbb{R}$ and $g(\theta) = \theta$, the bound implies that $\mathrm{Var}_\theta(T) \geq \mathcal{I}(\theta)^{-1} = \frac{1}{n} i(\theta)^{-1}$. This shows that the lowest achievable variance for any unbiased estimator is of order $\frac{1}{n}$. By showing that an unbiased estimator achieves the lower bound it is possible to prove that it is a UMVU estimator.

1.2.3 Classical asymptotic theory - large n

Classical asymptotic theory deals with settings in which the observation space $\mathcal{X}$ remains fixed and the number of observations n tends to infinity. The idea behind this asymptotic regime is to make statements about what effect additional data collection has on the statistical analysis. For example, one might be interested in answering the following question.

How much does an estimator improve if the sample size is doubled?

To answer such questions, one assumes that one can construct a sequence of estimators $(T_n)_{n\in\mathbb{N}}$ and then analyzes the limit of $T_n(X_1, \ldots, X_n)$ as n tends to infinity. The following two properties are important:

- **Consistency** A sequence $(T_n)_{n\in\mathbb{N}}$ is called a *(weakly) consistent* estimator of $g(\theta)$ if it holds that
$$T_n(X_1, \ldots, X_n) \xrightarrow{P_\theta} g(\theta) \quad \text{as } n \to \infty.$$
If the convergence holds P_θ-almost surely instead, this is also called strongly consistent. Proving these types of statements generally relies on an application of a version of the (strong) law of large numbers.

- **Asymptotic normality** A sequence $(T_n)_{n\in\mathbb{N}}$ is called an *asymptotically normal* estimator of $g(\theta)$ if there exists a positive semi-definite $\Sigma \in \mathbb{R}^{d\times d}$ such that
$$\sqrt{n}\big(T_n(X_1, \ldots, X_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}(0, \Sigma) \quad \text{as } n \to \infty.$$
The matrix Σ is called the *asymptotic variance*. This type of statement is generally proved by using a version of the central limit theorem.

It is possible to prove consistency and asymptotic normality for quite general classes of estimators. The following theorem shows that given sufficient regularity MLEs are asymptotically normal.

Theorem 1.4 *(Asymptotic normality of MLE).*
Let $(\widehat{\theta}_n^{\mathrm{MLE}})_{n\in\mathbb{N}}$ be a sequence of maximum likelihood estimators of θ and assume sufficient regularity on the statistical model. Then it holds that

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{MLE}} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, i(\theta)^{-1}\right) \quad \text{as } n \to \infty.$$

1.2.4 Outlook: Statistical decision theory

The ideas introduced above are extended and unified in statistical decision theory. The details go beyond the scope of this course but for the sake of completeness we provide a short introduction here. The mathematical setup of statistical decision theory is similar to what we introduced in Section 1.1 but the terminology is slightly different. The main ingredients are the following.

(i) A statistical model as described in Section 1.1.

(ii) An action space $\mathcal{A}$ that depends on the statistical objective (e.g., $\mathcal{A} = \mathbb{R}$ for estimating a real-valued parameter or $\mathcal{A} = \{0, 1\}$ for testing a hypothesis). This corresponds to the space of the parameter of interest Γ in our previous terminology.

(iii) A set of decisions $\mathcal{D} \subseteq \{\delta \mid \delta : \mathcal{X}^n \to \mathcal{A} \text{ measurable}\}$. Decisions correspond to statistics in our previous terminology.

(iv) A loss function $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}$, where $\mathcal{L}(\theta, a)$ quantifies the loss of taking action a when the true parameter is θ. For example, if $\mathcal{A} = \mathbb{R}$ and $\Theta = \mathbb{R}$ the squared error loss is given for all $\theta \in \mathbb{R}$ and $a \in \mathbb{R}$ by

$$\mathcal{L}(\theta, a) = (\theta - a)^2.$$

The decision theory loss function is slightly different from the one defined in Section 1.2.1, which quantifies how well a data observation x_i fits to a given parameter instead of an action.

(v) The risk function $\mathcal{R} : \Theta \times \mathcal{D} \to \mathbb{R}$ is defined for all $\theta \in \Theta$ and $\delta \in \mathcal{D}$ by

$$\mathcal{R}(\theta, \delta) \coloneqq \mathbb{E}_\theta[\mathcal{L}(\theta, \delta(X))].$$

Hence $\mathcal{R}(\theta, \delta)$ quantifies the risk of decision δ if θ is the true parameter. For example, for the squared error loss and $\mathcal{A} = \mathbb{R}$ and $\Theta = \mathbb{R}$ the risk correspond to the MSE, that is,

$$\mathcal{R}(\theta, \delta) \coloneqq \mathbb{E}_\theta[(\theta - \delta(X))^2].$$

Based on these ingredients, one can formally define various desirable properties and optimality conditions of a decision (e.g., an estimator).

- A decision $\delta \in \mathcal{D}$ is called *inadmissible* if there exists $\delta' \in \mathcal{D}$ such that

$$\forall \theta \in \Theta : \mathcal{R}(\theta, \delta') \leq \mathcal{R}(\theta, \delta) \quad \text{and} \quad \exists \theta \in \Theta : \mathcal{R}(\theta, \delta') < \mathcal{R}(\theta, \delta).$$

Otherwise it is called *admissible*. A surprising result due to Stein is that the ordinary least squares estimator can be inadmissible in certain settings (see James-Stein estimator).

- A decision $\delta \in \mathcal{D}$ is called *minimax optimal* if

$$\sup_{\theta\in\Theta} R(\theta, \delta) = \inf_{\delta'\in\mathcal{D}} \sup_{\theta\in\Theta} R(\theta, \delta').$$

In Section 2.3.2 we will see that the kernel ridge regression estimator is minimax optimal up to a constant factor.

- A decision $\delta \in \mathcal{D}$ is called *Bayes estimator* for a distribution π over Θ if

$$\mathbb{E}_{\theta\sim\pi}[\mathcal{R}(\theta,\delta)] = \inf_{\delta'\in\mathcal{D}} \mathbb{E}_{\theta\sim\pi}[\mathcal{R}(\theta,\delta')].$$

  Instead of considering the worst-case parameter as for minimax optimality, a Bayes estimator minimizes an averaged risk (called Bayes risk) with respect to a distribution π over the parameter space (the prior).

1.3 Linear regression and ordinary least squares

In this section, we introduce the *linear regression model*, which will appear several times as a case-study throughout the course. Consider paired data $(X_1, Y_1), \dots, (X_n, Y_n)$, where $X_i \in \mathbb{R}^p$ is called the vector of predictors and Y_i the response. We use the matrix notation $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times 1}$ in which the observations have been aggregated row-wise and call X the *design matrix*. A linear regression model assumes that there exists a regression parameter $\beta_0 \in \mathbb{R}^{p\times 1}$ and a noise distribution $\mu_0 \in \mathcal{P}_{\text{noise}}$, where $\mathcal{P}_{\text{noise}}$ is a set of distributions on $\mathbb{R}^n$, such that the joint distribution of (X, Y) satisfies that there exists $\varepsilon \sim \mu_0$ and

$$Y = X\beta_0 + \varepsilon \quad \text{with } \mathbb{E}[\varepsilon|X] = 0. \tag{1.3.1}$$

Formally, the stochastic model $\mathcal{P}$ consists of all distributions over (X, Y) for which there exists a parameter $\beta_0 \in \mathbb{R}^{p\times 1}$ and a noise distribution $\mu_0 \in \mathcal{P}_{\text{noise}}$ such that the data are generated by (1.3.1). This type of model is called *semi-parametric* since parts of it are specified by the parameter β_0 while the remaining parts, in this case the distributions of μ_0 and X, are not parameterized. In the condition (1.3.1), the noise ε is stated explicitly, this can be avoided by simply assuming the joint distribution of (X, Y) satisfies

$$\mathbb{E}[Y|X] = X\beta_0.$$

In this formulation the noise is indirectly specified by $\varepsilon := Y - \mathbb{E}[Y|X]$. We distinguish between two types of approaches when analyzing regression models.

(1) The *fixed-design* approach which assumes X is non-random.

(2) The *random-design* approach which explicitly models the randomness of X.

In this course, we will mostly focus on the fixed-design setting as this makes many technical considerations a lot easier. In both the fixed- and random-design case, it is common to further constrain the allowed noise distributions $\mathcal{P}_{\text{noise}}$ to make technical considerations easier.

Linear model with i.i.d. noise These are linear regression models in which the noise terms are i.i.d. across observations, i.e., one assumes there exists a noise distribution μ_0 on $\mathbb{R}$ such that the noise term $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ satisfies that $\varepsilon_i \overset{i.i.d.}{\sim} \mu_0$. Under this assumption it is sufficient to specify the equation of a single observation,

$$Y_i = X_i^\top \beta_0 + \varepsilon_i \quad \text{with } E[\varepsilon_i|X_i] = 0.$$

Gaussian linear model These are linear regression models in which the noise terms follow a multivariate Gaussian distribution. More specifically, assume there exists a positive definite matrix $\Sigma_0 \in \mathbb{R}^{p\times p}$ such that

$$\varepsilon \sim \mathcal{N}(0, \Sigma_0).$$

In the fixed-design case this model is fully parameterized by the regression parameter β_0 and the noise covariance Σ_0. The Gaussian assumption makes many theoretical considerations substantially easier. For example, it allows for a simple expression of the MLE for β_0 (see Section 1.3.1

below). Often one additionally combines the Gaussian assumption with the i.i.d. noise assumption. The model is then fully specified – using a single additional noise parameter $\sigma_0 \in (0, \infty)$ – by the equation

$$Y_i = X_i^\top \beta_0 + \varepsilon_i \quad \text{with } E[\varepsilon_i | X_i] = 0 \text{ and } \varepsilon_i \sim \mathcal{N}(0, \sigma_0^2).$$

The parameter of interest in linear regression is generally the regression parameter β_0. The *ordinary least squares* estimator is the most fundamental approach for estimating this parameter and is an important building block for the theory discussed in this course.

1.3.1 Ordinary (linear) least squares

Consider the fixed-design setting and assume $X^\top X$ is invertible (which implies $n > p$) and denote by β_0 the true underlying regression parameter. Ordinary least squares (OLS) estimates the regression parameter β_0 by minimizing the mean squared prediction error,

$$\widehat{\beta}^{\mathrm{OLS}} := \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 = \left(X^\top X\right)^{-1} X^\top Y.$$

Assuming that the noise terms are i.i.d. and satisfy $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_0^2$, it holds that

- $\mathbb{E}\big[\widehat{\beta}^{\mathrm{OLS}}\big] = \mathbb{E}\big[(X^\top X)^{-1} X^\top (X\beta^0 + \varepsilon)\big] = \beta_0$
- $\mathrm{Var}\left(\widehat{\beta}^{\mathrm{OLS}}\right) = (X^\top X)^{-1} X^\top \mathrm{Cov}(\varepsilon) X (X^\top X)^{-1} = \sigma_0^2 (X^\top X)^{-1}$.

Moreover, it can be shown that the OLS estimator has minimum variance among all linear unbiased estimators. This result is known as the Gauss-Markov theorem and makes no further assumptions on the noise distributions.

Theorem 1.5 *(Gauss-Markov Theorem).*
Assume a linear regression model with i.i.d. noise $(\varepsilon_1, \ldots, \varepsilon_n)$ that satisfies $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_0^2$. Then, for all unbiased estimators $\widehat{\beta} := AY$, where $A \in \mathbb{R}^{p \times n}$, it holds that

$$\mathrm{Var}\left(\widehat{\beta}\right) - \mathrm{Var}\left(\widehat{\beta}^{\mathrm{OLS}}\right)$$

is positive semi-definite.

If one additionally assumes a Gaussian noise distribution, i.e., $\varepsilon_i \sim \mathcal{N}(0, \sigma_0^2)$, it can be shown that $\widehat{\beta}^{\mathrm{OLS}}$ is the MLE. To see this, first compute the log-likelihood which is given by

$$\ell((\beta, \sigma^2), (X, Y)) = -\tfrac{n}{2} \log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2$$

and then show that the maximum is attained at $\widehat{\beta}^{\mathrm{OLS}}$. Furthermore, the Fisher information is given by

$$\mathcal{I}(\beta, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} X^\top X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Observe that we used σ^2 as a parameter and not σ. Therefore, by Theorem 1.4 it immediately follows that

$$\sqrt{n}\big(\widehat{\beta}^{\mathrm{OLS}} - \beta_0\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma_0^2 \Sigma_X^{-1}\big) \quad \text{as } n \to \infty,$$

where $\Sigma_X := \lim_{n\to\infty} \frac{1}{n} X^\top X$ is assumed to exist. One can also directly use the central limit theorem to prove this statement – also without assuming Gaussian noise distributions. However, if the Gaussian noise assumption is true the much stronger finite sample result

$$\widehat{\beta}^{\mathrm{OLS}} \sim \mathcal{N}\big(\beta_0, \sigma_0^2 (X^\top X)^{-1}\big)$$

is also true, which can be shown the closed-form expression of $\widehat{\beta}^{\mathrm{OLS}}$ and basic properties of the Gaussian distribution.

Chapter 2

Kernel Methods

In this chapter, we introduce a versatile class of non-parametric statistical procedures called *kernel methods*. They extend many classical parametric methods such as linear regression, principal component analysis and clustering to high-dimensional or otherwise complex data structures. The key ingredient for these methods is a kernel function (introduced in Section 2.2.2) that maps the input data into a well-behaved feature space and allows for efficient computations in that space.

Here, we focus on how to use kernel methods to extend linear regression to (additive noise) non-linear regression, that is, models of the form $Y = f_0(X) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$, X non-random and f_0 a non-linear function. The theory, however, is much more general and in Section 2.4 we discuss some of the further applications that can be solved with kernel methods.

2.1 Ridge regression

We start with the fixed-design linear regression model with i.i.d. noise, that is

$$Y = X\beta_0 + \varepsilon,$$

where $Y \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^{n\times 1}$ with ε_i i.i.d. distributed with mean zero distribution μ_0. As seen in Section 1.3.1, β_0 can be estimated with the OLS estimator

$$\widehat{\beta}^{\mathrm{OLS}} = (X^\top X)^{-1} X^\top Y$$

as long as $X^\top X$ is invertible. This estimator has two shortcomings:

(i) $X^\top X$ is not necessarily invertible. In particular, since $\operatorname{rank}(X^\top X) \leq \min(p, n)$, the matrix $X^\top X$ can only be invertible if $p < n$.

(ii) The Gauss-Markov theorem (Theorem 1.5) only ensures that the OLS estimator is MSE optimal among all linear and unbiased estimators, but there could be non-linear or biased estimator with lower MSE.

Ridge regression estimator Both shortcomings can be addressed by adapting the OLS estimator to include a penalty term on the parameter. Formally, we consider for all $\lambda \in [0, \infty)$ the following estimators

$$\widehat{\beta}^{\mathrm{R}}_\lambda := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{2.1.1}$$

We call $\widehat{\beta}^{\mathrm{R}}_\lambda$ the ridge regression estimator. Several remarks are in order:

- The term $\lambda\|\beta\|_2^2$ is called the *penalty* or *regularization* term and pushes β towards zero. In particular, for increasing λ the ridge regression estimator converges to 0, while for λ going to zero it converges to the OLS estimator (if it exists).

- As in the case of the OLS, we assume $\mathbb{E}[\varepsilon_i] = 0$ and hence ignore fitting an intercept term. If this is not satisfied, we can first center (X_i, Y_i), that is, $(\tilde{X}_i, \tilde{Y}_i) = (X_i, Y_i) - \frac{1}{n}\sum_{\ell=1}^{n}(X_\ell, Y_\ell)$ and then proceeding with the centered data $(\tilde{X}, \tilde{Y})$ as discussed below. This corresponds to adding an intercept term c to the mean squared error part of the loss in (2.1.1) (i.e., $\|Y - X\beta - c\|_2^2$) but not to the penalty term.
- By explicitly minimizing the objective, it can be shown (exercise) that
$$\widehat{\beta}_\lambda^{\mathrm{R}} = (X^\top X + \lambda \mathrm{I})^{-1} X^\top Y, \tag{2.1.2}$$
where I denotes the $(n \times n)$-identity matrix.
- The ridge regression estimator is a linear and (generally) biased estimator, which can be seen from (2.1.2).
- The ridge regression estimator is not invariant to scaling of X. More specifically, scaling a single coordinate of X affects the solution of (2.1.1) since the mean squared error term remains fixed, e.g.,
$$(cX^1, X^2, \ldots, X^d) \cdot (\tfrac{1}{c}\beta^1, \beta^2, \ldots, \beta^d)^\top = X\beta$$
but the penalty term changes. To avoid this non-invariance, we use the convention that each column of X is centered and scaled such that for all $j \in \{1, \ldots, p\}$ it holds that $\|X^j\|_2 = \sqrt{n}$.

Does the ridge regression estimator solve shortcomings (i) and (ii) above?

First, consider (i) and assume we are in the high-dimensional case (i.e., $n < p$). Then, for the ridge regression estimator to exist, we require
$$(\underbrace{X^\top X}_{\text{rank} \leq n} + \lambda \mathrm{I}) \in \mathbb{R}^{p \times p}$$
to be invertible. This is actually true for all $\lambda > 0$ as the following argument shows. Let $A \in \mathbb{R}^{p \times p}$ be an arbitrary symmetric matrix and define
$$R_A(x) \coloneqq \frac{x^\top A x}{x^\top x}. \qquad \text{("Rayleigh-Ritz" quotient)}$$
By the min-max theorem it holds that
$$\lambda_{\min}(A) = \min_{x \neq 0} R_A(x),$$
$$\lambda_{\max}(A) = \max_{x \neq 0} R_A(x),$$
where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimal and maximal eigenvalue of A, respectively. Now fix $A = X^\top X$ and $B = \lambda \mathrm{I}$, then
$$\begin{aligned}
\lambda_{\min}(A + B) &= \min_{x \neq 0} R_{A+B}(x) \\
&= \min_{x \neq 0} \left(\frac{x^\top A x}{x^\top x} + \frac{x^\top B x}{x^\top x} \right) \\
&\geq \min_{x \neq 0} R_A(x) + \min_{x \neq 0} R_B(x) \\
&= \lambda_{\min}(A) + \lambda_{\min}(B) \\
&\geq \lambda > 0.
\end{aligned} \tag{2.1.3}$$
This implies that the minimal eigenvalue of $A + B$ is strictly positive, which further implies that $A + B$ is invertible and the ridge regression estimator exists.

Second, consider (ii) and assume we are in the low-dimensional case (i.e., $n > p$) and $X^\top X$ is invertible. Then, the OLS estimator exists and we can compare the MSE of both estimators. The following theorem shows that for appropriately chosen λ the ridge regression estimator can be lead to a smaller MSE than the OLS estimator.

Theorem 2.1.
For all $\lambda \in (0, 2\frac{\sigma_0^2}{\|\beta_0\|_2^2})$, it holds that

$$\mathbb{E}\left[(\widehat{\beta}^{\mathrm{OLS}} - \beta_0)(\widehat{\beta}^{\mathrm{OLS}} - \beta_0)^\top\right] - \mathbb{E}\left[(\widehat{\beta}_\lambda^{\mathrm{R}} - \beta_0)(\widehat{\beta}_\lambda^{\mathrm{R}} - \beta_0)^\top\right]$$

is (strictly) positive definite.

Proof. Denote by $\theta_0 = (\beta_0, \sigma_0^2)$ the two (true) parameters. Define the term we are interested in and expand it as follows

$$\begin{aligned}(*) &\coloneqq \mathbb{E}\left[(\widehat{\beta}^{\mathrm{OLS}} - \beta_0)(\widehat{\beta}^{\mathrm{OLS}} - \beta_0)^\top\right] - \mathbb{E}\left[(\widehat{\beta}_\lambda^{\mathrm{R}} - \beta_0)(\widehat{\beta}_\lambda^{\mathrm{R}} - \beta_0)^\top\right] \\ &= \mathrm{Var}_{\theta_0}(\widehat{\beta}^{\mathrm{OLS}}) - \left(\mathrm{Var}_{\theta_0}(\widehat{\beta}_\lambda^{\mathrm{R}}) - \mathrm{bias}_{\theta_0}(\widehat{\beta}_\lambda^{\mathrm{R}})\,\mathrm{bias}_{\theta_0}(\widehat{\beta}_\lambda^{\mathrm{R}})^\top\right). \end{aligned} \tag{2.1.4}$$

The last expression is a version of the bias-variance trade-off discussed in Section 1.2.2. We now consider the bias and variance terms separately. First, the bias term is given by,

$$\begin{aligned}\mathrm{bias}_{\theta_0}(\widehat{\beta}_\lambda^{\mathrm{R}}) &= \mathbb{E}[\widehat{\beta}_\lambda^{\mathrm{R}}] - \beta_0 \\ &= (X^\top X + \lambda \mathrm{I})^{-1} X^\top (X\beta_0 + \mathbb{E}[\varepsilon]) - \beta_0 \\ &= (X^\top X + \lambda \mathrm{I})^{-1} (X^\top X + \lambda \mathrm{I} - \lambda \mathrm{I})\beta_0 - \beta_0 \\ &= -\lambda (X^\top X + \lambda \mathrm{I})^{-1} \beta_0.\end{aligned}$$

Second, the variance term is given by,

$$\begin{aligned}\mathrm{Var}_{\theta_0}(\widehat{\beta}_\lambda^{\mathrm{R}}) &= \mathbb{E}\left[(\widehat{\beta}_\lambda^{\mathrm{R}} - \mathbb{E}[\widehat{\beta}_\lambda^{\mathrm{R}}])(\widehat{\beta}_\lambda^{\mathrm{R}} - \mathbb{E}[\widehat{\beta}_\lambda^{\mathrm{R}}])^\top\right] \\ &= \mathbb{E}\left[(X^\top X + \lambda \mathrm{I})^{-1} X^\top \varepsilon \varepsilon^\top X (X^\top X + \lambda \mathrm{I})^{-1}\right] \\ &= \sigma_0^2 (X^\top X + \lambda \mathrm{I})^{-1} X^\top X (X^\top X + \lambda \mathrm{I})^{-1}.\end{aligned}$$

Hence, combining expressions for the bias and variance terms in (2.1.4), we get that

$$\begin{aligned}(*) &= \sigma_0^2 (X^\top X)^{-1} - \sigma_0^2 (X^\top X + \lambda \mathrm{I})^{-1} X^\top X (X^\top X + \lambda \mathrm{I})^{-1} \\ &\qquad - \lambda^2 (X^\top X + \lambda \mathrm{I})^{-1} \beta_0 \beta_0^\top (X^\top X + \lambda \mathrm{I})^{-1} \\ &= \sigma_0^2 M M^{-1} (X^\top X)^{-1} M^{-1} M - \sigma_0^2 M X^\top X M - \lambda^2 M \beta_0 \beta_0^\top M \\ &= M\left[\sigma_0^2 (X^\top X + 2\lambda \mathrm{I} + \lambda^2 (X^\top X)^{-1}) - \sigma_0^2 X^\top X - \lambda^2 \beta_0 \beta_0^\top\right] M \\ &= \lambda M\left[\sigma_0^2 (2\,\mathrm{I} + \lambda (X^\top X)^{-1}) - \lambda \beta_0 \beta_0^\top\right] M,\end{aligned}$$

where $M = (X^\top X + \lambda \mathrm{I})^{-1}$. So $(*)$ is strictly positive definite if and only if

$$\sigma_0^2 (2\,\mathrm{I} + \lambda (X^\top X)^{-1}) - \lambda \beta_0 \beta_0^\top$$

is strictly positive definite. Using similar arguments as in (2.1.3), we get that

$$\begin{aligned}\lambda_{\min}(\sigma_0^2 (2\,\mathrm{I} + \lambda (X^\top X)^{-1}) - \lambda \beta_0 \beta_0^\top) &\geq \sigma_0^2 \lambda_{\min}(2\,\mathrm{I} + \lambda (X^\top X)^{-1}) - \lambda \cdot \lambda_{\max}(\beta_0 \beta_0^\top) \\ &\geq \sigma_0^2 2 - \lambda \|\beta_0\|_2^2,\end{aligned}$$

where in the last step we used that $X^\top X$ is positive semi-definite and $\lambda_{\max}(\beta_0 \beta_0^\top) = \|\beta_0\|_2^2$. Therefore, we get that $(*)$ is strictly positive definite if $0 < \lambda < 2\frac{\sigma_0^2}{\|\beta_0\|_2^2}$. This completes the proof of Theorem 2.1. □

To see why Theorem 2.1 addresses problem (ii), that is, it improves over the OLS in terms of MSE, define $M(\widehat{\beta}) \coloneqq \mathbb{E}[(\widehat{\beta} - \beta_0)(\widehat{\beta} - \beta_0)^\top]$. Then, it holds that

$$\mathrm{MSE}(\widehat{\beta}) = \mathrm{trace}(M(\widehat{\beta})) = \sum_{j=1}^{d} e_j^\top M(\widehat{\beta}) e_j.$$

Hence by Theorem 2.1 we get that

$$\mathrm{MSE}(\widehat{\beta}^{\mathrm{OLS}}) - \mathrm{MSE}(\widehat{\beta}^{\mathrm{R}}_\lambda) = \sum_{j=1}^{d} e_j^\top (M(\widehat{\beta}^{\mathrm{OLS}}) - M(\widehat{\beta}^{\mathrm{R}}_\lambda)) e_j > 0.$$

2.1.1 Comparison of OLS and ridge regression based on SVD

We now provide further intuition on the ridge regression estimator by comparing it to the OLS estimator. For this, we use the singular value decomposition (SVD) summarized in Appendix A.1.

We consider the low dimensional setting and assume $X^\top X$ is invertible. Then, using the thin SVD $X = UDV^\top$, we can express the predicted values of the OLS estimator as follows

$$\begin{aligned} X\widehat{\beta}^{\mathrm{OLS}} &= X(X^\top X)^{-1} X^\top Y \\ &= UDV^\top (VDU^\top UDV^\top)^{-1} VDU^\top Y \\ &= UDV^\top (VD^2V^\top)^{-1} VDU^\top Y \\ &= UU^\top Y \\ &= \sum_{j=1}^{p} \underbrace{U^j (U^j)^\top Y}_{\text{projection of Y onto } U^j}. \end{aligned}$$

Similarly we can express the predicted values of the ridge regression estimator as

$$\begin{aligned} X\widehat{\beta}^{\mathrm{R}} &= X(X^\top X + \lambda \mathrm{I})^{-1} X^\top Y \\ &= UDV^\top (VD^2V^\top + \lambda \mathrm{I})^{-1} VDU^\top Y \\ &= UDV^\top V (D^2 + \lambda \mathrm{I})^{-1} V^\top VDU^\top Y \\ &= UD(D^2 + \lambda \mathrm{I})^{-1} DU^\top Y \\ &= \sum_{j=1}^{p} U^j \underbrace{\frac{d_j^2}{d_j^2 + \lambda}}_{\text{scaling factor}} (U^j)^\top Y, \end{aligned}$$

where $d_j = D_j^j$ is the j-th diagonal element. Now if d_j is small the scaling factor is also small, while in contrast if d_j is large the scaling factor is close to one. Therefore the ridge regression estimator shrinks the contribution of principal components corresponding to small singular values and focuses on the principal components with large singular values.

2.2 Non-linear feature maps and kernels

We now extend ridge regression to non-linear regression. The extension is based on transforming the predictors X using non-linear feature maps. The idea is that the non-linear regression function becomes a linear function on the transformed predictors. The two-step approach of transforming the predictors and then applying linear methodology has many applications in statistics, e.g., smoothing splines and generalized additive models.

As our motivating example, we consider the fixed-design non-linear regression model with i.i.d. noise given for all $i \in \{1, \ldots, n\}$ by

$$Y_i = f_0(X_i) + \varepsilon_i, \tag{2.2.1}$$

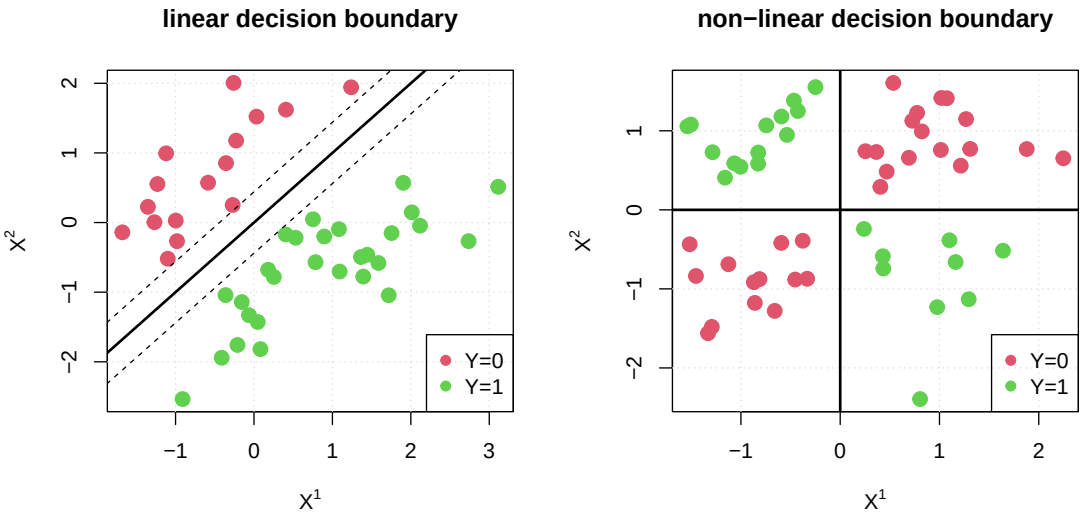

Figure 2.1: Data for two classification tasks given in Example 2.2. In the left plot, corresponding to (i), the decision boundary is linear, while in the right plot, corresponding to (ii), it is non-linear.

with $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} \mu_0$ zero mean and $f_0 \in \mathcal{F}$ with $\mathcal{F} \subseteq \{f \mid f : \mathbb{R}^p \to \mathbb{R}\}$ a fixed Banach space, e.g., Hölder-continuous functions with exponent α. Using a similar loss as in the case of ridge regression, we can define a minimal risk estimator of the function f_0 by

$$\widehat{f} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{F}}^2 \right).$$

This estimator is called the *penalized (non-linear) least-squares estimator* and is equal to the ridge regression estimator when $\mathcal{F}$ is the space of linear functions (with appropriate norm). While we have seen that this estimator is easy to compute for linear functions (ridge regression), this is no longer the case for more general function classes $\mathcal{F}$.

Can this optimization be reduced to the linear case?

Example 2.2. *Consider two classification settings where $X \in \mathbb{R}^2$ and $Y \in \{0, 1\}$.*

(i) Assume $X \sim \mu_0$ for some distribution μ_0 on $\mathbb{R}^2$ and

$$Y = \mathbb{1}(X^1 - X^2 > 0).$$

Data from such a model is shown in Figure 2.1 (left). The decision boundary (solid black line) is linear in this case.

(ii) Assume $X \sim \nu_0$ for some distribution ν_0 on $\mathbb{R}^2$ and

$$Y = \mathbb{1}(X^1 X^2 < 0).$$

Data from such a model is shown in Figure 2.1 (right). In this case it is not possible to separate the two classes ($Y = 0$ and $Y = 1$) with a linear decision boundary. However, if one adds the auxiliary predictor $\tilde{X} := X^1 X^2$, it is possible to separate the points linearly with $\tilde{X} < 0$.

The example suggests to map the observed predictors X into a sufficiently rich feature space such that the regression function f_0 is linear in that space. Formally, we want to find a feature map $\Phi : \mathbb{R}^p \to \mathbb{R}^d$ and use linear methods on $\Phi(X)$ instead of X. In Example 2.2 (ii), we could for example define for all $x \in \mathbb{R}^p$ the feature map

$$\Phi(x) \coloneqq (x^1, x^2, x^1x^2, x^1x^1, x^2x^2).$$

This comes with two added difficulties: (1) The feature space might be high-dimensional and (2) it can be computationally costly (the above example with p predictors leads to $(p^2+3p)/2$ features).

While (1) can be handled by ridge regression, we will use the so-called *kernel-trick* to solve the computational bottleneck (2).

2.2.1 Kernel trick

Assume we have a fixed feature map $\Phi : \mathbb{R}^p \to \mathbb{R}^d$, such that the non-linear regression model in (2.2.1) reduces to

$$Y_i = \Phi(X_i)^\top \beta_0 + \varepsilon_i,$$

with $\beta_0 \in \mathbb{R}^d$. For a fixed-design matrix $X \in \mathbb{R}^{n\times p}$ denote by $\Phi(X) \in \mathbb{R}^{n\times d}$ the design matrix in the feature space. The ridge regression estimator in this case is given by

$$\widehat{\beta}^{\mathrm{R}}_\lambda = (\Phi(X)^\top\Phi(X) + \lambda \mathrm{I}_d)^{-1}\Phi(X)^\top Y. \tag{2.2.2}$$

The computational cost of computing this estimator and similarly also for the predicted values $\widehat{Y} = \Phi(X)\widehat{\beta}^{\mathrm{R}}_\lambda$ is $\mathcal{O}(d^3 + d^2n)$.[1] When $d \gg n$ we can reduce the computational complexity by manipulating the expression for the ridge regression estimator as follows

$$\begin{aligned}
\widehat{\beta}^{\mathrm{R}}_\lambda &= (\Phi(X)^\top\Phi(X) + \lambda \mathrm{I}_d)^{-1}\Phi(X)^\top Y \\
&= (\Phi(X)^\top\Phi(X) + \lambda \mathrm{I}_d)^{-1}\Phi(X)^\top(\Phi(X)\Phi(X)^\top + \lambda \mathrm{I}_n)(\Phi(X)\Phi(X)^\top + \lambda \mathrm{I}_n)^{-1}Y \\
&= (\Phi(X)^\top\Phi(X) + \lambda \mathrm{I}_d)^{-1}(\Phi(X)^\top\Phi(X) + \lambda \mathrm{I}_d)\Phi(X)^\top(\Phi(X)\Phi(X)^\top + \lambda \mathrm{I}_n)^{-1}Y \\
&= \Phi(X)^\top(\Phi(X)\Phi(X)^\top + \lambda \mathrm{I}_n)^{-1}Y.
\end{aligned}$$

In particular, this implies that the predicted values can be expressed as

$$\widehat{Y} = \Phi(X)\widehat{\beta}^{\mathrm{R}}_\lambda = \Phi(X)\Phi(X)^\top(\Phi(X)\Phi(X)^\top + \lambda \mathrm{I}_n)^{-1}Y,$$

which now only depend on the matrix $K \coloneqq \Phi(X)\Phi(X)^\top \in \mathbb{R}^{n\times n}$. In particular, the computational complexity is reduced to $\mathcal{O}(n^3 + n^2d)$ which is substantially smaller if $d \gg n$.

Remark 2.3. *The entries in the matrix K only depend on the feature map Φ via inner-products, i.e., $K_{ij} = \langle \Phi(X_i), \Phi(X_j)\rangle$. Therefore, if these inner-products can be computed efficiently one does not need to compute the feature map $\Phi(X)$ explicitly, which often allows to compute the matrix K more efficiently. For example, for the feature map*

$$\Phi(x) \coloneqq (1, \sqrt{2}x^1, \dots, \sqrt{2}x^p, x^1x^1, \dots x^1x^p, x^2x^1, \dots, x^px^p),$$

it holds that

$$\begin{aligned}
\langle \Phi(X_i), \Phi(X_j)\rangle &= 1 + 2\sum_{k=1}^p X_i^k X_j^k + \sum_{k,\ell=1}^p X_i^k X_i^\ell X_j^k X_j^\ell \\
&= (1 + \sum_{k=1}^p X_i^k X_j^k)^2 \\
&= (1 + X_i^\top X_j)^2.
\end{aligned}$$

Hence, each entry in K can be computed in p operations instead of d, reducing the computational cost to $\mathcal{O}(pn^2)$ instead of $\mathcal{O}(dn^2)$.

[1] A short introduction on how to compute computational costs is given in Appendix A.2.

2.2.2 Kernels

In this section, we introduce kernel functions which will help us generalize the ideas of the previous section. To this end, let $\mathcal{X}$ denote an arbitrary set (this will be the set on which the original predictors live, i.e., $\mathbb{R}^p$ in the section before) and let $\Phi : \mathcal{X} \to \mathcal{H}$ be a feature map, where $\mathcal{H}$ is an inner-product space with inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. As we saw in Section 2.2.1, it is possible to compute the predicted values of the non-linear ridge regression estimator in (2.2.2) by only evaluating feature map via the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined for all $x, x' \in \mathcal{X}$ by

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}. \tag{2.2.3}$$

The function k is an example of a kernel function which is the building block of any kernel method. Formally, a kernel function is defined as follows.

Definition 2.4.
A *positive definite kernel* (or *pd kernel* for short) is a symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in \mathcal{X}$ the matrix $K \in \mathbb{R}^{n \times n}$ defined for all $i, j \in \{1, \dots, n\}$ by

$$K_{ij} = k(x_i, x_j)$$

is positive semi-definite.

It is helpful to think of pd kernels as similarity measures. For example, in (2.2.3) the value $k(x, x')$ quantifies how similar the two features $\Phi(x)$ and $\Phi(x')$ are in $\mathcal{H}$. Pd kernels have some important properties that are summarized in the following proposition.

Proposition 2.5 *(Properties of pd kernels)***.**

(i) Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a pd kernel, then for all $x, x' \in \mathcal{X}$ it holds that

$$k(x, x')^2 \leq k(x, x) k(x', x').$$

(ii) Let $\Phi : \mathcal{X} \to \mathcal{H}$ be a feature map with $\mathcal{H}$ an inner-product space, then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined for all $x, x' \in \mathcal{X}$ by

$$k(x, x') \coloneqq \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

is a pd kernel.

(iii) Let $k_1, k_2, \dots : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be pd kernels, then

- for $\alpha_1, \alpha_2 \geq 0$, it holds that $k \coloneqq \alpha_1 k_1 + \alpha_2 k_2$ is a pd kernel,
- the point-wise limit $k \coloneqq \lim_{n \to \infty} k_n$ (if it exists) is a pd kernel and
- the product $k \coloneqq k_1 k_2$ is a pd kernel.

Proof. We only prove (i) and (ii). The proof of (iii) is part of the exercises.

(i) Fix $x, x' \in \mathcal{X}$, then

$$K \coloneqq \begin{pmatrix} k(x, x) & k(x, x') \\ k(x', x) & k(x', x') \end{pmatrix}$$

is positive semi-definite because k is a kernel. This implies, that $\det(K) \geq 0$ which implies that

$$k(x, x) k(x', x') \geq k(x, x') k(x', x) = k(x, x')^2$$

where we used that k is symmetric since it is a kernel. This proves the result.

(ii) Firstly, the symmetry of k follows from the symmetry of the inner-product. Next, fix $n \in \mathbb{N}$, $x_1, \dots, x_n \in \mathcal{X}$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. Then,

$$\begin{aligned}\alpha^\top K \alpha &= \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \\ &= \sum_{i,j=1}^n \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} \\ &= \Big\langle \sum_{i=1}^n \alpha_i \Phi(x_i), \sum_{j=1}^n \alpha_j \Phi(x_j) \Big\rangle_{\mathcal{H}} \geq 0.\end{aligned}$$

This proves that k is positive definite and hence a pd kernel.

□

In the following example, we introduce some commonly used pd kernels on $\mathbb{R}^p$ and explain how to prove that they are positive definite. Many more kernels exist and often the choice depends on the application at hand.

Example 2.6 (pd kernels on $\mathcal{X} = \mathbb{R}^p$)**.**

- Linear kernel:

$$k : (x, x') \mapsto x^\top x'$$

Proposition 2.5 (ii) immediately implies that this is a pd kernel.

- Polynomial kernel (of order d):

$$k : (x, x') \mapsto (1 + x^\top x')^d$$

To see that this is a pd kernel, use that 1 and $x^\top x'$ are pd kernels which by Proposition 2.5 (iii) implies that $1 + x^\top x'$ is a pd kernel. Now, applying Proposition 2.5 (iii) d times implies that k is a pd kernel.

- Gaussian kernel (with bandwidth $\sigma > 0$):

$$k : (x, x') \mapsto \exp\Big(- \frac{\|x - x'\|_2^2}{2\sigma^2} \Big)$$

This is an important kernel in many applications. To see that it is positive definite, we use $\|x - x'\|_2^2 = \|x\|_2^2 + \|x'\|_2^2 - 2\langle x, x' \rangle$ to get

$$k(x, x') = \underbrace{\sum_{\ell=0}^\infty \tfrac{1}{k!} \sigma^{-2k} \langle x, x' \rangle^k}_{\text{pd by Prop. 2.5 (iii)}} \underbrace{\exp\Big(- \frac{\|x\|_2^2}{2\sigma^2} \Big) \exp\Big(- \frac{\|x'\|_2^2}{2\sigma^2} \Big)}_{\text{pd by Prop. 2.5 (ii)}}$$

Using once more that the product of pd kernels is positive definite, we get that the Gaussian kernel is positive definite.

A key advantage of kernel methods is that they can be defined on arbitrary sets, so they are not restricted to Euclidean space. Three examples are given below.

Example 2.7 (pd kernels on more general spaces)**.**

- Sobolev kernel (on $\mathcal{X} = [0, 1]$):

$$k : (x, x') \mapsto \min(x, x')$$

This kernel corresponds to the covariance function of a Brownian motion, which directly implies that it is pd. In the exercises, you will see how to prove it directly.

- Jaccard similarity (on $\mathcal{X} = \mathcal{P}(\{1, \dots, p\})$):

$$k : (x, x') \mapsto \begin{cases} \frac{|x \cap x'|}{|x \cup x'|} & \text{if } x \cup x' \neq \varnothing \\ 1 & \text{otherwise} \end{cases}$$

A proof that this is a pd kernel is given in the exercises.

- Aitchison kernel (on the simplex, i.e., $\mathcal{X} = \Delta^{p-1} \coloneqq \{x \in [0,1]^p \mid \sum_{j=1}^p x^j = 1\}$):

$$k : (x, x') \mapsto \sum_{j=1}^{p} \log\left(\frac{x^j + c}{\sqrt[p]{\prod_{j=1}^p (x^j + c)}}\right) \log\left(\frac{x'^j + c}{\sqrt[p]{\prod_{j=1}^p (x'^j + c)}}\right),$$

for $c > 0$. This kernel can be used to analyze compositional data and has connections to the log-contrast model [Huang et al., 2022]. To see that it is positive definite, use the feature map $\Phi : \Delta^{p-1} \to \mathbb{R}^p$ defined for all $x \in \Delta^{p-1}$ by

$$\Phi(x) = \left(\log\left(\frac{x^1 + c}{\sqrt[p]{\prod_{j=1}^p (x^j + c)}}\right), \dots, \log\left(\frac{x^p + c}{\sqrt[p]{\prod_{j=1}^p (x^j + c)}}\right)\right)$$

and apply Proposition 2.5 (ii).

- Gaussian kernel (on a general Banach space $\mathcal{X}$):

$$k : (x, x') \mapsto \exp\left(-\frac{\|x - x'\|_{\mathcal{X}}^2}{2\sigma^2}\right)$$

The only difference to the Gaussian kernel on $\mathbb{R}^p$ is that now we use the norm $\|\cdot\|_{\mathcal{X}}$ corresponding to the Banach space. This allows applying the kernel methodology in settings where X is a random function in for example $L^2([0,1])$. A practical example of this is given in chemistry when measuring chemical compounds with a mass spectrometer. In that case each mass spectrometry profile can be modeled as a random function.

2.2.3 Reproducing kernel Hilbert spaces

In this section, we introduce a class of feature spaces that has many desirable properties, in particular computationally. Formally, we make the following definition.

Definition 2.8 *(Reproducing kernel Hilbert space).*
Let $\mathcal{H} \subseteq \{f \mid f : \mathcal{X} \to \mathbb{R}\}$ be a Hilbert space. Then, $\mathcal{H}$ is called a *reproducing kernel Hilbert space (RKHS)* if there exists a pd kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

(i) $\forall x \in \mathcal{X} : k(x, \cdot) \in \mathcal{H}$,

(ii) $\forall f \in \mathcal{H}, x \in \mathcal{X} : \langle f, k(x, \cdot)\rangle_{\mathcal{H}} = f(x)$.

A pd kernel k satisfying these properties is called a *reproducing kernel* for the RKHS $\mathcal{H}$. Property (ii) is called the *reproducing property.*

Given an RKHS $\mathcal{H}$ on $\mathcal{X}$ with reproducing kernel k, we can construct a feature map $\Phi : \mathcal{X} \to \mathcal{H}$ defined for all $x \in \mathcal{X}$ by

$$\Phi(x) \coloneqq k(x, \cdot).$$

By property (i) in the definition of an RKHS, Φ indeed maps into $\mathcal{H}$. Moreover, by the reproducing property (property (ii)) we also get that (2.2.3) is true.

Example 2.9 (Space of linear functions on $\mathbb{R}^p$ is an RKHS)**.** *Consider the space of linear functions on $\mathbb{R}^p$ defined by*

$$\mathcal{H} \coloneqq \{f : \mathbb{R}^p \to \mathbb{R} \mid \exists \beta \in \mathbb{R}^p \text{ such that } \forall x \in \mathbb{R}^p : \ f(x) = \beta^\top x\}.$$

Moreover, define the inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ for all $f, g \in \mathcal{H}$ by

$$\langle f, g \rangle_{\mathcal{H}} = \beta^\top \alpha,$$

where $\beta, \alpha \in \mathbb{R}^p$ such that $f(\cdot) = \beta^\top(\cdot)$ and $g(\cdot) = \alpha^\top(\cdot)$, respectively.[2] *The space $\mathcal{H}$ together with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ forms a Hilbert space. To show that it is an RKHS, we now need to find a pd kernel $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ satisfying properties (i) and (ii) in Definition 2.8. By property (i), we know that $k(x, \cdot)$ is a linear function for all $x \in \mathbb{R}^p$. Let $\beta_x \in \mathbb{R}^p$ be the representation of this function, i.e., $k(x, \cdot) = \beta_x^\top(\cdot)$. Then, by property (ii), we get for all $\alpha \in \mathbb{R}^p$ and $x \in \mathbb{R}^p$ that*

$$\alpha^\top x = \langle \alpha^\top(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} = \alpha^\top \beta_x.$$

Therefore, it holds that $\beta_x = x$. Hence, k has to have the form $k : (x, x') \mapsto x^\top x'$. This is however just the linear kernel and therefore pd. Hence, we have shown that $\mathcal{H}$ together with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is an RKHS with reproducing kernel equal to the linear kernel.

As we show in the following theorem, every fixed pd kernel k corresponds to an RKHS. This guarantees that if one starts with a pd kernel k, then there is a well-behaved feature space $\mathcal{H} \subseteq \{f \mid f : \mathcal{X} \to \mathbb{R}\}$ and a feature map $\Phi : \mathcal{X} \to \mathcal{H}$ such that k satisfies (2.2.3). The following theorem makes this statement precise.

Theorem 2.10 *(Every pd kernel induces an inner-product space and a feature map)***.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a pd kernel. Then there exists an inner-product space $\mathcal{H}_k$ and a feature map $\Phi_k : \mathcal{X} \to \mathcal{H}_k$ such that for all $x, x' \in \mathcal{X}$ it holds that

$$k(x, x') = \big\langle \Phi_k(x), \Phi_k(x') \big\rangle_{\mathcal{H}_k}.$$

Proof. We prove the result by explicitly constructing the inner-product space as

$$\mathcal{H}_k \coloneqq \{f : \mathcal{X} \to \mathbb{R} \mid \exists n \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_n \in \mathcal{X} : \ f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)\}$$

and the feature map $\Phi_k : \mathcal{X} \to \mathcal{H}$ for all $x \in \mathcal{X}$ by

$$\Phi_k(x) \coloneqq k(x, \cdot).$$

It now remains to be shown that there exists an inner-product on $\mathcal{H}_k$ which satisfies the desired properties. To construct such an inner-product $\langle \cdot, \cdot \rangle : \mathcal{H}_k \times \mathcal{H}_k \to \mathbb{R}$, fix two arbitrary functions $f, g \in \mathcal{H}_k$ with the expansions

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \quad \text{and} \quad g(\cdot) = \sum_{i=1}^{m} \beta_i k(x'_i, \cdot) \tag{2.2.4}$$

and define

$$\langle f, g \rangle \coloneqq \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j).$$

This is well-defined (i.e., does not depend on the explicit expansion of f or g) since

$$\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(x'_j).$$

[2] This is a well-defined inner-product, because the coefficients β and α are unique.

Furthermore, $\langle \cdot, \cdot \rangle$ satisfies for all $x, x' \in \mathcal{X}$ that

$$\langle \Phi_k(x), \Phi_k(x') \rangle = \langle k(x, \cdot), k(x', \cdot) \rangle = k(x, x')$$

as desired. Next, we show that the function $\langle \cdot, \cdot \rangle$ is indeed an inner-product, i.e., it is symmetric, linear and positive definite, by explicitly checking the conditions.

- *Symmetry:*
$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) = \langle g, f \rangle$$

- *Linearity:* For $a \in \mathbb{R}$ and $f, g, h \in \mathcal{H}_k$ with f and g as in (2.2.4) it holds that
$$\begin{aligned} \langle ag + h, f \rangle &= \sum_{i=1}^{n} \alpha_i (a g(x_i) + h(x_i)) \\ &= a \sum_{i=1}^{n} \alpha_i g(x_i) + \sum_{i=1}^{n} \alpha_i h(x_i) \\ &= a \langle g, f \rangle + \langle h, f \rangle. \end{aligned}$$

- *Positive-definiteness:*

  (1) First we show that $\langle f, f \rangle \geq 0$ for all $f \in \mathcal{H}_k$. Fix $f \in \mathcal{H}_k$ with the expansion as in (2.2.4), then
  $$\langle f, f \rangle = \sum_{i,j=1}^{n} \alpha_i k(x_i, x_j) \alpha_j \geq 0,$$
  where we used that k is a pd kernel.

  (2) Next, we show that $\langle f, f \rangle = 0 \Rightarrow f \equiv 0$. To see this, we first observe that $\langle \cdot, \cdot \rangle$ is itself a pd kernel on $\mathcal{H}_k$ since for all $\gamma_1, \ldots, \gamma_m \in \mathbb{R}$ and $f_1, \ldots, f_m \in \mathcal{H}_k$ it holds that
  $$\sum_{i,j=1}^{m} \gamma_i \langle f_i, f_j \rangle \gamma_j = \Big\langle \sum_{i=1}^{m} \gamma_i f_i, \sum_{i=1}^{m} \gamma_i f_i \Big\rangle \geq 0.$$
  Hence, for a fixed $f \in \mathcal{H}_k$ with expansion given in (2.2.4), we can apply Proposition 2.5 (i) to get that
  $$(\langle k(x, \cdot), f \rangle)^2 \leq \langle k(x, \cdot), k(x, \cdot) \rangle \langle f, f \rangle = k(x, x) \langle f, f \rangle. \tag{2.2.5}$$
  Moreover, it holds that
  $$\langle k(x, \cdot), f \rangle = \sum_{i=1}^{n} \alpha_i k(x, x_i) = f(x) \tag{2.2.6}$$
  Finally, combining (2.2.5) and (2.2.6) implies
  $$f(x)^2 \leq k(x, x) \langle f, f \rangle,$$
  and hence $\langle f, f \rangle = 0$ also implies $f \equiv 0$.

This completes the proof of Theorem 2.10. □

Remark 2.11. *The inner-product space in Theorem 2.10 can be turned into a Hilbert space (and hence an RKHS) by including the limits of all Cauchy sequences. This requires verifying (left as an exercise) that the inner-product is well-defined and satisfies the desired properties on the extended space.*

Starting from a pd kernel, the construction in the proof of Theorem 2.10 can be used to derive an explicit representation of the corresponding RKHS. This is done for the linear and Sobolev kernel in the following example.

Example 2.12 (RKHSs with explicit representations)**.**

- Linear kernel with $\mathcal{X} = \mathbb{R}^p$: *We now consider the reverse direction of Example 2.9 and start from the linear kernel* $k : (x, x') \mapsto x^\top x'$*. Using the representation from the proof of Theorem 2.10 and taking the closure leads to the RKHS*

$$\begin{aligned}\mathcal{H}_k &= \overline{\{f : \mathbb{R}^p \to \mathbb{R} \mid \exists \alpha \in \mathbb{R}^n, x_1, \dots, x_n \in \mathbb{R}^p \text{ s.t. } f(\cdot) = \textstyle\sum_{i=1}^n \alpha_i k(x_i, \cdot)\}} \\ &= \overline{\{f : \mathbb{R}^p \to \mathbb{R} \mid \exists \alpha \in \mathbb{R}^n, x_1, \dots, x_n \in \mathbb{R}^p \text{ s.t. } f(\cdot) = (\textstyle\sum_{i=1}^n \alpha_i x_i)^\top (\cdot)\}} \\ &= \overline{\{f : \mathbb{R}^p \to \mathbb{R} \mid \exists \beta \in \mathbb{R}^p \text{ s.t. } f(\cdot) = \beta^\top (\cdot)\}} \\ &= \{f : \mathbb{R}^p \to \mathbb{R} \mid \exists \beta \in \mathbb{R}^p \text{ s.t. } f(\cdot) = \beta^\top (\cdot)\}\end{aligned}$$

  with inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ *given for all* $f, g \in \mathcal{H}_k$ *with* $f(\cdot) = \beta^\top(\cdot)$ *and* $g(\cdot) = \gamma^\top(\cdot)$ *by*

$$\langle f, g \rangle_{\mathcal{H}_k} = \beta^\top \gamma.$$

  The RKHS for the linear kernel is therefore isomorphic with $\mathbb{R}^p$*.*

- Sobolev kernel with $\mathcal{X} = [0, 1]$: *Recall, that the Sobolev kernel is given by* $k : (x, x') \mapsto \min(x, x')$*. The RKHS resulting from Theorem 2.10 can be expressed as*

$$\mathcal{H}_k = \{f : [0, 1] \to \mathbb{R} \mid f \text{ a.e. differentiable}, f(0) = 0 \text{ and } \textstyle\int_0^1 f'(x)^2 dx < \infty\}$$

  with inner-product

$$\langle f, g \rangle_{\mathcal{H}_k} = \int_0^1 f'(x) g'(x) dx.$$

  A proof of this result can be found in Wainwright [2019, Example 12.6].

Depending on the kernel it can be difficult to get an easy representation of the corresponding RKHS beyond the one given in the proof of Theorem 2.10.

RKHSs are useful because they are easy to optimize over. In particular, it can be shown that optimization over a (potentially infinite dimensional) RKHS can be expressed as a finite dimensional optimization. This result is known as the representer theorem.

Theorem 2.13 *(Representer theorem)***.**
Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a pd kernel and $\mathcal{H}$ the corresponding RKHS. Furthermore, let

- $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be a loss function,
- $J : [0, \infty) \to \mathbb{R}$ strictly increasing and
- $Y \in \mathbb{R}^n, X = (X_1, \dots, X_n) \in \mathcal{X}^n$ and $f \in \mathcal{H}$ and $K = (k(X_i, X_j))_{i,j}$ the kernel matrix.[a]

Then,

$$\widehat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \underbrace{c(Y, X, f(X)) + J(\|f\|_{\mathcal{H}}^2)}_{=: Q_1(f)}$$

if and only if $\widehat{f} = \sum_{i=1}^n \widehat{\alpha}_i k(X_i, \cdot)$ with

$$\widehat{\alpha} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \underbrace{c(Y, X, K\alpha) + J(\alpha^\top K \alpha)}_{=: Q_2(\alpha)}.$$

[a]For $x \in \mathcal{X}^m$ and $f : \mathcal{X} \to \mathbb{R}$ we use the slight abuse of notation $f(x) = (f(x_1), \dots, f(x_m))$.

Proof. We begin with the "only if" direction. Assume $\widehat{f} \in \mathcal{H}$ minimizes Q_1. Then, the decomposition

$$\widehat{f} = u + v \tag{2.2.7}$$

with $v \in V \coloneqq \operatorname{span}(\{k(X_1, \cdot), \dots, k(X_n, \cdot)\})$ and $u \in V^\perp$ exists. This implies for all $i \in \{1, \dots, n\}$ that

$$\widehat{f}(X_i) = \langle k(x_i, \cdot), u + v \rangle_{\mathcal{H}} = \langle k(X_i, \cdot), v \rangle_{\mathcal{H}} = v(X_i). \tag{2.2.8}$$

Moreover, it holds that

$$\|\widehat{f}\|_{\mathcal{H}}^2 = \langle \widehat{f}, \widehat{f} \rangle_{\mathcal{H}} = \langle u, u \rangle_{\mathcal{H}} + \langle v, v \rangle_{\mathcal{H}} = \|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2,$$

which since J is increasing implies that

$$J(\|\widehat{f}\|_{\mathcal{H}}^2) = J(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2) \geq J(\|v\|_{\mathcal{H}}^2). \tag{2.2.9}$$

Now, since (2.2.8) implies that Q_1 only depends on u via the penalty term $J(\|f\|_{\mathcal{H}}^2)$, optimality of $\widehat{f}$ together with (2.2.9) imply that $J(\|\widehat{f}\|_{\mathcal{H}}^2) = J(\|v\|_{\mathcal{H}}^2)$. Since J is strictly increasing this however also implies $u \equiv 0$. Hence, $\widehat{f} = v$ and there exist $\alpha_1, \dots, \alpha_n$ such that

$$\widehat{f}(\cdot) = \sum_{i=1}^n \alpha_i k(X_i, \cdot).$$

This implies that $\|\widehat{f}\|_{\mathcal{H}}^2 = \alpha^\top K \alpha$ and hence $Q_1(\widehat{f}) = Q_2(\alpha)$. Therefore, we have proved the "only if" direction with $\alpha = \widehat{\alpha}$.

For the "if" direction, assume $\widehat{\alpha} \in \mathbb{R}^n$ minimizes Q_2 and define $\widehat{f}(\cdot) \coloneqq \sum_{i=1}^n \widehat{\alpha}_i k(X_i, \cdot)$. Then, it holds that $Q_1(\widehat{f}) = Q_2(\widehat{\alpha})$. Next, fix $\tilde{f} \in \mathcal{H}$ with

$$Q_1(\tilde{f}) \leq Q_1(\widehat{f}). \tag{2.2.10}$$

By the same argument as in the "only if" direction, we get for $v \in V$ and $u \in V^\perp$ with $\tilde{f} = u + v$ that

$$Q_1(v) \leq Q_1(\tilde{f}). \tag{2.2.11}$$

Moreover, since $v \in V$ there exists $\alpha \in \mathbb{R}^n$ such that $v(x) = K\alpha$ and $\|v\|_{\mathcal{H}}^2 = \alpha^\top K \alpha$. Hence, $Q_1(v) = Q_2(\alpha)$ and by the optimality of $\widehat{\alpha}$ we also have that

$$Q_1(\widehat{f}) \leq Q_1(v). \tag{2.2.12}$$

Together (2.2.10), (2.2.11) and (2.2.12) imply that $Q_1(\widehat{f}) = Q_1(\tilde{f})$ which implies that $\widehat{f}$ minimizes Q_1. This completes the proof of Theorem 2.13. □

2.3 Kernel ridge regression

We now apply the theory on kernels and RKHSs to construct a non-linear regression procedure. Consider again the fixed-design non-linear regression problem

$$Y = f_0(X) + \varepsilon,$$

with $\varepsilon \sim \mu_0$ mean zero and $f_0 \in \mathcal{F}$ with $\mathcal{F} \subseteq \{f \mid f : \mathcal{X} \to \mathbb{R}\}$ a fixed function class. As we saw in Section 2.2, we can construct an estimator for f_0 using the following approach: First, assume there exists a known feature map $\Phi : \mathcal{X} \to \mathbb{R}^d$ such that

$$f_0(\cdot) = \beta_0^\top \Phi(\cdot).$$

Second, apply standard ridge regression as shown in Figure 2.2 (top). In Section 2.2.1, we saw that this approach can be sped up computationally using the kernel trick. This makes the approach feasible even if the dimension of the feature space $\mathbb{R}^d$ is large. Nevertheless, two practical issues of this approach remain: It can only be applied for finite-dimensional feature spaces and it requires to explicitly construct the feature map Φ.

The theory we developed in Sections 2.2.2 and 2.2.3 suggests a second approach: Assume we are given a kernel k such that the corresponding RKHS (given by Theorem 2.10) is equal to the function class $\mathcal{F}$. Then, we can apply the representer theorem to efficiently estimate the corresponding penalized regression estimator, see Figure 2.2 (bottom). This approach is known as *kernel ridge-regression*. The two approaches are mathematically equivalent whenever the RKHS is finite dimensional and the feature map and kernel satisfy for all $x, x' \in \mathcal{X}$ that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle.$$

However, kernel ridge-regression can also be applied with infinite dimensional feature spaces and only requires specification of the kernel.

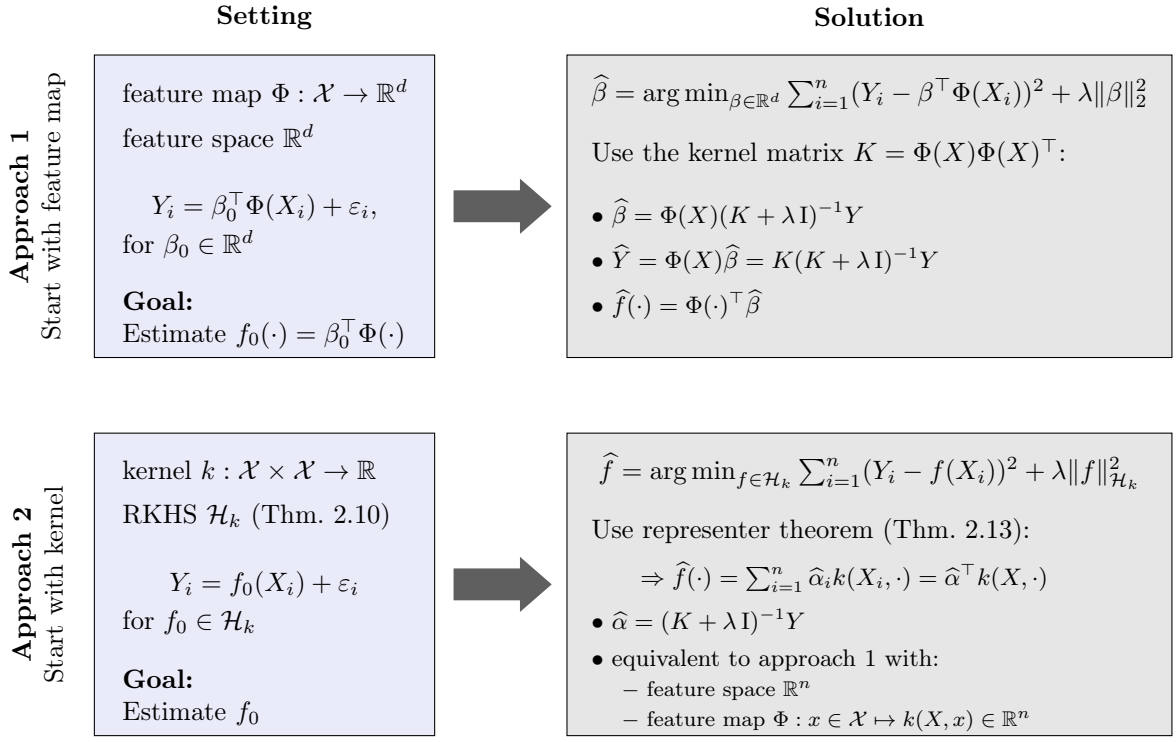

Figure 2.2: Overview of non-linear regression with kernels. Two approaches are possible: Either starting with a feature map (top) or starting with a kernel (bottom). Approach 2 is called kernel ridge regression and is mathematically equivalent to approach 1 whenever the RKHS is finite dimensional. Moreover, in practice, it is often easier to specify a pd kernel instead of a feature map, making the second approach more versatile.

2.3.1 Theoretical properties of kernel ridge regression

Assume we observe data $(Y_1, X_1), \ldots, (Y_n, X_n)$ from

$$Y_i = f_0(X_i) + \varepsilon_i.$$

Throughout this section, we use the following model assumptions and notational conventions:

(i) Fixed-design, that is, $X_1, \dots, X_n \in \mathcal{X}$ are deterministic.

(ii) $\varepsilon_1, \dots, \varepsilon_n$ are real-valued i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_0^2 > 0$.

(iii) $f_0 \in \mathcal{H}$, where $\mathcal{H}$ is RKHS with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

(iv) $\|f_0\|_{\mathcal{H}} \leq 1$.[3]

(v) $K \coloneqq (k(X_i, X_\ell))_{i,\ell}$ kernel matrix with

$$K = UDU^\top$$

where $d_j \coloneqq D_j^j$ with $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$ and U is orthogonal.

Recall that the kernel ridge regression estimator with penalty $\lambda > 0$ is defined by

$$\widehat{f}_\lambda \coloneqq \operatorname*{arg\,min}_{f \in \mathcal{H}} \left(\sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

Using the representer theorem (Theorem 2.13) the estimator has the following closed form representation

$$\widehat{f}_\lambda = \sum_{i=1}^n \widehat{\alpha}_i^\lambda k(X_i, \cdot) \quad \text{with} \quad \widehat{\alpha}^\lambda = (K + \lambda \mathrm{I})^{-1} Y.$$

We now want to assess theoretically, whether this is a good estimator. The following theorem provides an upper bound on the mean squared prediction error.

Theorem 2.14 *(Upper bound on MSPE).*
Given the setting described at the beginning of Section 2.3.1, the mean squared prediction error (MSPE) can be bounded from above by

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (f_0(X_i) - \widehat{f}_\lambda(X_i))^2 \right] \leq \frac{\sigma_0^2}{n} \sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda}{4n} \leq \underbrace{\frac{\sigma_0^2}{n} \frac{1}{\lambda} \sum_{i=1}^n \min\left(\frac{d_i}{4}, \lambda \right)}_{\sim \text{ variance}} + \underbrace{\frac{\lambda}{4n}}_{\sim \text{ bias}} =: \delta_n(\lambda).$$

The separation of the upper bound into a bias and a variance term follows from the decomposition in the proof.

Proof. We make use of two statements that follow from the closed form solution for the ridge regression estimator and properties of the RKHS $\mathcal{H}$. Firstly,

$$(\widehat{f}_\lambda(X_1), \dots, \widehat{f}_\lambda(X_n))^\top = K(K + \lambda \mathrm{I})^{-1} Y \tag{2.3.1}$$

and secondly, there exists $\alpha \in \mathbb{R}^n$ such that

$$(f_0(X_1), \dots, f_0(X_n))^\top = K\alpha \quad \text{and} \quad \alpha^\top K \alpha \leq \|f_0\|_{\mathcal{H}}^2 \leq 1. \tag{2.3.2}$$

Both (2.3.1) and (2.3.2) are proved in the exercises.

Next, define $\theta \coloneqq U^\top K \alpha = DU^\top \alpha$, which also implies $U\theta = K\alpha$, and expand the MSPE using (2.3.1) and (2.3.2) as follows

$$\begin{aligned}
\mathbb{E}\big[\|\widehat{f}_\lambda(X) - f_0(X)\|_2^2\big] &= \mathbb{E}\big[\|K(K + \lambda \mathrm{I})^{-1}(U\theta + \varepsilon) - U\theta\|_2^2\big] \\
&= \mathbb{E}\big[\|UDU^\top (UDU^\top + \lambda UU^\top)^{-1}(U\theta + \varepsilon) - U\theta\|_2^2\big] \\
&= \mathbb{E}\big[\|U(D(D + \lambda \mathrm{I})^{-1} U^\top (U\theta + \varepsilon) - \theta)\|_2^2\big] \\
&= \mathbb{E}\big[\|D(D + \lambda \mathrm{I})^{-1}(\theta + U^\top \varepsilon) - \theta\|_2^2\big] \\
&= \underbrace{\|(D(D + \lambda \mathrm{I})^{-1} - \mathrm{I})\theta\|_2^2}_{\text{term 1}} + \underbrace{\mathbb{E}\big[\|D(D + \lambda \mathrm{I})^{-1} U^\top \varepsilon\|_2^2\big]}_{\text{term 2}}.
\end{aligned}$$

[3] This can always be achieved by rescaling the kernel k.

We now consider the two terms separately.
(term 1): Expanding term 1 leads to

$$\|(D(D+\lambda \mathrm{I})^{-1}-\mathrm{I})\theta\|_2^2 = \sum_{i=1}^n \left(\frac{d_i}{d_i+\lambda}-1\right)^2 \theta_i^2 = \sum_{i=1}^n \frac{\lambda^2}{(d_i+\lambda)^2}\theta_i^2. \tag{2.3.3}$$

Next, define D^+ to be the diagonal matrix with d_i^{-1} if $d_i > 0$ and zero otherwise. Then, it holds that

$$\sum_{i:d_i>0} \frac{\theta_i^2}{d_i} = \|\sqrt{D^+}\theta\|_2^2 = \alpha^\top U D D^+ D U^\top \alpha = \alpha^\top K \alpha \leq 1,$$

where we used that $\theta = DU^\top\alpha$. Hence, the expression (2.3.3) can be simplified to

$$\begin{aligned}
\sum_{i=1}^n \frac{\lambda^2}{(d_i+\lambda)^2}\theta_i^2 &= \sum_{i:d_i>0} \frac{\theta_i^2}{d_i}\frac{d_i\lambda^2}{(d_i+\lambda)^2} \\
&\leq \left(\sum_{i:d_i>0}\frac{\theta_i^2}{d_i}\right)\left(\max_{i:d_i>0}\frac{d_i\lambda^2}{(d_i+\lambda)^2}\right) \\
&\leq 1\left(\max_{i\in\{1,\dots,n\}}\frac{d_i\lambda^2}{(d_i+\lambda)^2}\right) \\
&\leq \frac{\lambda}{4},
\end{aligned}$$

where for the first inequality we used Hölder's inequality and for the last inequality we used $(a+b)^2 \geq 4ab$.
(term 2): Term 2 can be bounded as follows

$$\begin{aligned}
\mathbb{E}\big[\|D(D+\lambda \mathrm{I})^{-1}U^\top\varepsilon\|_2^2\big] &= \mathbb{E}\big[\{D(D+\lambda \mathrm{I})^{-1}U^\top\varepsilon\}^\top\{D(D+\lambda \mathrm{I})^{-1}U^\top\varepsilon\}\big] \\
&= \mathbb{E}\big[\operatorname{trace}\big(D(D+\lambda \mathrm{I})^{-1}U^\top\varepsilon\varepsilon^\top U D(D+\lambda \mathrm{I})^{-1}\big)\big] \\
&= \sigma_0^2 \operatorname{trace}\big(D^2(D+\lambda \mathrm{I})^{-2}\big) \\
&= \sigma_0^2 \sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2} \\
&\leq \sigma_0^2 \sum_{i=1}^n \min\left(1, \frac{d_i}{4\lambda}\right) \\
&= \frac{\sigma_0^2}{\lambda}\sum_{i=1}^n \min\left(\lambda, \frac{d_i}{4}\right),
\end{aligned}$$

where we again used $(a+b)^2 \geq 4ab$ for the inequality.

Combining the bounds for term 1 and term 2, completes the proof of Theorem 2.14. □

Theorem 2.14 provides an upper bound on how large the MSPE can be. To better understand the bound, we can express it in terms of a shrinking penalty $\lambda_n = \frac{\lambda}{n}$ and $\widehat{\mu}_i \coloneqq \frac{d_i}{n}$ (which corresponds to the eigenvalues of the scaled kernel matrix $\frac{1}{n}K$) as follows

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (f_0(X_i) - \widehat{f}_\lambda(X_i))^2\right] \leq \frac{\sigma_0^2}{\lambda_n}\frac{1}{n}\sum_{i=1}^n \min\left(\frac{\widehat{\mu}_i}{4}, \lambda_n\right) + \frac{\lambda_n}{4} =: \delta_n(\lambda_n).$$

Using this parametrization, we can explicitly bound the term $\delta_n(\lambda_n)$ for different kernels.

Linear kernel

For the linear kernel it holds that $\frac{1}{n}K = \frac{1}{n}XX^\top$. As we saw the RKHS in this case corresponds to space of linear functions and hence the model given at the beginning of Section 2.3.1 is a linear regression model. Assume that we are in the low-dimensional and invertible case (i.e., $\text{rank}(XX^\top) = p$). Then, it holds that

$$\widehat{\mu}_{p+1}, \ldots, \widehat{\mu}_n = 0.$$

This in turn implies that

$$\sum_{i=1}^{n} \min\left(\frac{\widehat{\mu}_i}{4}, \lambda_n\right) \leq p\lambda_n,$$

which further implies that the bound satisfies

$$\delta_n(\lambda_n) \leq \sigma_0^2 \frac{p}{n} + \frac{\lambda_n}{4}. \tag{2.3.4}$$

It is easy to check that simply applying OLS leads to a similar rate of $\mathcal{O}(\frac{p}{n})$. As we will discuss in Chapter 3 this is in a certain sense the optimal achievable rate in the non-sparse setting.

Sobolev kernel

Bounding δ_n becomes more complicated for kernels corresponding to infinite dimensional RKHSs. As an example we consider the Sobolev kernel. For the analysis we use a random-design version of the setting at the beginning of Section 2.3.1, where $X_1, \ldots, X_n \overset{iid}{\sim} \nu$ for some distribution ν. Since the empirical eigenvalues $\widehat{\mu}_i$ are now random quantities, we will first connect them to population quantities that only depend on the kernel. To this end, we use Mercer's theorem, which provides a series expansion of every kernel and connects it to a bounded linear operator.

Theorem 2.15 *(Mercer's theorem).*
Let $(\mathcal{X}, \nu)$ be a compact measured space. Then, for any positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $\int_{\mathcal{X}\times\mathcal{X}} k(x,y)^2 \nu(dx)\nu(dy) < \infty$, there exits $(\mu_j)_{j\in\mathbb{N}} \subseteq \mathbb{R}$ with $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ and an orthonormal basis $(e_j)_{j\in\mathbb{N}} \subseteq L^2(\mathcal{X}, \nu)$ satisfying for ν-almost all $x, x' \in \mathcal{X}$ that

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x'),$$

and that the infinite series converges absolutely and uniformly. Furthermore, the linear bounded operator $\mathcal{K} : L^2(\mathcal{X}, \nu) \to L^2(\mathcal{X}, \nu)$ defined for all $f \in L^2(\mathcal{X}, \nu)$ by

$$\mathcal{K}(f) = \int_{\mathcal{X}} k(x, \cdot) f(x) \nu(dx)$$

has eigenvalues $(\mu_j)_{j\in\mathbb{N}}$ and eigenvectors $(e_j)_{j\in\mathbb{N}}$.

A proof of this result can be found in Wainwright [2019, Theorem 12.20]. It furthermore holds that the empirical eigenvalues $\widehat{\mu}_j$ approach the eigenvalues of the population kernel operator μ_j as n tends to infinity. Moreover it is possible to prove [see e.g., Lundborg et al., 2022, Lemma 18, Supporting information] that for all $\lambda > 0$ it holds that

$$\mathbb{E}\left[\sum_{i=1}^{n} \min\left(\frac{\widehat{\mu}_i}{4}, \lambda\right)\right] \leq \sum_{i=1}^{\infty} \min\left(\frac{\mu_i}{4}, \lambda\right). \tag{2.3.5}$$

Finally, in order to bound δ_n for the Sobolev kernel, we need to explicitly compute the eigenvalues of the corresponding kernel operator. For the Sobolev kernel together with the measure $\nu = \text{Unif}(0,1)$ this computation is given in Wainwright [2019, Example 12.23] and results in

$$\mu_j = \frac{4}{\pi^2(2j-1)^2} \quad \text{and} \quad e_j : x \mapsto \sin\left(\frac{x}{\sqrt{\mu_j}}\right).$$

Combining all these results, we end up with the following corollary of Theorem 2.15.

Corollary 2.16 *(Upper bound for MSPE with Sobolev kernel).*
Assume a random-design version of the setting at the beginning of Section 2.3.1, where $X_1, \dots, X_n \overset{iid}{\sim} \text{Unif}(0,1)$ and independent of $\varepsilon_1 \dots, \varepsilon_n$. Then, it holds that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f_0(X_i) - \widehat{f}_\lambda(X_i))^2\right] = \mathcal{O}\left(\frac{\sigma_0^2}{n\sqrt{\lambda_n}} + \lambda_n\right).$$

In particular, the optimal choice $\lambda_n \sim \left(\frac{\sigma_0^2}{n}\right)^{2/3}$ leads to

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f_0(X_i) - \widehat{f}_\lambda(X_i))^2\right] = \mathcal{O}\left(\left(\frac{\sigma_0^2}{n}\right)^{2/3}\right).$$

Proof. First, we plug-in the expression for the eigenvalues μ_j and then use an integral bound as follows,

$$\sum_{j=1}^{\infty}\min\left(\frac{\mu_j}{4}, \lambda_n\right) = \sum_{j=1}^{\infty}\min\left(\frac{1}{\pi^2(2j-1)^2}, \lambda_n\right) \leq \int_0^{\infty}\min\left(\frac{1}{\pi^2(2x-1)^2}, \lambda_n\right)dx.$$

Next, defining $c^* \coloneqq \frac{1}{2}\left(\frac{1}{\sqrt{\pi^2\lambda_n}} + 1\right)$ we can further bound the integral to get

$$\int_0^{\infty}\min\left(\frac{1}{\pi^2(2x-1)^2}, \lambda_n\right)dx \leq \lambda_n c^* + \frac{1}{\pi^2}\int_{c^*}^{\infty}\frac{1}{(2x-1)^2}dx = \frac{\sqrt{\lambda_n}}{\pi} + \frac{\lambda_n}{2} = \mathcal{O}(\sqrt{\lambda_n}),$$

as $n \to \infty$. Finally, combining this bound with (2.3.5) and the definition of $\delta(\lambda_n)$ leads to

$$\mathbb{E}[\delta(\lambda_n)] = \mathcal{O}\left(\frac{\sigma_0^2}{n\sqrt{\lambda_n}} + \lambda_n\right),$$

which together with Theorem 2.14 proves the first part of Corollary 2.16. The optimal choice λ_n is then found by minimizing the this bound with respect to λ_n.

□

Gaussian kernel

For the Gaussian kernel on $[-1,1]$ and $\nu = \text{Unif}(-1,1)$, the eigenvalues of the kernel operator can be shown to satisfy

$$\mu_j \asymp \exp(-cj\log(j)),$$

for some constant $c > 0$ and where $\asymp$ denotes that the two quantities are asymptotically equivalent up to constants, i.e., $\lim_{j\to\infty}\frac{\mu_j}{\exp(-cj\log(j))}$ and $\lim_{j\to\infty}\frac{\exp(-cj\log(j))}{\mu_j}$ exist and are finite. A similar argument as in the proof of Corollary 2.16 leads to the bound

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f_0(X_i) - \widehat{f}_\lambda(X_i))^2\right] = \mathcal{O}\left(\sigma_0^2\frac{\log(cn)}{n}\right),$$

as n goes to infinity [Wainwright, 2019, Example 13.21]. While the convergence rate is substantially faster than for the Sobolev kernel, the eigenvalues μ_j also decay a lot faster indicating that the RKHS for the Gaussian kernel is less rich. A formal result connecting the eigenvalues of a pd kernel with its RKHS can be found in Wainwright [2019, Corollary 12.26].

2.3.2 Lower bound

An obvious follow up question to Theorem 2.14 is whether it is possible to achieve better rates. The following result answers this question with "no" by providing a lower bound, that achieves (up to constants) the same asymptotic rate.

Theorem 2.17 *(Lower bound on MSPE [Yang et al., 2017])*.
Assume the setting described at the beginning of Section 2.3.1. Then, it holds that

$$\inf_{\widehat{f}\in\mathcal{H}} \sup_{f_0\in\mathcal{H}:\|f_0\|_{\mathcal{H}}\leq 1} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f_0(X_i)-\widehat{f}(X_i))^2\right] \geq c \inf_{\lambda>0} \delta_n(\lambda).$$

2.4 Outlook: Non-parametric hypothesis testing

So far we have only considered kernel methods for estimating non-linear regression functions. The theory on pd kernels and RKHSs is however much more general and can be applied to various different tasks. In this section, we shortly present one additional application to which kernel methods have been successfully applied. Specifically, we consider the problem of non-parametric hypothesis testing. Two highly used kernel-based tests are the *Hilbert-Schmidt Independence Criterion* (HSIC) test for independence testing [Gretton et al., 2005] and the *Maximum Mean Discrepancy* (MMD) test for two-sample testing [Gretton et al., 2012]. Both tests are based on a method for embedding distributions into RKHSs, called *kernel mean embedding*.

Let P be a distribution on $\mathcal{X}$ and let $\mathcal{H}$ be an RKHS with reproducing kernel k. Then, the kernel mean embedding of P into $\mathcal{H}$, is defined by

$$\Pi(P) \coloneqq \int_{\mathcal{X}} k(x,\cdot)P(dx) = \mathbb{E}_{X\sim P}[k(X,\cdot)],$$

where the integral/expectation is a Bochner integral.[4] It can be shown that $\Pi(P)$ is an element in the RKHS $\mathcal{H}$. Using the kernel mean embedding thus allows to compare different distributions using the RKHS norm $\|\cdot\|_{\mathcal{H}}$. We now discuss how this embedding can be applied to two-sample and independence testing, respectively.

- *Two-sample testing:* Assume we are given $(X_1,\dots,X_n)$ i.i.d. copies of $X\in\mathcal{X}$ and $(Z_1,\dots,Z_m)$ i.i.d. copies from $Z\in\mathcal{X}$. Denote by P_X and P_Z the distributions of X and Z, respectively. We now want to determine whether the two data sets come from the same distribution, i.e., whether $P_X = P_Z$. The idea behind MMD is to embed both distributions into an RKHS and measure their difference in the RKHS norm. Formally, let $k:\mathcal{X}\to\mathbb{R}$ be a pd kernel and $\mathcal{H}$ the corresponding RKHS, then MMD is defined by

  $$\mathrm{MMD}(P_X,P_Z) \coloneqq \|\Pi(P_X)-\Pi(P_Z)\|_{\mathcal{H}}^2.$$

  Under additional assumptions on the kernel k it is possible to show that the kernel mean embedding is injective, which means that $\mathrm{MMD}(P_X,P_Z)=0$ if and only if $P_X=P_Z$. We can thus see the MMD as a non-parametric distance between the two distributions P_X and P_Z. The MMD can be used to construct a hypothesis test for the null hypothesis that $X\overset{d}{=}Z$.

[4] Bochner integrals extend the Lebesgue integral to functions with values in Banach spaces.

- *Independence testing:* Assume we are given $(X_1^1, X_1^2), \dots, (X_n^1, X_n^2)$ i.i.d. copies of $X = (X^1, X^2) \in \mathcal{X}^1 \times \mathcal{X}^2$ and let P_X denote the joint distribution of X and P_{X^1} and P_{X^2} the respective marginal distributions. We now want to determine whether the two variables X^1 and X^2 are independent, i.e., $X^1 \perp\!\!\!\perp X^2$. Formally, $X^1 \perp\!\!\!\perp X^2$ if and only if $P_X = P_{X^1} \otimes P_{X^2}$. The idea behind the HSIC test is to embed both the joint P_X and the product of the marginals $P_{X^1} \otimes P_{X^2}$ into an RKHS and measure their difference in the RKHS-norm. Formally, let $k : \mathcal{X}^1 \times \mathcal{X}^2 \to \mathbb{R}$ be a pd kernel and $\mathcal{H}$ the corresponding RKHS, then HSIC is defined by

$$\mathrm{HSIC}(P_X) \coloneqq \|\Pi(P_X) - \Pi(P_{X^1} \otimes P_{X^2})\|_{\mathcal{H}}^2.$$

  Under additional assumptions on the kernel k it is possible to show that the kernel mean embedding is injective, which means that $\mathrm{HSIC}(P_X) = 0$ if and only if $X^1 \perp\!\!\!\perp X^2$. We can thus see the HSIC as a non-parametric measure of dependence between the variables X^1 and X^2. HSIC can be used to construct a hypothesis test for the null hypothesis that $X^1 \perp\!\!\!\perp X^2$.

Chapter 3

Lasso and Sparsity

3.1 Motivation of sparsity

Consider the following (fixed-design) linear regression model

$$Y = X\beta_0 + \varepsilon \quad \text{with } \mathbb{E}[\varepsilon] = 0.$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta_0 \in \mathbb{R}^p$, $\mathrm{Cov}(\varepsilon) = \sigma_0^2 \mathrm{I}$. If $X^\top X$ is invertible, we can compute the MSPE of the OLS estimator as follows

$$\begin{aligned}
\tfrac{1}{n}\mathbb{E}\big[\|X\beta_0 - X\widehat{\beta}^{\mathrm{OLS}}\|_2^2\big] &= \tfrac{1}{n}\mathbb{E}\big[(\beta_0 - \widehat{\beta}^{\mathrm{OLS}})^\top X^\top X(\beta_0 - \widehat{\beta}^{\mathrm{OLS}})\big] \\
&= \tfrac{1}{n}\mathbb{E}\big[\operatorname{trace}\big(X(\beta_0 - \widehat{\beta}^{\mathrm{OLS}})(\beta_0 - \widehat{\beta}^{\mathrm{OLS}})^\top X^\top\big)\big] \\
&= \tfrac{1}{n}\operatorname{trace}\big(\mathbb{E}\big[(\beta_0 - \widehat{\beta}^{\mathrm{OLS}})(\beta_0 - \widehat{\beta}^{\mathrm{OLS}})^\top\big] X^\top X\big) \\
&= \tfrac{1}{n}\operatorname{trace}\big(\operatorname{Var}\big(\widehat{\beta}^{\mathrm{OLS}}\big) X^\top X\big) \\
&= \tfrac{1}{n}\operatorname{trace}\big(\sigma_0^2 (X^\top X)^{-1} X^\top X\big) \\
&= \tfrac{p}{n}\sigma_0^2.
\end{aligned}$$

Here, we used the cyclic property and linearity of the trace. This shows that the MSPE of the OLS estimator depends on the number of parameters that need to be estimated. As shown in Section 2.3.1 (Theorem 2.14 and (2.3.4)), the MSPE for the ridge regression estimator (i.e., kernel ridge regression with linear kernel) has a similar form. In high-dimensional settings, where p is large such a dependence on p is however problematic.

What happens if we know that most of the coordinates in β_0 are equal to zero?

To formalize this question, we define

$$S_0 \coloneqq \{k \in \{1,\dots,p\} \mid \beta_0^k \neq 0\}$$

and assume that $s \coloneqq |S_0|$ is much smaller than p. Moreover, for all $\beta \in \mathbb{R}^p$ define $\|\beta\|_0 \coloneqq \big|\{k \in \{1,\dots,p\} \mid \beta^k \neq 0\}\big|$ that counts the non-zero entries of β.[1] Intuitively, if we would know the set S_0 we could simply apply the OLS estimator using the predictors X^{S_0} which would lead to an MSPE of $\frac{s}{n}\sigma_0^2$ which again would be small.

A naive way of extending this idea to the case where S_0 is unknown is to use the *best subset selection* estimator, which is defined by

$$\widehat{\beta}^{\mathrm{BSS}} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathcal{F}_{\mathrm{sparse}}} \|Y - X\beta\|_2^2,$$

[1] The function $\|\cdot\|_0$ is not actually a norm because it is not homogeneous (i.e., $\|c\beta\|_0 \neq c\|\beta\|_0$ for $c > 0$).

where $\mathcal{F}_{\text{sparse}} \coloneqq \{\beta \in \mathbb{R}^p \mid \|\beta\|_0 \leq s\}$. It is possible to show with high probability that

$$\tfrac{1}{n}\|X\beta_0 - X\widehat{\beta}^{\text{BSS}}\|_2^2 \leq \inf_{\beta \in \mathcal{F}_{\text{sparse}}} \tfrac{1}{n}\|X\beta_0 - X\beta\|_2^2 + \sigma_0^2 \tfrac{s}{n} \log(\tfrac{pe}{s}), \tag{3.1.1}$$

where e is Euler's number. A derivation of this bound is given in Wainwright [2019, Example 13.16]. In fact, it can even be shown that this is the minimax optimal rate for sparse linear regression [Wainwright, 2019, see Example 15.16]. While the best subset selection estimator has many desirable theoretical properties, it is only of limited practical use because it requires computation of $\binom{p}{s}$ OLS-estimators which is computationally infeasible even for relatively small p. In the following section, we will introduce an estimator that achieves this optimal rate (under additional assumptions) and is computationally efficient.

3.2 Lasso estimator

Instead of using the best subset selection estimator directly, we attempt to approximate it. To this end, define for all $q \in (0, \infty)$ the ℓ_q-norm[2] $\|\cdot\|_q$ for all $\beta \in \mathbb{R}^p$ by

$$\|\beta\|_q \coloneqq \left(\sum_{j=1}^{p} |\beta|^q\right)^{\frac{1}{q}}.$$

Instead of the best subset selection estimator, we can now consider all ℓ_q–norm estimators of the form

$$\widehat{\beta}_{q,s} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p : \|\beta\|_q \leq s} \|Y - X\beta\|_2^2.$$

If we use $\|\cdot\|_0$ (defined in the previous section) instead of $\|\cdot\|_q$ this corresponds to the best subset selection estimator. The optimization problem is convex (and hence relatively simple) whenever the s-balls $\{\beta \in \mathbb{R}^p \mid \|\beta\|_q \leq s\}$ are convex, which is the case if and only if $q \geq 1$.

This motivates using $q = 1$, which can be seen as the closest convex relaxation of the best subset selection estimator. Expressed in terms of the dual optimization problem[3], the estimator for $q = 1$ is given for a penalty parameter $\lambda > 0$ by

$$\widehat{\beta}^{\text{L}}_{\lambda} \coloneqq \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \tfrac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{3.2.1}$$

This estimator is called *least absolute shrinkage and selection operator* or *Lasso* for short. The minimization in (3.2.1) does not necessarily have a unique solution. In those cases any solution to the minimization is called a Lasso estimator or Lasso solution. The optimization in (3.2.1) can be efficiently solved using coordinate sub-gradient descent. As we will see later, the Lasso estimator achieves similar error rates as the best subset selection estimator. A key property of the ℓ_1-penalty is that it is actually able to set coordinates exactly equal to zero. This is particularly useful in the context of variable selection.

3.2.1 Prediction error

Using concentration bounds, which we discuss in Section 3.3, it is possible to provide a bound on the MSPE of the Lasso estimator, without making any assumptions on the design matrix X. The formal result is given in following theorem.

[2]This is only a true norm if $q \geq 1$, otherwise the functions are not subadditive.

[3]This terminology comes from Lagrangian optimization. The key idea is to switch between a constrained optimization and a penalized optimization.

Theorem 3.1 *(Lasso slow rate).*
Assume a fixed-design linear regression model

$$Y = X\beta_0 + \varepsilon,$$

where ε is sub-Gaussian[a] with parameter σ_0^2 and $\mathbb{E}[\varepsilon] = 0$. For $\lambda > 0$, let $\widehat{\beta}_\lambda^{\mathrm{L}}$ be a Lasso estimator. Then, with probability at least $1 - 2p^{-\left(\frac{1}{2}A_{n,p}^2 - 1\right)}$ it holds that

$$\tfrac{1}{n}\|X(\beta_0 - \widehat{\beta}_\lambda^L)\|_2^2 \leq 4\sigma_0 A_{n,p}\sqrt{\tfrac{\log(p)}{n}}\|\beta_0\|_1,$$

where $A_{n,p} \coloneqq \frac{\lambda}{\sigma_0}\sqrt{\frac{n}{\log(p)}}$.

[a] Sub-Gaussian random variables are formally defined in Definition 3.3 and include Gaussian random variables as a special case.

Proof. Since $\widehat{\beta}_\lambda^{\mathrm{L}}$ minimizes

$$\tfrac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

it holds that

$$\tfrac{1}{2n}\underbrace{\|Y - X\widehat{\beta}_\lambda^{\mathrm{L}}\|_2^2}_{=\|\varepsilon\|_2^2 + \|X(\beta_0 - \widehat{\beta}_\lambda^L)\|_2^2 + 2\varepsilon^\top X(\beta_0 - \widehat{\beta}_\lambda^{\mathrm{L}})} + \lambda\|\widehat{\beta}_\lambda^{\mathrm{L}}\|_1 \leq \tfrac{1}{2n}\underbrace{\|Y - X\beta_0\|_2^2}_{=\|\varepsilon\|_2^2} + \lambda\|\beta_0\|_1.$$

Therefore, rearranging the terms and applying Hölder's inequality leads to

$$\begin{aligned}\tfrac{1}{2n}\|X(\beta_0 - \widehat{\beta}_\lambda^L)\|_2^2 &\leq \tfrac{1}{n}\varepsilon^\top X(\widehat{\beta}_\lambda^L - \beta_0) + \lambda\|\beta_0\|_1 - \lambda\|\widehat{\beta}_\lambda^{\mathrm{L}}\|_1 \\ &\leq \tfrac{1}{n}\|\varepsilon^\top X\|_\infty\|\beta_0 - \widehat{\beta}_\lambda^L\|_1 + \lambda\|\beta_0\|_1 - \lambda\|\widehat{\beta}_\lambda^{\mathrm{L}}\|_1.\end{aligned}$$

Now, let $\Omega \coloneqq \{\frac{1}{n}\|\varepsilon^\top X\|_\infty \leq \lambda\}$, then on Ω it holds that

$$\tfrac{1}{2n}\|X(\beta_0 - \widehat{\beta}_\lambda^L)\|_2^2 \leq 2(\lambda\|\beta_0 - \widehat{\beta}_\lambda^L\|_1 + \lambda\|\beta_0\|_1 - \lambda\|\widehat{\beta}_\lambda^{\mathrm{L}}\|_1) \leq 4\lambda\|\beta_0\|_1.$$

To prove the theorem it remains to show that $\mathbb{P}(\Omega) \geq 1 - 2p^{-\left(\frac{1}{2}A_{n,p}^2 - 1\right)}$. This result requires concentration inequalities and makes use of the assumption that ε is sub-Gaussian. The formal result is given in Lemma 3.7 below. □

We can use Theorem 3.1 to investigate the asymptotic behavior of the Lasso estimator for large n and p. For this assume that $p = p_n$ depends on n and set $\lambda_n \coloneqq \frac{\sigma_0}{\sqrt{2}}\sqrt{\frac{\log(p_n)}{n}}a_n$, for an arbitrary sequence a_n. Then, the MSPE bound in Theorem 3.1 is equal to

$$4\sigma_0 A_{n,p}\sqrt{\tfrac{\log(p_n)}{n}}\|\beta_0\|_1 = \tfrac{4}{\sqrt{2}}\sigma_0\|\beta_0\|_1 \cdot a_n\sqrt{\tfrac{\log(p_n)}{n}} \tag{3.2.2}$$

and it holds with probability at least $1 - 2p_n^{-(a_n^2 - 1)}$. Therefore, as long as $\sqrt{\frac{\log(p_n)}{n}}$ converges to zero it is possible to select a_n to be divergent such that the MSPE bound (3.2.2) converges to zero while the probability of it being true converges to one.

For any $\delta \in (0, 1)$ we can also select a_n to be a fixed constant such that the probability of the MSPE bound being true is at least $1 - \delta$. In that case the MSPE bound converges at a rate of $\sqrt{\frac{\log(p_n)}{n}}$. However, in the high-dimensional regime in which $\sqrt{\frac{\log(p_n)}{n}}$ converges to zero this is slower than the minimax-optimal rate $\frac{\log(p_n)}{n}$ that the BSS achieves, see (3.1.1).

3.3 Basic concentration bounds

In the analysis of statistical estimators, it is often required to control the tail behavior of random variables. For example, in the proof of the central limit theorem in its simplest form, one assumes that the random variables have finite variance in order to ensure that convergence holds.

In this section, we explore some basic concentration bounds that ensure that random variables with well-behaved moments concentrate close to their mean with high probability.

3.3.1 Markov-type inequalities

The building block for a whole array of concentration bounds is Markov's inequality, which is stated in the following theorem.

Theorem 3.2 *(Markov's inequality).*
Given a non-negative random variable X it holds for all $t \in (0, \infty)$ that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Starting from the expectation of X (and assuming $\mathbb{E}[X] < \infty$) we get that

$$\mathbb{E}[X] = \mathbb{E}[X\mathbb{1}_{\{X<t\}}] + \mathbb{E}[X\mathbb{1}_{\{X\geq t\}}] \geq \mathbb{E}[X\mathbb{1}_{\{X<t\}}] + t\mathbb{E}[\mathbb{1}_{\{X\geq t\}}] \geq t\mathbb{P}(X \geq t).$$

Dividing by t proves the result. For the case that $\mathbb{E}[X] = \infty$ the result holds trivially, which completes the proof of Theorem 3.2. □

For any non-negative random variable with finite expectation, Markov's inequality implies that the probability of X being larger than t decays linearly as t grows. This can be improved, given that the random variable additionally has finite moments of higher order. For example, if X is a random variable with finite second moment, then we can apply Markov's inequality to $(X-\mathbb{E}[X])^2$ and t^2 to get *Chebyshev's inequality* which states for all $t \in (0, \infty)$ that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

In this case, large deviations from the mean decay at a rate of t^2. The same argument can be used in the case where the k-th absolute centered moment $\mathbb{E}[|X - \mathbb{E}[X]|^k]$ is finite. This implies that for all $t \in (0, \infty)$ it holds that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^k]}{t^k}.$$

A further way of exploiting finite higher moments, is to use the moment generating function

$$\Psi_X(\alpha) \coloneqq \mathbb{E}[e^{\alpha(X-\mu)}],$$

where $\mu \coloneqq \mathbb{E}[X]$. This function has the following properties

- it always exists at $\alpha = 0$ and $\Psi_X(0) = 1$,
- it may not exist for all α (or in fact for any $\alpha \neq 0$)
- $\Psi_X^{(k)}(0) = \mathbb{E}[(X - \mu)^k]$ (if it exists in an open neighborhood of 0) and
- for $X \sim \mathcal{N}(\mu, \sigma_0^2)$ the moment generating function is given by $\Psi(\alpha) = e^{\alpha^2 \frac{\sigma_0^2}{2}}$.

Assume X has a moment generating function that exists for all $\alpha \in [0, c]$. Using Markov's inequality we get for all $\alpha \in [0, c]$ and $t \in (0, \infty)$ that

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\alpha(X-\mu)} \geq e^{\alpha t}) \leq \frac{\mathbb{E}[e^{\alpha(X-\mu)}]}{e^{\alpha t}} = \frac{\Psi(\alpha)}{e^{\alpha t}}.$$

Since, the inequality holds for all $\alpha \in [0, c]$, we can choose it to make the bound as tight as possible. This leads to *Chernoff's bound*, which states for all $t \in (0, \infty)$ that

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\alpha \in [0,c]} \frac{\mathbb{E}[e^{\alpha(X-\mu)}]}{e^{\alpha t}}. \tag{3.3.1}$$

3.3.2 Sub-Gaussian bounds

As we saw in the previous section, it is possible to use both higher-order moments and the moment generating function of a random variable to derive bounds on how slow the tail of its distribution can decay. The intuition is that the more finite moments a random variable has, the faster its tails decay.

The Gaussian distribution is known to have finite moments of arbitrary degree and hence has fast-decaying tails. The following definition characterizes a class of random variables that have at least the same rate of decay in the tails as a Gaussian random variable.

Definition 3.3 *(Sub-Gaussian).*
A random variable X with mean μ is *sub-Gaussian* with parameter $\sigma > 0$ if for all $\alpha \in \mathbb{R}$ it holds that

$$\Psi_X(\alpha) = \mathbb{E}[e^{\alpha(X-\mu)}] \leq e^{\frac{\alpha^2}{2}\sigma^2}.$$

The upper bound corresponds to the moment generating function of a Gaussian random variable with variance σ^2. Hence, a Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$ is trivially sub-Gaussian with parameter σ. Using Chernoff's bound (3.3.1) on a sub-Gaussian random variable leads to the sub-Gaussian deviation bound.

Proposition 3.4 *(Sub-Gaussian deviation bound).*
Let X be sub-Gaussian with parameter σ and mean μ, then it holds for all $t \in (0, \infty)$ that

$$\mathbb{P}(|X - \mu| \geq t) \leq 2\exp(-\tfrac{t^2}{2\sigma^2}).$$

Proof. Using Chernoff's bound (3.3.1) and the definition of a sub-Gaussian random variable leads for all $t \in (0, \infty)$ to

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\alpha \in [0,\infty)} \frac{\mathbb{E}[e^{\alpha(X-\mu)}]}{e^{\alpha t}} \leq \inf_{\alpha \in [0,\infty)} \exp(\tfrac{\alpha^2}{2}\sigma^2 - \alpha t) = \exp(-\tfrac{t^2}{2\sigma^2}),$$

where we used that the infimum is attained at $\alpha = t/\sigma^2$. Moreover, we can apply the same argument to $-X$ since it is also sub-Gaussian with parameter σ, to get the same upper bound for $(\mu - X)$. Finally, combining both we get

$$\mathbb{P}(|X - \mu| \geq t) \leq \mathbb{P}(X - \mu \geq t) + \mathbb{P}(\mu - X \geq t) \leq 2\exp(-\tfrac{t^2}{2\sigma^2}),$$

which completes the proof of Proposition 3.4. □

Example 3.5 (Sub-Gaussian random variables)**.**

- Rademacher random variables: *If X is a Rademacher random variable, i.e., $X \in \{-1, 1\}$ with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 0.5$, then it is sub-Gaussian with parameter $\sigma = 1$.*

  A proof is given in the exercises (Assignment 2, exercise 1).

- Bounded random variables: *If X is a random variable such that $-\infty < a \leq X \leq b < \infty$, then X is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$.*

  A proof is given in the exercises (Week 5, exercise 2). Sometimes this result is also called Hoeffding's lemma.

- Sums of random variables: *If X and Y are sub-Gaussian (and not necessarily independent) with parameters σ_X and σ_Y, respectively, then $X+Y$ is sub-Gaussian with parameter $(\sigma_X + \sigma_Y)$.*

Proposition 3.6 *(Hoeffding's bound).*
Let $W_1, \ldots, W_n$ be independent mean zero random variables that are sub-Gaussian with parameter σ_i, respectively. Then, for all $\gamma \in \mathbb{R}^n$ it holds that

$$\sum_{i=1}^{n} \gamma_i W_i$$

is sub-Gaussian with parameter $(\sum_{i=1}^n \gamma_i^2 \sigma_i^2)^{1/2}$.

Proof. Using the independence of $W_1, \ldots, W_n$ and the definition of a sub-Gaussian random variable we get that

$$\mathbb{E}\Big[\exp\Big(\alpha \sum_{i=1}^{n} \gamma_i W_i\Big)\Big] = \prod_{i=1}^{n} \mathbb{E}\Big[\exp\Big(\alpha \gamma_i W_i\Big)\Big] \leq \prod_{i=1}^{n} \exp\Big(\frac{\alpha^2 \gamma_i^2}{2}\sigma^2\Big) = \exp\Big(\frac{\alpha^2}{2}\sum_{i=1}^{n}\gamma_i^2\sigma^2\Big),$$

which completes the proof of Proposition 3.6. □

Based on this result, we can prove that the max-norm of the weighted average of a sequence of n sub-Gaussian random variables concentrates around zero as n increases, which is the missing result we required in the proof of Theorem 3.1 above.

Lemma 3.7 *(Concentration bound for Lasso).*
Let $X \in \mathbb{R}^{n \times p}$ be a fixed (non-random) matrix with $\|X^j\|_2^2 = \sum_{i=1}^n (X_i^j)^2 = n$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ a vector of independent, mean-zero and sub-Gaussian random variables with parameter σ. Let $A = \frac{\lambda}{\sigma}\sqrt{\frac{n}{\log(p)}}$, then

$$\mathbb{P}(\tfrac{1}{n}\|X^\top \varepsilon\|_\infty > \lambda) \leq 2p^{-\left(\frac{1}{2}A^2 - 1\right)}.$$

Proof. First, use a union bound to get that

$$\mathbb{P}(\tfrac{1}{n}\|X^\top \varepsilon\|_\infty > \lambda) = \mathbb{P}\left(\bigcup_{j \in \{1,\ldots,p\}} \{\tfrac{1}{n}|\varepsilon^\top X^j| > \lambda\}\right) \leq \sum_{j=1}^{p} \mathbb{P}(\tfrac{1}{n}|\varepsilon^\top X^j| > \lambda).$$

Next, by Proposition 3.6 it holds that $\frac{1}{n}\varepsilon^\top X^j$ is sub-Gaussian with parameter $\sigma/n\|X^j\|_2 = \sigma/\sqrt{n}$. Hence, using the sub-Gaussian deviation bound (Proposition 3.4) and the definition of λ, it holds that

$$\sum_{j=1}^{p} \mathbb{P}(\tfrac{1}{n}|\varepsilon^\top X^j| > \lambda) \leq 2p \exp\Big(-\lambda^2 \frac{n}{2\sigma^2}\Big) = 2p^{-\left(\frac{1}{2}A^2 - 1\right)},$$

which completes the proof of Proposition 3.7. □

3.4 Outlook: Extended theory

The results discussed in this chapter only scratch the surface of how sparsity can be used to perform statistical analyses in high-dimensional data. In the following we discuss how the theory for Lasso can be extended.

Two immediate questions, that have not yet been answered by our analysis are (1) whether the rate in Theorem 3.1 can be improved and (2) whether the Lasso estimator can be used for variable selection, that is, whether it can recover the true set of active variables S_0.

(1) *Lasso fast rate (and random-design):* As discussed previously, the upper bound for the best-subset-selector is of order $\frac{s\log(p)}{n}$ which in the high-dimensional regime is smaller than the bound of order $\|\beta_0\|_1\sqrt{\frac{\log(p)}{n}}$ from Theorem 3.1. This begs the questions, whether the bound in Theorem 3.1 can be improved. With further assumptions on the design matrix X this is indeed possible, the result is known as the Lasso fast-rate (in contrast to the slow-rate given in Theorem 3.1). It states that for $\lambda = A\sqrt{\frac{\log(p)}{n}}$, with $A > 0$ constant, it holds with high probability that

$$\tfrac{1}{n}\|X(\beta_0 - \widehat{\beta}^{\mathrm{L}}_\lambda)\|_2^2 \leq C\sigma_0^2\tfrac{s\log(p)}{n}, \tag{3.4.1}$$

for some constant $C > 0$. The formal result and proof can be found in [Wainwright, 2019, Theorem 7.20]. In contrast to Theorem 3.1 the proof of this result uses assumptions on design matrix X. These assumptions on X can be shown to hold with high probability if X is for example multivariate Gaussian (random-design) [Wainwright, 2019, Theorem 7.16].

(2) *Lasso for variable selection:* In some applications the target of interest is the set of active variables S_0. One of the advantages of Lasso, over ridge regression, is that the ℓ_1-penalty is capable of setting some of the coordinates in the estimator $\widehat{\beta}^{\mathrm{L}}_\lambda$ exactly to zero. A potential estimator for S_0 is therefore given by

$$\widehat{S} \coloneqq \{j \in \{1,\dots,p\} \mid (\widehat{\beta}^{\mathrm{L}}_\lambda)^j \neq 0\}.$$

It turns out that this is indeed a consistent estimator of S_0 under additional assumptions on the design matrix. Details can be found in Wainwright [2019, Chapter 7.5].

Chapter 4

Double Machine Learning

In this chapter, we consider a method for semi-parametric inference with the aim of estimating a parameter θ_0 and constructing confidence intervals in the presence of a high-dimensional nuisance parameter η_0. These types of problems have a long history in statistics and have recently become popular in combination with modern machine learning techniques. We follow the double machine learning approach given in Chernozhukov et al. [2018].

4.1 Motivating example - Partially linear model

To illustrate the type of problems and the need for advanced methods, we start with a *partially linear model*. Specifically, let $\theta_0 \in \mathbb{R}$ and $g_0, m_0 \in \mathcal{F} \subseteq \{f \mid f : \mathbb{R}^p \to \mathbb{R}\}$, where $\mathcal{F}$ is a fixed function class.[1] We consider a random vector $Z = (D, X, Y) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$ which satisfies

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U \\ D &= m_0(X) + V, \end{aligned} \tag{4.1.1}$$

with $\mathbb{E}[U|X, D] = 0$ and $\mathbb{E}[V|X] = 0$. This is called a partially linear model because only the parameter of interest θ_0 is assumed to enter the assignment of Y linearly. It is important that in this model U and V are allowed to be dependent. Given n i.i.d. observations $Z_1, \dots, Z_n$ of Z our goal is to perform inference (potentially also beyond estimation, e.g., by constructing confidence intervals) on the parameter θ_0. Here, the nuisance parameter is $\eta_0 = (g_0, m_0)$.

Example 4.1 (Treatment effect in the presence of confounding)**.** *Assume we are interested in estimating the effect of a specific type of sleeping pill on the quality of sleep. We have access to data consisting of n patients. For each patient $i \in \{1, \dots, n\}$, the following variables are recorded:*

- *$D_i \in \{0, 1\}$ is a treatment indicator with $D_i = 1$ if the patient was treated with the drug and $D_i = 0$ otherwise.*
- *$Y_i \in \mathbb{R}$ is a score measuring the quality of sleep, with large values corresponding to good and small to bad sleep.*
- *$X_i \in \mathbb{R}^p$ are p different patient characteristics (e.g., age, sex, etc.).*

This data can be modeled by the partially linear model given in (4.1.1)*. It may help to think about this model as a sequential scheme with which the data is generated as follows: First, the patient characteristics X_i are drawn, then depending on the characteristics the treatment D_i is set and depending on the treatment and characteristics the sleep quality Y_i is drawn. In that case, it can be argued using the language of causality that the treatment effect of the sleeping pill corresponds to θ_0, see Remark 4.2. An obvious goal in this example is to estimate θ_0 and determine whether it is significantly larger than zero.*

[1] We actually do not require that g_0 and m_0 are from the same function space $\mathcal{F}$, we only make this assumption to ease notation.

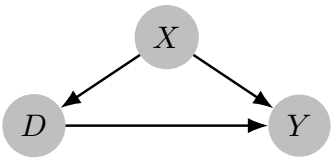

Figure 4.1: Directed causal graphical model for Example 4.1. Arrows correspond to direct causal effects.

Remark 4.2 (Causality)**.** *Causal models provide a mathematical language to describe, in a rigorous way, what is meant by a treatment effect. They go beyond standard statistical models, which only model a fixed data generating process, by additionally modeling the distributions of certain (mechanistic) changes (commonly called interventions) to the data generating process.*

The causal effect of the sleeping pill on the quality of sleep, given in Example 4.1, can be formally expressed using Pearl's do-notation [Pearl, 2009] as

$$\tau_{D\to Y} := \mathbb{E}[Y_i|\operatorname{do}(D_i = 1)] - \mathbb{E}[Y_i|\operatorname{do}(D_i = 0)].$$

The "do" in this notation indicates that the value of the treatment has been set (or intervened) to a specific value. This is different from conditioning:

> *For example, assume that X^j indicates whether a patient lives next to an airport. Now, if all people next to an airport sleep badly and therefore take the sleeping pill and all patients that do not live next to an airport sleep well and do not take the pills it can happen that*
>
> $$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] < 0 < \mathbb{E}[Y_i|\operatorname{do}(D_i = 1)] - \mathbb{E}[Y_i|\operatorname{do}(D_i = 0)].$$
>
> *That is, looking at the conditional expectations could lead to the mistaken conclusion that the sleeping pill worsens patients' sleep, while in fact it improves sleep quality.*

Causal models are often visualized using directed graphical models, in the case of Example 4.1 the corresponding graph is given in Figure 4.1. The causal details are not important for this course. The key point is that given the partially linear model and additional causal assumptions (such as no unobserved confounding), it holds that the causal treatment effect $\tau_{D\to Y}$ is equal to θ_0.

4.1.1 Estimating conditional expectations

There is a close connection between the conditional expectation, projections, minimizing the squared error loss and regression. More formally, let $(X, Y) \in \mathcal{X} \times \mathbb{R}$ be a random vector where $\mathcal{X}$ is a measurable space and Y has finite second moment.[2] Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote the underlying probability space, $L^2(\Omega, \mathcal{A}, \mathbb{P})$ the Hilbert space of square integrable real-valued random variables and $\sigma(X) \subseteq \mathcal{A}$ the sigma-algebra generated by X. Then, using that $L^2(\Omega, \sigma(X), \mathbb{P}) \subseteq L^2(\Omega, \mathcal{A}, \mathbb{P})$ is a closed subspace, the Hilbert-space projection theorem [e.g., Luenberger, 1997, Section 3, Theorem 2] implies that there is a unique minimizer $Z^* \in L^2(\Omega, \sigma(X), \mathbb{P})$ to

$$\inf_{Z\in L^2(\Omega,\sigma(X),\mathbb{P})} \mathbb{E}\big[(Y - Z)^2\big].$$

Furthermore, the minimizer Z^* satisfies that $Y - Z^*$ is orthogonal to $L^2(\Omega, \sigma(X), \mathbb{P})$, that is, for all $Z \in L^2(\Omega, \sigma(X), \mathbb{P})$ it holds that $\mathbb{E}[Z(Y - Z^*)] = 0$. This property implies that Z^* is an orthogonal projection of Y on X in the space $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

The conditional expectation $\mathbb{E}[Y|X]$, in the case that Y has finite second moments, is defined as the unique minimizer Z^*. Using that for any random variable $Z \in L^2(\Omega, \sigma(X), \mathbb{P})$ there exists

[2] Finite second moment is not required for the conditional expectation to exist but it makes its interpretation as a projection easier to understand.

a function $g \in \mathcal{G} \coloneqq \{g : \mathcal{X} \to \mathbb{R} \mid g \text{ is measurable and } \mathbb{E}[g(X)^2] < \infty\}$ such that $g(X) = Z$ almost surely, we get that $\mathbb{E}[Y|X]$ satisfies that

$$\inf_{g \in \mathcal{G}} \mathbb{E}\big[(Y - g(X))^2\big] = \mathbb{E}\big[(Y - \mathbb{E}[Y|X])^2\big] \tag{4.1.2}$$

and for all $g \in \mathcal{G}$ that

$$\mathbb{E}[g(X)(Y - \mathbb{E}[Y|X])] = 0.$$

This characterization of the conditional expectation, connects regression to estimating conditional expectations as follows. Consider a regression model

$$Y = f_0(X) + \varepsilon \quad \text{with} \quad \mathbb{E}[\varepsilon|X] = 0,$$

where $f_0 \in \mathcal{F}$ with $\mathcal{F} \subseteq \mathcal{G}$ a fixed function class. Then, by the linearity of the conditional expectation it immediately holds that $\mathbb{E}[Y|X] = f_0(X)$ almost surely and by (4.1.2) also that

$$f_0(X) = \mathbb{E}[Y|X] = \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbb{E}\big[(Y - f(X))^2\big].$$

Therefore, minimizing the mean squared error over a sufficiently large function class $\mathcal{F}$ corresponds to estimating the conditional expectation. The field of machine learning has made substantial progress on creating efficient and powerful tools to solve such optimization tasks with minimal assumptions on the function classes $\mathcal{F}$. These methods, however, rely on various complex regularization techniques and are in general hard to analyze theoretically.

4.1.2 Challenges when estimating θ_0

Our goal is to estimate and construct confidence intervals for the linear parameter θ_0 in the partially linear model (4.1.1). While this may appear straight-forward at first glance, there are two pitfalls that need to be accounted for.

Confounding bias A first attempt at estimating θ_0 might be to perform a linear regression of Y on D. However, the population OLS θ^{OLS} (i.e., the best linear approximation to $\mathbb{E}[Y|D]$) is equal to

$$\theta^{\text{OLS}} = \frac{\mathbb{E}[YD]}{\mathbb{E}[D^2]} = \theta_0 + \frac{\mathbb{E}[g_0(X)D]}{\mathbb{E}[D^2]},$$

which is not equal to θ_0 if $\mathbb{E}[g_0(X)D] \neq 0$. Therefore, a simple OLS estimator in general leads to a biased estimate of θ_0. This type of bias is called *confounding bias*, because it stems from the fact that X and D are dependent. This type of bias is a central object of study in causality. In the partially linear model there are two fundamental ways of accounting for it; (1) adjustment and (2) inverse probability weighting. In this course we focus only on (1).

Adjustment is based solely on the outcome model, that is, the equation for Y in the partially linear model (4.1.1). It estimates the full conditional expectation $\mathbb{E}[Y|D, X]$ (see Section 4.1.1) and then marginalize out X. Formally, using the partially linear model, we get that the conditional expectation $\mathbb{E}[Y|D, X]$ satisfies

$$\mathbb{E}[Y|D, X] = D\theta_0 + g_0(X).$$

Hence, once we have estimated the conditional expectation $\mathbb{E}[Y|D, X]$, we can construct estimators to extract θ_0. One option is to observe that θ_0 satisfies

$$\theta_0 = \mathbb{E}\left[\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X]\right].$$

Given an estimator $\widehat{f}$ of the conditional mean this leads to the estimator

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{f}(1, X_i) - \widehat{f}(0, X_i)). \tag{4.1.3}$$

A second option is to consider the function class

$$\mathcal{G}_{\text{plm}} \coloneqq \{f : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R} \mid \exists \theta \in \mathbb{R}, g \in \mathcal{F} \text{ s.t. } \forall d \in \mathbb{R}, x \in \mathbb{R}^p : \ f(d,x) = d\theta + g(x)\}.$$

Then, it holds that $\mathbb{E}[Y|D,X]$ is the almost surely unique minimizer of $\inf_{f \in \mathcal{G}_{\text{plm}}} \mathbb{E}[(Y - f(D,X))]$. Optimizing over this class directly results in an estimator of θ_0. In practice, one might consider a penalized estimator of the form

$$(\widehat{\theta}, \widehat{g}) = \underset{(\theta,g) \in \mathbb{R} \times \mathcal{F}}{\arg\min} \sum_{i=1}^{n} (Y_i - D_i\theta - g(X_i))^2 + \lambda \|g\|_{\mathcal{F}}^2. \tag{4.1.4}$$

Both (4.1.3) and (4.1.4) can be shown to be consistent estimators of θ_0, under relatively mild assumptions on the model and estimating procedure.

Regularization and overfitting bias While the adjustment based estimators asymptotically account for the confounding bias, we may also be interested in constructing confidence intervals (e.g., to test whether the parameter θ_0 is significantly larger than zero). To this end, it would be desirable for the estimators to be asymptotically normal, i.e., that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Unfortunately, for the adjustment estimators proposed above, this is in general not the case if the function class $\mathcal{F}$ and hence the conditional expectation $\mathbb{E}[Y|D,X]$ is too complex, see Example 4.3 below. In those cases, we need to employ more advanced machine learning procedures such as random forests, kernel ridge estimators, neural networks or Lasso. Employing these methods, leads to two further sources of bias in the estimation:

(1) *Regularization bias:* Machine learning methods are able to estimate flexible functions by regularizing in various ways (e.g., kernel-ridge regression or Lasso). This adds a bias to the estimation of the part of the conditional expectation $\mathbb{E}[Y|D,X]$ corresponding to X that then spills over to the asymptotic behavior of a naive estimator of θ_0.

(2) *Overfitting bias:* A further, but separate, source of bias is incurred if the same data is used twice. For example, in (4.1.3), if the same data was used for estimating $\widehat{f}$ and computing the average, then the evaluations $\widehat{f}(X_i)$ are too optimistic leading to an overfitting bias.

As we show in the next section, DML addresses both of these sources of bias: For (1) it uses an orthogonalization technique to estimate θ_0 and for (2) it uses a type of sample splitting called cross-fitting. These are detailed in the following section.

4.1.3 DML for the partially linear model

Let us formalize the discussion by introducing for each sample size $n \in \mathbb{N}$, two machine learning estimators $\widehat{g}_n$ and $\widehat{m}_n$ that estimate g_0 and m_0, respectively, based on i.i.d. observations $Z_1, \ldots, Z_n$. We make no assumptions on the type of machine learning estimators, for example, they could be random forests, kernel ridge estimators, neural networks or Lasso.

The partially linear model (4.1.1) is parameterized by the parameter of interest θ_0 and the nuisance parameters $\eta_0 = (g_0, m_0)$. DML proposes a three step procedure that first uses $\widehat{g}_n$ and $\widehat{m}_n$ to estimate η_0 and then use them to estimate θ_0 efficiently. A simplified version is given by the following step-wise procedure:

(i) Split the data into two data sets I_1 and I_2.

(ii) Estimate g_0 and m_0 using only the data in I_1 resulting in the estimators $\widehat{g}_n$ and $\widehat{m}_n$.

(iii) On the remaining data I_2, estimate θ_0 as the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n}\Psi(Z_i,\widehat{\theta},(\widehat{g}_n,\widehat{m}_n))=0, \tag{4.1.5}$$

where Ψ is an appropriately chosen *score function* of the data and the parameters θ and η.

The two ingredients that make this procedure work are the sample splitting and the appropriately chosen function Ψ that ensures that the estimation of θ_0 is separated from the estimation of η_0.

Constructing the score function Ψ Let us begin by attempting to construct a function Ψ for which the resulting estimator $\widehat{\theta}$ is asymptotically normal. Motivated by the confounding bias, a first attempt might be to use the estimate $\widehat{g}_n$ to remove the effect due to confounding from Y and then linearly regress the resulting residuals on D. Formally, we define the residuals

$$R \coloneqq Y-\widehat{g}_n(X). \tag{4.1.6}$$

Then, the OLS estimator of R on D is given by

$$\widehat{\theta}_A \coloneqq \frac{\sum_{i=1}^{n}R_iD_i}{\sum_{i=1}^{n}D_i^2}, \tag{4.1.7}$$

which in particular implies that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\widehat{\theta}_AD_i-\widehat{g}_n(X_i))D_i=0.$$

Hence, this corresponds to using the estimator in step (iii) with the score function $\Psi:\mathbb{R}^{d+1}\times\mathbb{R}\times\mathcal{F}^2\to\mathbb{R}$ defined for all $(d,x)\in\mathbb{R}^{d+1}$, $\theta\in\mathbb{R}$ and $(g,m)\in\mathcal{F}^2$ by

$$\Psi((d,x),\theta,(g,m))=(y-d\theta-g(x))d.$$

Unfortunately, this naive method is not asymptotically normal. We illustrate this empirically based on the following numerical example, slightly adapted from Chernozhukov et al. [2018].

Example 4.3 (Numerical example of partially linear model)**.** *Define the functions $m_0:\mathbb{R}^{20}\to\mathbb{R}$ and $g_0:\mathbb{R}^{20}\to\mathbb{R}$ for all $x\in\mathbb{R}^{20}$ by*

$$m_0(x)=x^1+\frac{\exp(x^3)}{4+4\exp(x^3)} \quad \textit{and} \quad g_0(x)=\frac{\exp(x^1)}{1+\exp(x^1)}+\frac{x^3}{4},$$

respectively. Moreover, let $\Sigma\in\mathbb{R}^{20\times 20}$ be the matrix satisfying for all $k,j\in\{1,\dots,20\}$ that

$$\Sigma_{k,j}=0.7^{|j-k|}.$$

We then consider the partially linear model given by

$$\begin{aligned}Y&=D\theta_0+g_0(X)+U\\ D&=m_0(X)+V,\end{aligned}$$

with $X\sim\mathcal{N}(0,\Sigma)$, $U\sim\mathcal{N}(0,1)$ and $V\sim\mathcal{N}(0,1)$ independent. Assume now we observe n i.i.d. samples $(Y_1,X_1,D_1),\dots,(Y_n,X_n,D_n)$ from this model. We now apply the naive estimator described above using a random forest to estimate g_0. To assess whether this estimator is empirically asymptotically normal, we then perform 500 repetitions of this experiment and record the estimated value $\widehat{\theta}$ in each repetition. The resulting plot is shown in Figure 4.2 (top and bottom left) for $n=50$ and $n=500$. The estimates do not seem to converge to a normal distribution, indicating that it is not an asymptotically normal estimator. The DML estimator (introduced below) corrects appropriately for the bias and leads to an asymptotically normal estimator (see Figure 4.2 (right)).

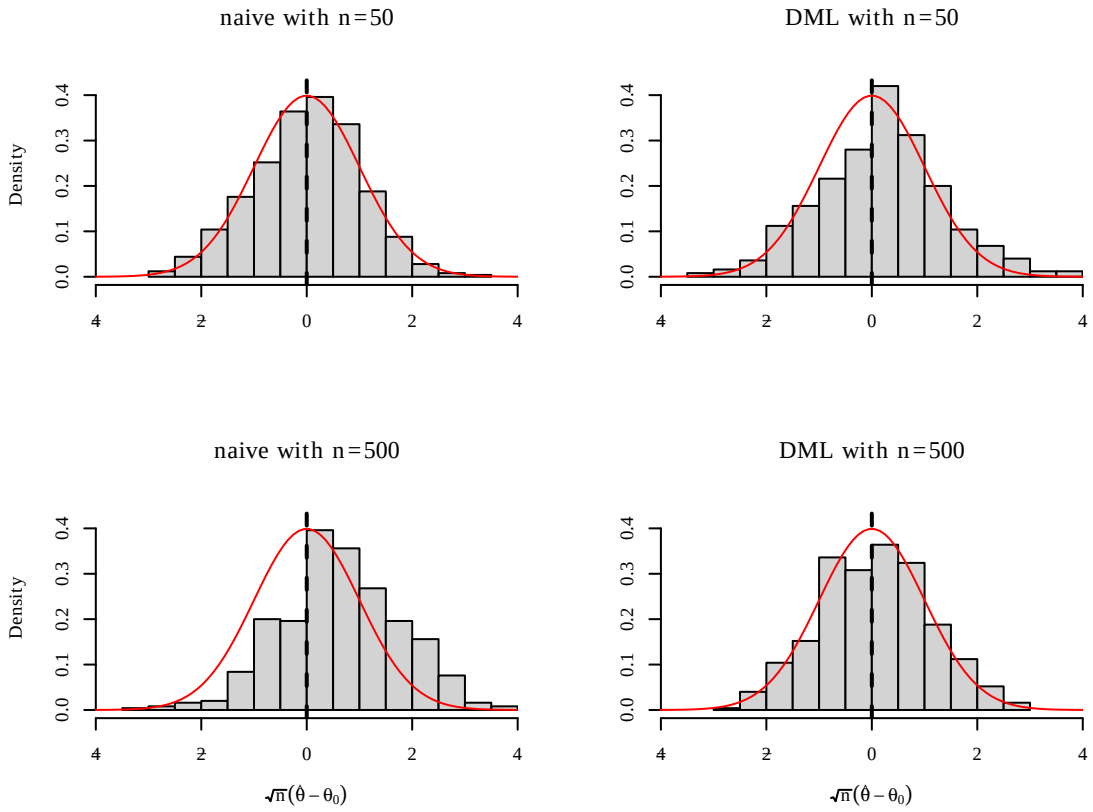

Figure 4.2: Histogram of estimate $\widehat{\theta}$ for 1000 repetitions from the same data generating model. In red the density of a standard normal distribution corresponding to the theoretically correct asymptotic distribution of the DML estimator. The naive estimator is clearly biased.

Why does the naive estimator fail to be asymptotically normal?

Let's look at the naive approach in more detail. Expanding the residuals R, defined in (4.1.6), leads to

$$R = (g_0(X) - \widehat{g}_n(X)) + D\theta_0 + U. \tag{4.1.8}$$

The problem in the subsequent OLS regression is that it can happen that $\mathbb{E}[(g_0(X)-\widehat{g}_n(X))D] \neq 0$ since D depends on X via m_0, which then leads to a bias in the estimate of θ_0. This bias persists asymptotically if the estimator $\widehat{g}_n$ converges too slowly. To make this more precise, we can compute the asymptotic bias of $\widehat{\theta}_A$. Consider for simplicity that $\widehat{g}_n$ is non-random. Then, the asymptotic bias can be computed using the law of large numbers on both the numerator and denominator in (4.1.7) combined with the continuous mapping theorem. This results in

$$\text{bias}(\widehat{\theta}_A) \sim \frac{\mathbb{E}[DR]}{\mathbb{E}[D^2]} - \theta_0 = \frac{\mathbb{E}[(g_0(X) - \widehat{g}_n(X))D]}{\mathbb{E}[D^2]}, \tag{4.1.9}$$

where we used the partially linear model (4.1.1) and (4.1.8) to simply the expression. For this bias term to not affect that asymptotic normality, we therefore require that (4.1.9) decays at a rate that is faster than the $\sqrt{n}$-parametric rate. This rate is, however, unachievable for complex model classes, such as non-parametric or high-dimensional models (see for example Theorem 2.17 applied with the Sobolev kernel).

To avoid this, one can use the estimate $\widehat{m}_n$ of m_0 to additionally remove the confounding in D. Consider the following estimator

$$\widehat{\theta}_B \coloneqq \frac{\sum_{i=1}^n (D_i - \widehat{m}_n(X_i))R_i}{\sum_{i=1}^n (D_i - \widehat{m}_n(X_i))D_i}.$$

Similarly as before with $\widehat{\theta}_A$, the asymptotic bias of $\widehat{\theta}_B$ can be computed (assuming both $\widehat{m}_n$ and $\widehat{g}_n$ are non-random for simplicity) to be

$$\text{bias}(\widehat{\theta}_B) \sim \frac{\mathbb{E}[(D - \widehat{m}_n(X))R]}{\mathbb{E}[(D - \widehat{m}_n(X))D]} - \theta_0 = \frac{\mathbb{E}[(m_0(X) - \widehat{m}_n(X))(g_0(X) - \widehat{g}_n(X))]}{\mathbb{E}[(D - \widehat{m}_n(X))D]},$$

where we again used the partially linear model (4.1.1) to simplify the expression. While this still results in a non-zero bias, the term depends on the product of the two terms $m_0(X) - \widehat{m}_n(X)$ and $g_0(X) - \widehat{g}_n(X)$. This is important because even if $\widehat{m}_n$ and $\widehat{g}_n$ do not converge at the parametric rate, their product may converge sufficiently fast. The DML estimator is based on the score function corresponding to the estimator $\widehat{\theta}_B$. A formal definition of the DML estimator for the partially linear model is given in the following definition.

Definition 4.4 *(DML estimator for the partially linear model)*.
Let $\widehat{g}_n$ and $\widehat{m}_n$ be machine learning estimators for g_0 and m_0, respectively. For all $n \in \mathbb{N}$, split the observation indices into K disjoint sets $I_1, \dots, I_K$ such that $I_1 \cup \dots \cup I_K = \{1, \dots, n\}$ and $|I_k| \approx n/K$. Furthermore, for all $n \in \mathbb{N}$ and all $k \in \{1, \dots, K\}$ define

$$\widehat{m}_n^k \coloneqq \widehat{m}_{|I_k^c|}((Z_i)_{i \in I_k^c}) \quad \text{and} \quad \widehat{g}_n^k \coloneqq \widehat{g}_{|I_k^c|}((Z_i)_{i \in I_k^c}),$$

where $I_k^c = \{1, \dots, n\} \setminus I_k$. Moreover, for $k \in \{1, \dots, K\}$, let $\widehat{\theta}_n^k$ be the solution to

$$\frac{1}{|I_k|} \sum_{i \in I_k} (D_i - \widehat{m}_n^k(X_i))(Y_i - D_i\widehat{\theta}_n^k - \widehat{g}_n^k(X_i)) = 0. \tag{4.1.10}$$

Then the *DML estimator* $\widehat{\theta}_n^{\mathrm{DML}}$ is defined by

$$\widehat{\theta}_n^{\mathrm{DML}} := \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}_n^k.$$

The following theorem shows that given that the partially linear model is sufficiently well-behaved and given that the machine learning estimators have a sufficiently fast rate of convergence, the DML estimator will indeed be asymptotically normal.

Theorem 4.5 *(Asymptotic normality of DML for partially linear model).*
Assume that the partially linear model in (4.1.1) satisfies $\mathbb{E}[V^2U^2] < \infty$, $\mathbb{E}[D^2] < \infty$, $\mathbb{E}[V^2] > 0$ and that $\mathbb{E}[U^2|X]$ and $\mathbb{E}[V^2|X]$ are almost surely bounded. Moreover, assume for all $n \in \mathbb{N}$ that the machine learning estimators $\widehat{m}_n$ and $\widehat{g}_n$ trained on n i.i.d. observations $Z_1, \ldots, Z_n$ are L^2-consistent and satisfy for X independent of the training observations that

$$\lim_{n\to\infty} \sqrt{n} \cdot \mathbb{E}[(m_0(X) - \widehat{m}_n(X))^2]^{\frac{1}{2}} \mathbb{E}[(g_0(X) - \widehat{g}_n(X))^2]^{\frac{1}{2}} = 0.$$

Then, for $\sigma^2 := \mathbb{E}[V^2]^{-2}\mathbb{E}[V^2U^2]$ it holds that

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{DML}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as } n \to \infty.$$

Proof. For notational simplicity, we assume the full sample size is nK and that each split $I_1, \ldots, I_K$ contains exactly n observations.

We start by expanding (4.1.10) to see that $\widehat{\theta}_n^k$ satisfies

$$\widehat{\theta}_n^k \left(\frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))D_i\right) = \sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))(Y_i - \widehat{g}_n^k(X_i)).$$

Then, defining $J_n := \frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))D_i$, it holds that on the event $\Omega = \{|J_n| > 0\}$, the estimator $\widehat{\theta}_n^k$ exists and is given by

$$\widehat{\theta}_n^k = \left(\frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))D_i\right)^{-1}\left(\frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))(Y_i - \widehat{g}_n^k(X_i))\right).$$

Below, we show that J_n converges to zero in probability, which further implies that $\lim_{n\to\infty}\mathbb{P}(\Omega) = 1$. We can therefore assume that the event Ω holds for the remainder of the proof.

Next, we decompose the scaled estimation error using the structure of the partially linear model as follows

$$\begin{aligned}
&\sqrt{n}(\widehat{\theta}_n^k - \theta_0) \\
&= \sqrt{n}J_n^{-1}\left[\frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))(Y_i - \widehat{g}_n^k(X_i)) - J_n\theta_0\right] \\
&= \sqrt{n}J_n^{-1}\left[\frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))U_i + \frac{1}{n}\sum_{i\in I_k}(D_i - \widehat{m}_n^k(X_i))(g_0(X_i) - \widehat{g}_n^k(X_i))\right] \\
&= \sqrt{n}J_n^{-1}\Bigg[\frac{1}{n}\sum_{i\in I_k}(m_0(X_i) - \widehat{m}_n^k(X_i))U_i + \frac{1}{n}\sum_{i\in I_k}V_iU_i \\
&\qquad + \frac{1}{n}\sum_{i\in I_k}(m_0(X_i) - \widehat{m}_n^k(X_i))(g_0(X_i) - \widehat{g}_n^k(X_i)) + \frac{1}{n}\sum_{i\in I_k}V_i(g_0(X_i) - \widehat{g}_n^k(X_i))\Bigg] \\
&= J_n^{-1}a_n^* + J_n^{-1}b_n^* + J_n^{-1}c_n^* + J_n^{-1}d_n^*, \qquad (4.1.11)
\end{aligned}$$

where we defined

- $a_n^* = \frac{1}{\sqrt{n}} \sum_{i \in I_k} V_i U_i$,
- $b_n^* = \frac{1}{\sqrt{n}} \sum_{i \in I_k} (m_0(X_i) - \widehat{m}_n^k(X_i))(g_0(X_i) - \widehat{g}_n^k(X_i))$,
- $c_n^* = \frac{1}{\sqrt{n}} \sum_{i \in I_k} U_i (m_0(X_i) - \widehat{m}_n^k(X_i))$,
- $d_n^* = \frac{1}{\sqrt{n}} \sum_{i \in I_k} V_i (g_0(X_i) - \widehat{g}_n^k(X_i))$.

We will now analyze the convergence of the terms J_n, b_n^*, c_n^* and d_n^* separately. More precisely, we will show that $J_n \xrightarrow{\mathbb{P}} \mathbb{E}[V^2]$, $b_n^* \xrightarrow{\mathbb{P}} 0$, $c_n^* \xrightarrow{\mathbb{P}} 0$ and $d_n^* \xrightarrow{\mathbb{P}} 0$.
Term J_n: We want to show that this term converges in probability to $\mathbb{E}[V^2]$. To see this, fix $\varepsilon > 0$ and use the partially linear model together with the triangle inequality to get that

$$\begin{aligned}
&\mathbb{P}(|J_n - \mathbb{E}[V^2]| > \varepsilon) \\
&\quad = \mathbb{P}(|\tfrac{1}{n} \textstyle\sum_{i \in I_k} (D_i - \widehat{m}_n^k(X_i)) D_i - \mathbb{E}[V^2]| > \varepsilon) \\
&\quad = \mathbb{P}(|\tfrac{1}{n} \textstyle\sum_{i \in I_k} (m_0(X_i) - \widehat{m}_n^k(X_i)) D_i + \tfrac{1}{n} \textstyle\sum_{i \in I_k} V_i D_i - \mathbb{E}[V^2]| > \varepsilon) \\
&\quad \leq \mathbb{P}(|\tfrac{1}{n} \textstyle\sum_{i \in I_k} (m_0(X_i) - \widehat{m}_n^k(X_i)) D_i| > \varepsilon) + \mathbb{P}(|\tfrac{1}{n} \textstyle\sum_{i \in I_k} V_i D_i - \mathbb{E}[V^2]| > \varepsilon). \qquad (4.1.12)
\end{aligned}$$

Next, we consider the two summands separately. For the first summand in (4.1.12), use the triangle and Cauchy-Schwarz inequalities to get that

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n} \sum_{i \in I_k} (m_0(X_i) - \widehat{m}_n^k(X_i)) D_i\right|\right] &\leq \frac{1}{n} \sum_{i \in I_k} \mathbb{E}\left[|(m_0(X_i) - \widehat{m}_n^k(X_i)) D_i|\right] \\
&\leq \mathbb{E}\left[(m_0(X) - \widehat{m}_n^k(X))^2\right]^{\frac{1}{2}} \mathbb{E}\left[D^2\right]^{\frac{1}{2}}.
\end{aligned}$$

By our assumptions this implies that $\lim_{n \to \infty} \mathbb{E}\left[|\frac{1}{n} \sum_{i \in I_k} (m_0(X_i) - \widehat{m}_n^k(X_i)) D_i|\right] = 0$ and hence this term also converges to zero in probability, i.e., $\lim_{n \to \infty} \mathbb{P}(|\frac{1}{n} \sum_{i \in I_k} (m_0(X_i) - \widehat{m}_n^k(X_i)) D_i| > \varepsilon) = 0$. For the second summand in (4.1.12), the weak law of large numbers implies that $\frac{1}{n} \sum_{i \in I_k} V_i D_i$ converges in probability to $\mathbb{E}[V_i D_i] = \mathbb{E}[V_i m_0(X_i) + V_i V_i] = \mathbb{E}[V_i^2]$ (using the partially linear model). Hence, we have shown that

$$\lim_{n \to \infty} \mathbb{P}(|J_n - \mathbb{E}[V^2]| > \varepsilon) = 0,$$

or equivalently J_n converges in probability to $\mathbb{E}[V^2]$.
Term b_n^*: Applying the triangle and Cauchy-Schwarz inequality, we get that

$$\begin{aligned}
\mathbb{E}[|b_n^*|] &\leq \frac{1}{\sqrt{n}} \sum_{i \in I_k} \mathbb{E}[|(m_0(X_i) - \widehat{m}_n^k(X_i))(g_0(X_i) - \widehat{g}_n^k(X_i))|] \\
&\leq \sqrt{n} \cdot \mathbb{E}[(m_0(X) - \widehat{m}_n^k(X))^2]^{\frac{1}{2}} \mathbb{E}[(g_0(X) - \widehat{g}_n^k(X))^2]^{\frac{1}{2}}.
\end{aligned}$$

Hence, by assumption b_n^*, converges to zero.
Terms c_n^* and d_n^*: The method for showing that each term converges to zero uses the same arguments. Here, we only show it for c_n^*. First, using standard properties of the conditional expectation, independence of the samples in I_k and I_k^c and the assumption $\mathbb{E}[U|X, D] = 0$, we get

that

$$\begin{aligned}
\mathbb{E}[c_n^*] &= \mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i\in I_k} U_i(m_0(X_i) - \widehat{m}_n^k(X_i))\right] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i\in I_k} \mathbb{E}[U_i(m_0(X_i) - \widehat{m}_n^k(X_i))|X_i, (Z_\ell)_{\ell\in I_k^c}]\right] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i\in I_k} \mathbb{E}[U_i|X_i](m_0(X_i) - \widehat{m}_n^k(X_i))\right] \\
&= 0. \qquad (4.1.13)
\end{aligned}$$

Next, using independence of the samples and that I_k and I_k^c are disjoint sets of observations, we get

$$\begin{aligned}
\mathbb{E}[(c_n^*)^2] &= \frac{1}{n}\sum_{i,j\in I_k} \mathbb{E}\left[U_i(m_0(X_i) - \widehat{m}_n^k(X_i))U_j(m_0(X_j) - \widehat{m}_n^k(X_j))\right] \\
&= \frac{1}{n}\sum_{i\neq j\in I_k} \mathbb{E}\left[\mathbb{E}\left[U_i(m_0(X_i) - \widehat{m}_n^k(X_i))|(Z_\ell)_{\ell\in I_k^c}\right]\mathbb{E}\left[U_j(m_0(X_j) - \widehat{m}_n^k(X_j))|(Z_\ell)_{\ell\in I_k^c}\right]\right] \\
&\qquad + \frac{1}{n}\sum_{i\in I_k} \mathbb{E}\left[U_i^2(m_0(X_i) - \widehat{m}_n^k(X_i))^2\right] \\
&= \frac{1}{n}\sum_{i\neq j\in I_k} \mathbb{E}\left[\mathbb{E}\left[U_i(m_0(X_i) - \widehat{m}_n^k(X_i))|X_i, (Z_\ell)_{\ell\in I_k^c}\right]\right] \\
&\qquad\qquad \cdot \mathbb{E}\left[\mathbb{E}\left[U_j(m_0(X_j) - \widehat{m}_n^k(X_j))|X_i, (Z_\ell)_{\ell\in I_k^c}\right]\right] \\
&\qquad + \frac{1}{n}\sum_{i\in I_k} \mathbb{E}\left[\mathbb{E}\left[U_i^2(m_0(X_i) - \widehat{m}_n^k(X_i))^2|X_i, (Z_\ell)_{\ell\in I_k^c}\right]\right] \\
&= \frac{1}{n}\sum_{i\neq j\in I_k} \mathbb{E}\left[\mathbb{E}\left[U_i|X_i\right](m_0(X_i) - \widehat{m}_n^k(X_i))\right]\mathbb{E}\left[\mathbb{E}\left[U_j|X_i\right](m_0(X_j) - \widehat{m}_n^k(X_j))\right] \\
&\qquad + \frac{1}{n}\sum_{i\in I_k} \mathbb{E}\left[\mathbb{E}\left[U_i^2|X_i\right](m_0(X_i) - \widehat{m}_n^k(X_i))^2\right] \\
&\leq C\cdot\mathbb{E}\left[(m_0(X) - \widehat{m}_n^k(X))^2\right], \qquad (4.1.14)
\end{aligned}$$

where $C > 0$ is a constant that comes from the assumption that $\mathbb{E}[U^2|X]$ is bounded. Finally, using (4.1.13) and (4.1.14), we can apply Chebyshev's inequality to get for all $\varepsilon > 0$ that

$$\limsup_{n\to\infty} \mathbb{P}(|c_n^*| \geq \varepsilon) \leq \limsup_{n\to\infty} \frac{\mathbb{E}[(c_n^*)^2]}{\varepsilon^2} \leq \limsup_{n\to\infty} \frac{C\cdot\mathbb{E}\left[(m_0(X) - \widehat{m}_n^k(X))^2\right]}{\varepsilon^2} = 0.$$

Hence, we have shown that c_n^* converges to zero in probability.

The arguments above hold for all $\widehat{\theta}_n^k$, so using the definition of $\widehat{\theta}_n^{\mathrm{DML}}$, we get that

$$\begin{aligned}\sqrt{Kn}\left(\widehat{\theta}_n^{\mathrm{DML}} - \theta_0\right) &= \sqrt{Kn}\left(\frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}_n^k - \theta_0\right)\\ &= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{n}\left(\widehat{\theta}_n^k - \theta_0\right)\\ &= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left[J_n^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I_k}U_iV_i + o_{\mathbb{P}}(1)\right]\\ &= \frac{1}{\mathbb{E}[V^2]}\frac{1}{\sqrt{Kn}}\sum_{i=1}^{Kn}U_iV_i + o_{\mathbb{P}}(1),\end{aligned}$$

where we used the continuous mapping theorem to combine the convergence in probability of the different terms. Furthermore, defining $\sigma^2 \coloneqq \mathbb{E}[V^2]^{-2}\mathbb{E}[V^2U^2]$ and using the assumption $\mathbb{E}[V^2U^2] < \infty$, we can apply the central limit theorem to get that

$$\sqrt{Kn}\left(\widehat{\theta}_n^{\mathrm{DML}} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

This completes the proof of Theorem 4.5. □

4.1.4 Inference in sparse high-dimensional linear models

A useful feature of DML is that it can be used to construct confidence intervals in sparse high-dimensional linear models. Keep in mind that this is a non-trivial task because fitting sparse models (e.g., using Lasso) requires regularization (see Chapter 3) and regularization leads to biased estimates.

Assume we observe $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$ i.i.d. observations from a (random-design) sparse high-dimensional linear model given by

$$Y = \beta_0^\top X + U, \quad \text{with } \mathbb{E}[U|X] = 0.$$

Our goal is to construct confidence intervals for the parameter β_0^j for a fixed coordinate $j \in \{1, \ldots, p\}$. In order to apply DML, we assume that the covariates X additionally satisfy a linear model. Formally, assume the samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfy the joint high-dimensional sparse linear model

$$\begin{aligned} &Y = \beta_0^\top X + U \\ &X = B_0 X + V, \end{aligned} \quad \text{and} \quad \begin{aligned} &\mathbb{E}[U|X] = 0 \\ &\mathbb{E}[V^j|X^{-j}] = 0 \quad \forall j \in \{1, \ldots, p\} \end{aligned} \tag{4.1.15}$$

where $\beta_0 \in \mathbb{R}^p$, $B_0 \in \mathbb{R}^{p\times p}$ with zeros on the diagonal and $\mathrm{I} - B_0$ invertible.[3] To be able to estimate the parameters sufficiently well we assume that the model is sparse in the sense that for all $j \in \{1, \ldots, p\}$ it holds that

$$s_0^j \coloneqq \max\left(\|\beta_0^{-j}\|_0, \|(B_0)_j^{-j}\|_0\right)$$

is sufficiently small for all $j \in \{1, \ldots, p\}$. For a fixed coordinate $j \in \{1, \ldots, p\}$ we can use (4.1.15) to construct a partially linear model as in (4.1.1) as follows

$$\begin{aligned} Y &= X^j\beta_0^j + \gamma_0^\top X^{-j} + U \\ X^j &= \alpha_0^\top X^{-j} + V^j, \end{aligned} \tag{4.1.16}$$

[3] Here, the superscript $-j$ corresponds to selecting all but the j-th index.

where $\gamma_0 \coloneqq \beta_0^{-j}$ and $\alpha_0 \coloneqq (B_0)_j^{-j}$. Here, the nuisance parameter is $\eta_0 = (\gamma_0, \alpha_0)$, which can be estimated consistently and with known rates using Lasso. Hence, denote by $\widehat{\alpha}_n^j$ the Lasso estimator for the regression of X^j on X^{-j} (i.e., an estimator of α_0) and by $\widehat{\gamma}_n^j$ all but the last coordinate of a Lasso estimator for the regression of Y on (X^j, X^{-j}) (i.e., an estimator for γ_0). Denote by $\widehat{\beta}_n^{\mathrm{DML},j}$ the DML estimator defined in Definition 4.4 where $\widehat{m}_n = \widehat{\alpha}_n^j(\cdot)$ and $\widehat{g}_n = \widehat{\gamma}_n^j(\cdot)$.

Given that the Lasso estimators achieve the fast rate we can use the asymptotic normality given in Theorem 4.5 to construct confidence intervals. Formally, we get the following corollary.

Corollary 4.6 *(Confidence intervals for sparse high-dimensional linear models).* Assume we are given n i.i.d. samples $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$ from the high-dimensional sparse linear model (4.1.15) with U and V bounded random variables and $\mathbb{E}[VV^\top]$ strictly positive definite. Fix $j \in \{1, \dots, p\}$ and assume that $\widehat{\alpha}_n^j$ and $\widehat{\gamma}_n^j$ achieve the Lasso fast rate (3.4.1), then it holds for all $q \in (0, 1)$ that

$$\lim_{n\to\infty} \mathbb{P}\left(\beta_0^j \in \left[\widehat{\beta}_n^{\mathrm{DML},j} - \tfrac{\sigma_j}{\sqrt{n}}\Phi^{-1}(1-\tfrac{q}{2}),\ \widehat{\beta}_n^{\mathrm{DML},j} + \tfrac{\sigma_j}{\sqrt{n}}\Phi^{-1}(1-\tfrac{q}{2})\right]\right) = 1 - q, \tag{4.1.17}$$

where $\sigma_j^2 = \mathbb{E}[(V^j)^2]^{-2}\mathbb{E}[(V^j)^2U^2]$ and Φ^{-1} is the quantile function of the standard normal distribution.

Proof. Fix $j \in \{1, \dots, p\}$ such that $\widehat{\alpha}_n^j$ and $\widehat{\gamma}_n^j$ achieve the Lasso fast rate. We now want to apply Theorem 4.5. To do so, we need to check (1) that all moment assumptions are satisfied and (2) the regression estimators $\widehat{\alpha}_n^j$ and $\widehat{\gamma}_n^j$ converge sufficiently fast.

- For (1), observe that since we assumed that U and V are bounded it immediately holds that $\mathbb{E}[U^2(V^j)^2] < \infty$ and $\mathbb{E}[U^2|X^{-j}]$ and $\mathbb{E}[(V^j)^2|X^{-j}]$ are bounded. Furthermore, by $X = (\mathrm{I} - B_0)^{-1}V$ we get that X is also bounded and hence $\mathbb{E}[(X^j)^2] < \infty$. Lastly, since $E[VV^\top]$ is strictly positive definite, we get that $\mathbb{E}[(V^j)^2] = e_j^\top E[VV^\top]e_j^\top > 0$.

- For (2), we use that the $\widehat{\alpha}_n^j$ and $\widehat{\gamma}_n^j$ achieve the fast rate, i.e., with high-probability it holds for a constant $C > 0$ that

  $$\tfrac{1}{n}\|X^{-j}(\alpha_0 - \widehat{\alpha}_n^j)\|_2^2 \le Cs_0^j\tfrac{\log(p)}{n}, \quad \text{and} \quad \tfrac{1}{n}\|X^{-j}(\gamma_0 - \widehat{\gamma}_n^j)\|_2^2 \le Cs_0^j\tfrac{\log(p)}{n}, \tag{4.1.18}$$

  as $n \to \infty$. Since X is assumed to be bounded, this implies together with the dominated convergence theorem that

  $$\mathbb{E}[(\alpha_0^\top X_i^{-j} - (\widehat{\alpha}_n^j)^\top X_i^{-j})^2] = o(\tfrac{1}{n}) \quad \text{and} \quad \mathbb{E}[(\gamma_0^\top X_i^{-j} - (\widehat{\gamma}_n^j)^\top X_i^{-j})^2] = o(\tfrac{1}{n}).$$

  Hence, both estimators are L^2-consistent and we additionally get that

  $$\lim_{n\to\infty} \sqrt{n}\mathbb{E}[(\alpha_0^\top X_i^{-j} - \widehat{\alpha}_n X_i^{-j})^2]^{\frac{1}{2}}\mathbb{E}[(\gamma_0^\top X_i^{-j} - \widehat{\gamma}_n X_i^{-j})^2]^{\frac{1}{2}} = 0. \tag{4.1.19}$$

Therefore, all assumptions of Theorem 4.5 are satisfied and it holds that

$$\sqrt{n}\left(\widehat{\beta}_n^{\mathrm{DML},j} - \beta_0^j\right) \xrightarrow{d} \mathcal{N}(0, \sigma_j^2) \quad \text{as } n \to \infty,$$

for $\sigma_j \coloneqq \sqrt{\mathbb{E}[(V^j)^2U^2]}/\mathbb{E}[(V^j)^2]$. Using the definition of convergence in distribution implies (4.1.17), which completes the proof of Corollary 4.6. □

Remark 4.7 (High-dimensional regime with growing p). *You might have noticed that we kept p fixed in the statement and proof of Corollary 4.6. In the asymptotic regime where p is fixed and only n goes to infinity, it can be shown that the OLS adjustment estimator (i.e., regressing Y on X and picking the j-th coordinate) also satisfies* (4.1.17). *To analyze high-dimensional estimators*

one therefore generally considers an asymptotic regime in which $p = p_n$ is allowed to grow with n. For example, one can consider the asymptotic regime in which

$$\lim_{n\to\infty} \frac{\log(p_n)}{\sqrt{n}} = 0.$$

Given the Lasso fast rate (4.1.18)*, we then still have sufficiently fast convergence to ensure* (4.1.19)*.*

This intuitive argument can be made precise, but this requires modifying Theorem 4.5 to allow for the distributions to change with n. More specifically, one needs to look at triangular sampling schemes where for each n there is a new sample $X_1^{(n)}, \dots, X_n^{(n)}$ of size n.

Remark 4.8 (Bounded noise is not necessary)**.** *In Corollary 4.6 we assumed that U and V are bounded. This is not necessary and can be weakened. We mainly use the assumption to be able to convert convergence in probability into L^2-convergence. However, the L^2-convergence assumptions on the machine learning estimators in Theorem 4.5 are in fact also not strictly necessary. You can verify in the proof that the same statements hold with the weaker notion of convergence in probability.*

4.2 Beyond the partially linear model

The DML framework applies beyond the partially linear model and can be phrased in terms of more general semi-parametric inference problems. To get an idea of the generality, consider the following setting. Let $\mathcal{Z}$ be the observation space and Z a random variable on $\mathcal{Z}$ and assume we observe n i.i.d. copies $Z_1, \dots, Z_n$ of Z. Furthermore, we use the following notational convention.

- *Parameter space:* Let $\Theta \times \Lambda$ be the parameter space with $\theta \in \Theta$ the parameter of interest and $\eta \in \Lambda$ a nuisance parameter. Furthermore, assume that $\Theta \subseteq \mathbb{R}^p$ is an open and convex subset.
- *Stochastic model:* For all $(\theta, \eta) \in \Theta \times \Lambda$ there exists a distribution $\mathbb{P}_{\theta,\eta}$ on $\mathcal{Z}$ and there exists a fixed parameter $(\theta_0, \eta_0) \in \Theta \times \Lambda$ such that $Z \sim \mathbb{P}_{\theta_0,\eta_0}$.
- *Score function:* Let $\Psi : \mathcal{Z} \times \Theta \times \Lambda \to \mathbb{R}^p$ be a measurable function which satisfies

$$\mathbb{E}[\Psi(Z, \theta, \eta_0)] = 0 \quad \Leftrightarrow \quad \theta = \theta_0.$$

Our goal is to estimate θ_0 with an asymptotically normal estimator. Similar to the partially linear model, we will estimate the nuisance parameter using some machine learning method and then adjust for the regularization and overfitting bias. For this to work, we need to choose the correct score function which we then use to estimate θ_0. Specifically, we need this score function to satisfy what is called Neyman orthogonality. Intuitively, this means that the score function should be insensitive to small changes in the nuisance parameter. Formally, we require the following property.

Definition 4.9 *(Neyman orthogonality)***.**
A score function $\Psi : \mathcal{Z} \times \Theta \times \Lambda \to \mathbb{R}^p$ is said to satisfy the *Neyman orthogonality condition* at the true parameter (θ_0, η_0) with respect to a set $\widetilde{\Lambda} \subseteq \Lambda$ if $\eta \mapsto \mathbb{E}[\Psi(Z, \theta_0, \eta)]$ is Gateaux differentiable at η_0 and it holds for all $\eta \in \widetilde{\Lambda}$ that

$$\tfrac{\mathrm{d}}{\mathrm{d}r}\mathbb{E}[\Psi(Z, \theta_0, \eta_0 + r(\eta - \eta_0)]\Big|_{r=0} = 0.$$

The score function used in the DML estimator in Definition 4.4 for the partially linear model is given by

$$\Psi((D, X, Y), \theta, (m, g)) = (D - m(X))(Y - D\theta - g(X)).$$

It can be shown (exercise) that this score function is indeed Neyman orthogonal at the true parameters θ_0, m_0, g_0. The DML estimator for the general semi-parametric case with a fixed score function Ψ generalizes Definition 4.4 in the following way.

Definition 4.10 *(General DML estimator).*
Let $\widehat{\eta}_n$ be machine learning estimators for the nuisance parameter η_0. For all $n \in \mathbb{N}$, split the observation indices into K disjoint sets $I_1, \dots, I_K$ such that $I_1 \cup \dots \cup I_K = \{1, \dots, n\}$ and $|I_k| \approx n/K$. Furthermore, for all $n \in \mathbb{N}$ and all $k \in \{1, \dots, K\}$ define

$$\widehat{\eta}_n^k \coloneqq \widehat{\eta}_{|I_k^c|}((Z_i)_{i \in I_k^c})$$

where $I_k^c = \{1, \dots, n\} \setminus I_k$. Then, for $k \in \{1, \dots, K\}$, let $\widehat{\theta}_n^k$ be the solution to

$$\frac{1}{|I_k|} \sum_{i \in I_k} \Psi(Z_i, (\widehat{\theta}_n^k, \widehat{\eta}_k)) = 0. \tag{4.2.1}$$

Then the DML estimator $\widehat{\theta}_n^{\mathrm{DML}}$ is given by

$$\widehat{\theta}_n^{\mathrm{DML}} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\theta}_n^k.$$

Under some additional assumptions on the score function and the machine learning procedures this procedure will result in an asymptotically normal estimator. Most importantly, we require that the machine learning procedures converge sufficiently fast and that the score function is Neyman orthogonal. The formal result requires quite a few regularity conditions and goes beyond the scope of this course.

Bibliography

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

S. Huang, E. Ailer, N. Kilbertus, and N. Pfister. Supervised learning and model analysis with compositional data. *arXiv preprint arXiv:2205.07271*, 2022.

D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

A. R. Lundborg, R. D. Shah, and J. Peters. Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5):1821–1850, 2022.

J. Pearl. *Causality*. Cambridge university press, 2009.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

Appendix A

Background

A.1 Singular value decomposition

Let $X \in \mathbb{R}^{n\times p}$ and $m \coloneqq \min(n,p)$. Then the matrix X can be decomposed into orthogonal and diagonal matrices using the singular value decomposition (SVD). The SVD comes in two flavours.

- **full SVD:**
$$X = UDV^\top,$$
where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{p\times p}$ are orthogonal and $D \in \mathbb{R}^{n\times p}$ is diagonal with $d_1 \geq \cdots \geq d_m \geq 0$, where d_j is the j-th diagonal element of D.

- **thin SVD:**
$$X = UDV^\top,$$
where $U \in \mathbb{R}^{n\times m}$, $V \in \mathbb{R}^{m\times p}$ have orthogonal columns and $D \in \mathbb{R}^{m\times m}$ is diagonal with $d_1 \geq \cdots \geq d_m \geq 0$, where d_j is the j-th diagonal element of D.

A.1.1 Connection to principle component analysis

There is a close connection between SVD and principal component analysis (PCA). To see this consider the first principal component (PC), which is given by

$$Xw^*, \quad \text{where} \quad w^* \coloneqq \underset{w\in\mathbb{R}^p:\, \|w\|_2=1}{\arg\max}\ \widehat{\mathrm{Var}}(Xw).$$

Assuming X is a mean zero random variable, the variance is estimated by $\widehat{\mathrm{Var}}(Xw) = \frac{1}{n}w^\top X^\top Xw$. Hence, using the SVD we can bound the estimated variance from above as follows

$$\begin{aligned}
\widehat{\mathrm{Var}}(Xw) &= \tfrac{1}{n}w^\top X^\top Xw \\
&= \tfrac{1}{n}w^\top VD^2V^\top w \\
&= \tfrac{1}{n}a^\top D^2 a \\
&= \tfrac{1}{n}\sum_{j=1}^{p} a_j^2 d_j^2 \\
&\leq \tfrac{1}{n}d_1^2 \sum_{j=1}^{p} a_j^2 \\
&= \tfrac{1}{n}d_1^2, \qquad \text{(A.1.1)}
\end{aligned}$$

where $a \coloneqq V^\top w$ and we used that $\|w\| = 1$ in the last step. Next, observe that $V^\top V_1 = (1, 0, \ldots, 0)^\top$ implies that

$$\widehat{\mathrm{Var}}(XV_1) = \tfrac{1}{n} d_1. \tag{A.1.2}$$

Therefore, (A.1.1) and (A.1.2) together imply that the first PC is given by

$$XV_1 = U_1 d_1.$$

A similar arguments shows that $XV_2 = U_2 d_2$, ..., $XV_p = U_p d_p$ correspond to the remaining PCs.

A.2 Computational complexity

This section provides a rough overview of what to consider when assessing the computational cost of statistical estimation procedures. In modern applications, with large amounts of data, the computational cost is often a bottleneck and computational speed-ups are crucial. One can divide the computational complexity of an algorithm into two main parts:

(1) The number of arithmetic operations the algorithms makes.

(2) The required memory to run the algorithm.

Arithmetic operations The number of arithmetic operations can generally be computed by dividing the algorithm up into basic operations (e.g., addition, multiplication, taking logarithms, etc.) and then counting how many of these operations are required. For matrix algebra the following computational costs are often helpful:

- *Matrix multiplication:* Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times p}$ be two matrices, then the number of basic operations needed to compute the product AB (without using a specialized algorithm) is of order $\mathcal{O}(nmp)$. To see this, observe that computing $(AB)_{i,j} = \sum_{k=1}^{m} A_{ik} B_{kj}$ requires $2m$ computations (m additions and m multiplications) and hence the full matrix can be computed in $2mnp$ basic operations.
- *Matrix inversion:* Let $A \in \mathbb{R}^{p \times p}$ be an invertible matrix, then the inverse A^{-1} can be computed in $\mathcal{O}(p^3)$ operations using Gauss-eliminations.
- *Singular value decomposition:* Let $X \in \mathbb{R}^{n \times p}$ be a matrix, then computing the SVD $X = UDV^\top$ requires $\mathcal{O}(npm)$ operations using the QR-algorithm, where $m = \min(n, p)$.

For certain operations faster algorithms exist, for example two square matrices with dimension p can be multiplied with $\mathcal{O}(p^{2.807})$ operations using the Strassen algorithm, but for our purposes the above computational costs are sufficiently accurate to provide an approximation of the runtime of a baseline implementation.

Memory The memory usage can be approximated by counting all the doubles (or floats, depending on which precision is used) that need to be saved while the algorithm is executed. For a matrix $X \in \mathbb{R}^{n \times p}$ this means that we need to save np doubles or floats. One double uses 64 bits which corresponds to 8 bytes, while one float uses 32 bits which corresponds to 4 bytes. The memory usage in bytes is therefore given by

$$\#\text{doubles} \cdot 8\text{bytes} + \#\text{floats} \cdot 4\text{bytes}.$$