Designing and analyzing new early stopping rules for saving computational resources

Yaroslav Averyanov

L'UNIVERSITÉ DE LILLE
COMUE LILLE NORD DE FRANCE

ÉCOLE DOCTORALE RÉGIONALE SPI 72

Spécialité : Statistique

THÈSE DE DOCTORAT

Par

Yaroslav AVERYANOV

"Concevoir et analyser de nouvelles règles d'arrêt prématuré pour économiser les ressources de calcul"

Thèse présentée et soutenue publiquement à Lille, le 15/12/2020
Unité de recherche : équipe-projet MODAL, Inria Lille-Nord Europe

Rapporteurs avant soutenance :

Monsieur Patrice BERTAIL	Professeur des universités, Université Paris Nanterre
Monsieur Julien CHIQUET	Directeur de recherche, INRAE Paris

Composition du Jury :

Président du jury :	Madame Cristina BUTUCEA	Professeur des universités, ENSAE Paris
Examinateur :	Monsieur Julien MAIRAL	Directeur de recherche, Inria Grenoble Rhône-Alpes
Dir. de thèse :	Monsieur Cristian PREDA	Professeur des universités, Université de Lille
Co-dir. de thèse :	Monsieur Alain CELISSE	Professeur des universités, Université Paris 1

Abstract

This work develops and analyzes strategies for constructing instances of the so-called *early stopping rules* applied to some iterative learning algorithms for estimating the regression function. Such quantities are data-driven rules indicating when to stop the iterative learning process to reach a trade-off between computational costs and the statistical precision. Unlike a large part of the existing literature on early stopping, where these rules only depend on the data in a "weak manner", we provide *data-driven* solutions for the aforementioned problem without utilizing validation data.

The crucial idea exploited here is that of the minimum discrepancy principle (MDP), which shows when to stop an iterative learning algorithm. To the best of our knowledge, this idea dates back to the work of Vladimir A. Morozov in the 1960s-1970s who studied linear ill-posed problems and their regularization, mostly inspired by mathematical physics problems. Among different applications of this line of work, the so-called *spectral filter estimators* such as spectral cut-off, Landweber iterations, and Tikhonov (ridge) regularization have received quite a lot of attention (e.g., in statistical inverse problems). It is worth mentioning that the minimum discrepancy principle consists in controlling the residuals of an estimator (which are iteratively minimized) and properly setting a threshold for them such that one can achieve some (minimax) optimality.

The first part of this thesis is dedicated to theoretical guarantees of stopping rules based on the minimum discrepancy principle and applied to gradient descent, and Tikhonov (ridge) regression in the framework of reproducing kernel Hilbert space (RKHS). There, we show that this principle provides a minimax optimal functional estimator of the regression function when the rank of the kernel is finite. However, when one deals with infinite-rank reproducing kernels, the resulting estimator will be only suboptimal. While looking for a solution, we found the existence of the so-called *residuals polynomial smoothing* strategy. This strategy (combined with MDP) has been proved to be optimal for the spectral cut-off estimator in the linear Gaussian sequence model. We borrow this strategy, modify the stopping rule accordingly, and prove that the *smoothed minimum discrepancy principle* yields a minimax optimal functional estimator over a range of function spaces, which includes the well-known Sobolev function class.

Our second contribution consists in exploring the theoretical properties of the minimum discrepancy stopping rule applied to the more general family of *linear estimators*. The main difficulty of this approach is that, unlike the spectral filter estimators considered earlier, linear estimators do no longer lead to monotonic quantities (the bias and variance terms). Let us mention that this is also the case for famous algorithms such as Stochastic Gradient Descent. Motivated by further practical applications, we work with the widely used k-NN regression estimator as a reliable first example. We prove that

the aforementioned stopping rule leads to a minimax optimal functional estimator, in particular, over the class of Lipschitz functions on a bounded domain.

The third contribution consists in illustrating through empirical experiments that for choosing the tuning parameter in a linear estimator (the k-NN regression, Nadaraya-Watson, and variable selection estimators), the MDP-based early stopping rule performs comparably well with respect to other widely used and known model selection criteria.

Résumé

Ce travail développe et analyse des stratégies pour construire des instances de ce que l'on appelle les *règles d'arrêt prématurés* appliquées à certains algorithmes d'apprentissage itératif pour estimer la fonction de régression. Ces quantités sont des règles "data-driven" indiquant quand arrêter le processus d'apprentissage itératif pour parvenir à un compromis entre les coûts de calcul et la précision statistique. Contrairement à une grande partie de la littérature existante sur l'arrêt prématuré, où ces règles ne dépendent que des données de manière "faible", nous fournissons des solutions *data-driven* pour le problème susmentionné sans utiliser les données de validation.

L'idée cruciale exploitée ici est celle du principe d'écart minimal (MDP), qui montre où arrêter un algorithme d'apprentissage itératif. À notre connaissance, cette idée remonte aux travaux de Vladimir A. Morozov dans les années 1960-1970 qui a étudié des problèmes linéaires mal posés et leur régularisation, principalement inspirés par des problèmes de physique mathématique. Parmi les différentes applications de cette ligne de travail, les soi-disant *estimateurs de filtre spectral* tels que le "spectral cut-off", les itérations de Landweber, et la régularisation de Tikhonov (ridge) ont reçu beaucoup d'attention (par exemple, dans des problèmes statistiques inverses). Il est à noter que le principe d'écart minimal consiste à contrôler les résidus d'un estimateur (qui sont minimisés de manière itérative) et à leur fixer correctement un seuil tel que l'on puisse atteindre une certaine optimalité (minimax).

La première partie de cette thèse est consacrée aux garanties théoriques des règles d'arrêt basées sur le principe d'écart minimal et appliquées à la descente de gradient, et à la régression de Tikhonov (ridge) dans le cadre de l'espace de Hilbert à noyau reproduisant (RKHS). Là, nous montrons que ce principe fournit un estimateur fonctionnel optimal minimax de la fonction de régression lorsque le rang du noyau est fini. Cependant, quand nous traitons des noyaux reproduisants de rang infini, l'estimateur résultant sera seulement sous-optimal. En recherchant une solution, nous avons trouvé l'existence de la stratégie dite de *lissage polynomial des résidus*. Cette stratégie (combinée avec le MDP) s'est avérée optimale pour l'estimateur "spectral cut-off" dans le modèle de séquence gaussienne linéaire. Nous empruntons cette stratégie, modifions la règle d'arrêt en conséquence, et prouvons que le *principe d'écart minimal lissé* produira un estimateur fonctionnel optimal minimax sur une gamme d'espaces de fonctions, qui comprend la classe de fonctions Sobolev bien connue.

Notre deuxième contribution consiste à explorer des propriétés théoriques de la règle d'arrêt d'écart minimal appliquée à la famille plus générale des *estimateurs linéaires*. La principale difficulté de cette approche est que, contrairement aux estimateurs de filtre spectral considérés précédemment, les estimateurs linéaires ne conduisent plus à des quantités monotones (les biais et variance). Mentionnons que c'est également le cas des algorithmes célèbres tels que la descente de gradient stochastique.

Motivés par d'autres applications pratiques, nous travaillons avec l'estimateur de régression des k plus proches voisins largement utilisé, comme un premier exemple fiable. Nous montrons que la règle d'arrêt susmentionnée conduit à un estimateur fonctionnel optimal minimax, en particulier sur la classe des fonctions de Lipschitz sur un domaine borné.

La troisième contribution consiste à illustrer au moyen de simulations empiriques que, pour le choix du paramètre de réglage dans un estimateur linéaire (la méthode des k plus proches voisins, la régression de Nadaraya-Watson, et l'estimateur de sélection de variables), la règle d'arrêt prématuré basée sur le MDP se comporte comparativement bien par rapport à d'autres critères de sélection de modèles, largement utilisés et connus.

Table of Contents

INTRODUCTION

Nonparametric regression estimation has become a crucial question nowadays. Data are becoming more and more bulky and even more complex. For this reason, making parametric assumptions (e.g., Gaussian or Laplace) on the data generating process seems unrealistic and does no longer represent a reliable alternative. Furthermore, the resulting theoretical guarantees often strongly depend on these distributional assumptions, which somewhat restricts their validity domain.

In the present thesis, the main focus is given to the nonparametric estimation of the regression function through iterative learning algorithms. Iterative algorithms have become ubiquitous, for instance, in situations when some regularization is needed, or no closed-form expressions are available for the estimator. In practice, such iterative algorithms require the knowledge of the best iteration number at which one should interrupt the learning process. Our final goal is designing and analyzing a so-called *early stopping rule*, which aims at avoiding useless computations by interrupting the learning process "not too late" while outputting an almost optimal estimator of the regression function.

Early stopping can be seen as an (implicit) regularization strategy, which consists in stopping the learning process before "the convergence" (when applicable). For instance, it is likely the most commonly used strategy in (deep) neural network learning. This method is popular due to its effectiveness and simplicity. In the present document, we explore the theoretical (statistical) properties of the aforementioned method in two learning frameworks: with iterative spectral filter algorithms in reproducing kernel Hilbert space and for choosing the tuning parameter in linear estimators.

The content of the document is as follows.

Chapter 1: We introduce some basic mathematical concepts that will be essential for this work. There, we successively introduce nonparametric regression, the reproducing kernel Hilbert space (RKHS), early stopping rules for iterative learning algorithms, and linear estimators, among others.

Chapter 2: As a first contribution of the thesis, we present an early stopping rule for gradient descent and kernel ridge regression (cast as an iterative algorithm) in the framework of reproducing kernel Hilbert space. To be precise, we consider a stopping rule based on the minimum discrepancy principle (MDP) that was initially developed for solving ill-posed inverse problems. The main quantity MDP relies on is the empirical risk (the residuals) that is minimized and monitored throughout the iterative learning process. This stopping rule is proven to yield a minimax-rate optimal estimator of the regression function with finite-rank kernels. It appears that controlling the behavior of the empirical risk around its expectation is crucial for the analysis of the statistical optimality. Since with

infinite-rank kernels, this control is not tight enough, we develop a new stopping rule for this type of reproducing kernels. It relies on the so-called polynomial smoothing of the residuals, which allows reducing the variability of the empirical risk around its expectation. This stopping rule is called the smoothed discrepancy principle stopping rule. Combining all the assumptions on the eigenvalues of the normalized kernel matrix, we prove the minimax rate optimality of the resulting estimator based on the smoothed discrepancy principle stopping rule. It holds true for Sobolev smoothness classes, in particular.

Chapter 3: The second contribution of the thesis consists in applying the minimum discrepancy principle (MDP) to the more general class of linear estimators for choosing the tuning parameter. In this chapter, we choose to work with the well-known k-nearest neighbor regression estimator as a starting point. We prove that the MDP-based estimator achieves minimax optimality under the assumption that the regression function is bounded. This holds, in particular, over the class of Lipschitz functions on a bounded domain. The main goal of applying MDP for choosing k is lowering the computational time of the model selection procedure.

Chapter 4: The present chapter yields an extensive simulation study for the performance of the MDP stopping rule for the parameter tuning with several linear estimators such as the k-NN, Nadaraya-Watson regression estimators, and the variable selection (projection-based) estimator. These experiments have been carried out utilizing synthetic as well as real datasets. According to the results collected here, we can conclude that the MDP-based stopping rule performs comparably well to other considered model selection criteria such as cross-validation, Mallows' C_p,... while saving computational resources, unlike the former examples.

Conclusion and perspectives: In this chapter, we briefly summarize the main contributions and extensively discuss future research directions.

Let us finally emphasize that the content of this thesis (namely, Chapters 2–4) is based on the submissions we list below:

— Chapter 2 is based on the work "Early stopping and polynomial smoothing in regression with reproducing kernels", Y.Averyanov and A.Celisse, submitted to the Electronic Journal of Statistics.

— Chapter 3 and 4 are based on the work "Minimum discrepancy principle strategy for choosing k in k-NN regression", Y.Averyanov and A.Celisse, submitted to the Statistica Sinica.

Chapter 1

Overview

The main purpose of this chapter is to introduce the framework as well as leading notions the present work is based on. More precisely, we successively introduce the nonparametric regression context in Section 1.1, the reproducing kernel Hilbert spaces in Section 1.2 (which turns out to be crucial in our theoretical analysis of Chapter 2), the model selection problem in Section 1.3, and the main literature dedicated to the design of early stopping rules with iterative learning algorithms in Section 1.5. Finally, one can find some aspects of the parameter selection (via model selection) with linear estimators in Section 1.6.

1.1 Nonparametric regression

1.1.1 Formulation

A regression problem is defined by a set of covariate $X \in \mathcal{X}$, and a response $Y \in \mathcal{Y}$. In this work, we focus on the case of real-valued responses $\mathcal{Y} = \mathbb{R}$. The goal of the regression is to estimate a function $f : \mathcal{X} \mapsto \mathcal{Y}$ such that an error $Y - f(X)$ is as small as possible. Assume that both X and Y are random variables, and $p(x, y)$ is their joint probability distribution, then it is reasonable to measure the error in terms of the *mean-squared error* (MSE) as

$$\mathcal{L}(f) \coloneqq \mathbb{E}\left[(Y - f(X))^2\right],$$

where $\mathbb{E}$ denotes the expectation with respect to the joint $p(x, y)$.

The function f^*, which minimizes the criterion $\mathcal{L}(f)$, is called the *regression function* [67, Section 1.1] and defined as

$$f^*(x) \coloneqq \mathbb{E}\left[Y \mid X = x\right].$$

The goal of the regression problem is basically the same as constructing an estimator $\widehat{f}$ of f^* from i.i.d. samples $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R},\ i = 1, \ldots, n$. Note that for these samples, one can write

$$Y_i = f^*(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where ε_i are i.i.d. zero-mean random variables, meaning that $\mathbb{E}[\varepsilon_i \mid X_i] = 0,\ i = 1, \ldots, n$, usually (sub-)Gaussian $\mathcal{N}(0, \sigma^2)$ [110, Proposition 2.5.2], where $\sigma > 0$ is a standard deviation parameter. This

setting is called the *random design setting*.

Another common setup is called the *fixed design setting*, where in Eq. (1.1) $X_i,\ i = 1, \dots, n$, are fixed inputs, thus randomness comes only from the noise $\{\varepsilon_i\}_{i=1}^n$. In this setting, since the covariates are not random, we write them as $\{x_i\}_{i=1}^n$ (unless explicitly stated).

It is worth mentioning that in the present nonparametric regression context, we do not assume any particular parametric form of f^*, neither for the noise $\{\varepsilon_i\}_{i=1}^n$. However, several parameterizations have been thoroughly studied. For instance, the linear regression model arises by setting $f^*(x) = \langle x, \theta \rangle$, where $x \in \mathcal{X} \subseteq \mathbb{R}^d$, and $\theta \in \mathbb{R}^d$ is a parameter to estimate. Another possibility is the estimation of an additive/sparse regression function when one assumes that $f^*(x) = \sum_{j=1}^p \beta_j f_j(x)$, where $\beta_j \in \mathbb{R}$ is some coefficient, and $\{f_j\}_{j=1}^p$ are some (basis) functions (see, for instance, [38, 115]).

1.1.2 Quality measure of an estimator

In this section, we explain how to quantify the performance of a functional estimator $\widehat{f}$ of the regression function f^* from Eq. (1.1) (see, e.g., the monographs [108, 114] for more details).

Let us define an *empirical norm* in terms of the training points $X_i,\ i = 1, \dots, n$, acting on functions $f : \mathcal{X} \to \mathbb{R}$, by

$$\|f\|_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^n f^2(X_i). \tag{1.2}$$

Notice that in the fixed design case, this quantity is deterministic. In this case, the norm in Eq. (1.2) is denoted as $L_2(\mathbb{P}_n)$.

When the covariates $\{X_i\}_{i=1}^n$ are random, denoting the probability distribution of X as $\mathbb{P}_X$, we define the L_2 norm (in terms of $\mathbb{P}_X$), acting on functions $f : \mathcal{X} \to \mathbb{R}$, by

$$\|f\|_2^2 \coloneqq \|f\|_{L_2(\mathbb{P}_X)} = \int_{\mathcal{X}} f^2(x) d\mathbb{P}_X(x). \tag{1.3}$$

In addition to that, we denote (with a slight abuse of notation) the functional space $L_2(\mathbb{P}_X) \coloneqq \{f \mid \|f\|_2^2 < +\infty\}$ and the inner product in $L_2(\mathbb{P}_X)$ as $\langle f, g \rangle_{L_2(\mathbb{P}_X)} \coloneqq \int_{\mathcal{X}} f(x) g(x) d\mathbb{P}_X$ for any $f, g \in L_2(\mathbb{P}_X)$.

A quantity of interest will be the error of an estimator $\widehat{f}$ of f^*, which can be measured in terms of:

$$\|\widehat{f} - f^*\|_n^2 \quad \text{or} \quad \|\widehat{f} - f^*\|_2^2.$$

In this work, the expectation of either of these two errors will be called the *risk (prediction) error*.

Another choice to overcome the obstacle that f^* is unknown could be making a (relatively mild) assumption that it belongs to some (quite rich) functional space. Introducing an example of this space is the purpose of the next section.

1.2 Reproducing kernel Hilbert space

We now turn to the notion of a reproducing kernel Hilbert space [11], or RKHS for short. These spaces are particular instances of functional spaces that act from $\mathcal{X}$ to $\mathbb{R}$. We start to describe this notion by defining another notion of a positive semidefinite kernel function [99, 112]. After that, RKHS can be constructed based on this kernel function.

1.2.1 Positive semidefinite kernel functions

Let us begin with the notion of a positive semidefinite kernel function (reproducing kernel) [98, 112, 114].

Definition 1.2.1. A symmetric bivariate function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive semidefinite kernel if for all integers $n \geq 1$ and elements $\{x_i \in \mathcal{X}\}_{i=1}^n$, the $n \times n$ matrix with elements $K_{ij} = \mathbb{K}(x_i, x_j)$ is positive semidefinite, meaning that one has

$$\alpha^\top K \alpha \geq 0 \qquad \text{for all } \alpha \in \mathbb{R}^n.$$

Let us mention some well-known examples of positive semidefinite kernels [98, 112, 114].

Example 1 (Linear kernel)**.** *Assume that $\mathcal{X} = \mathbb{R}^d$ and define the linear kernel function as $\mathbb{K}(w, z) = \langle w, z \rangle_{\mathbb{R}^d} = \sum_{i=1}^d w_i z_i$. For any $\{x_i\}_{i=1}^n$ of arbitrary points from $\mathcal{X}$, define the matrix $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = \mathbb{K}(x_i, x_j),\ i, j \in \{1, \ldots, n\}$. Then for any vector $\alpha \in \mathbb{R}^n$,*

$$\alpha^\top K \alpha = \sum_{i,j=1}^n \alpha_i \alpha_j \langle x_i, x_j \rangle_{\mathbb{R}^d} = \| \sum_{i=1}^n \alpha_i x_i \|^2 \geq 0.$$

Thus, the linear kernel is positive semidefinite.

Example 2 (Polynomial kernel)**.** *Assume that $\mathcal{X} = \mathbb{R}^d$ and define the polynomial kernel function as $\mathbb{K}(x, z) = \langle x, z \rangle_{\mathbb{R}^d}^m$ for some natural number $m \geq 2$. Positive semi-definiteness (for $m = 2$) follows from*

$$\mathbb{K}(x, z) = \left(\sum_{i=1}^d x_i z_i \right)^2 = \sum_{i=1}^d x_i^2 z_i^2 + 2 \sum_{i<j} x_i x_j z_i z_j.$$

Set $D = d + \binom{d}{2}$ and define a map $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ as

$$\Phi(x) = \begin{cases} x_j^2, & \text{for } j \in \{1, \ldots, d\}, \\ \sqrt{2} x_i x_j, & \text{for } i < j. \end{cases}$$

Then, one can verify that $\mathbb{K}(x, z) = \langle \Phi(x), \Phi(z) \rangle_{\mathbb{R}^D}$. Finally, recall Example 1.

Example 3 (Gaussian kernel)**.** *Assume that* $\mathcal{X} \subseteq \mathbb{R}^d$ *and consider the Gaussian kernel* $\mathbb{K}(x,z) = \exp\left(-\frac{1}{2h^2}\|x-z\|^2\right)$ *for some parameter* $h > 0$*. Importantly, the Gaussian kernel is a very popular choice in practice [9, 98, 101].*

1.2.2 Mercer's theorem and consequences

Let us now construct an RKHS from a kernel function, as we claimed at the beginning of the section. More precisely, we turn to a useful representation of a broad class of positive semidefinite kernel functions in terms of their eigenfunctions. Given a symmetric positive semidefinite kernel function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ that is continuous (in both arguments), we can define a linear operator $T_k : L_2(\mathbb{P}_X) \mapsto L_2(\mathbb{P}_X)$ via

$$T_k(f)(z) \coloneqq \int_{\mathcal{X}} \mathbb{K}(x,z) f(z) d\mathbb{P}_X(x). \tag{1.4}$$

Assume that T_k is bounded on $L_2(\mathbb{P}_X)$, then we can say that T_k is a *Hilbert-Schmidt operator*. For instance, the boundness of T_k can be achieved by assuming

$$\int_{\mathcal{X}\times\mathcal{X}} \mathbb{K}^2(x,z) d\mathbb{P}_X(x) d\mathbb{P}_X(z) < +\infty. \tag{1.5}$$

It follows from

$$\begin{aligned}\|T_k(f)\|^2_{L_2(\mathbb{P}_X)} &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \mathbb{K}(x,y) f(x) d\mathbb{P}_X(x)\right)^2 d\mathbb{P}_X(y) \\ &\leq \|f\|^2_{L_2(\mathbb{P}_X)} \int_{\mathcal{X}\times\mathcal{X}} \mathbb{K}^2(x,y) d\mathbb{P}_X(x) d\mathbb{P}_X(y),\end{aligned}$$

where we used the Cauchy-Schwarz inequality.

Having gained some intuition about the kernel integral operator, we are ready to state Mercer's theorem.

Theorem 1.2.1 (Mercer's theorem; see Theorem 12.20 in [114])**.** *Suppose that* $\mathcal{X}$ *is compact, the kernel function* $\mathbb{K}$ *is continuous and positive semidefinite and satisfies Ineq. (1.5). Then, there exists a sequence of eigenfunctions* $\{\phi_j\}_{j=1}^{+\infty}$ *that forms an orthonormal basis of* $L_2(\mathbb{P}_X)$*, and non-negative eigenvalues* $\{\mu_j\}_{j=1}^{+\infty}$ *such that*

$$T_k(\phi_j) = \mu_j \phi_j, \qquad \text{for } j = 1, 2, \ldots \tag{1.6}$$

Moreover, the following expansion holds

$$\mathbb{K}(x,z) = \sum_{j=1}^{+\infty} \mu_j \phi_j(x) \phi_j(z), \tag{1.7}$$

where the convergence of the series holds absolutely and uniformly.

Given Mercer's kernel (1.7) and its associated eigenvalues $\{\mu_j\}_{j=1}^{+\infty}$, one can distinguish two cases: a) finite-rank kernels ($\mu_j = 0\ \forall j > r$, for some integer $r > 1$); b) infinite-rank kernels ($\mu_1 \geq \mu_2 \geq \ldots > 0$). We list some examples of these kernels below.

— *Finite-rank kernels*: examples of such kernels include the linear kernel $\mathbb{K}(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$, which has rank at most $r = d$; and the kernel $\mathbb{K}(x, x') = (1 + xx')^m$ generating polynomials of degree m, which has rank at most $r = m + 1$.

— *Infinite-rank kernels*:

— polynomial eigenvalue decay kernels:

$$cj^{-\beta} \leq \mu_j \leq Cj^{-\beta}, \qquad \text{for all } j = 1, 2, \ldots \tag{1.8}$$

where $0 < c \leq C$ are universal constants, and $\beta > 1$ parametrizes the decay rate. We note that Eq. (1.8) assumes a trace class operator $\mathrm{tr}(T_k) = \sum_{j=1}^{+\infty} \mu_j < +\infty$. Kernels with polynomial decaying eigenvalues include those that underlie the Sobolev spaces with different orders of smoothness (see, e.g., [66, 114]) that consist of functions that have weak derivatives being Lebesgue integrable. More formally, for some fixed integer $\alpha \geq 1$, consider the class $\mathbb{H}^\alpha[0, 1]$ of real-valued functions on $[0, 1]$ that are α-times differentiable, with the α-derivative $f^{(\alpha)}$ being Lebesgue-integrable, and such that $f(0) = f^{(1)}(0) = \ldots = f^{(\alpha-1)}(0) = 0$. Then, we may define an inner product

$$\langle f, g \rangle_{\mathbb{H}^\alpha} \coloneqq \int_0^1 f^{(\alpha)}(z) g^{(\alpha)}(z) dz, \quad \forall f, g \in \mathbb{H}^\alpha[0, 1]. \tag{1.9}$$

This inner product defines an RKHS and the reproducing kernel

$$\mathbb{K}(x, z) = \int_0^1 \frac{(x - y)_+^{\alpha-1}}{(\alpha - 1)!} \frac{(z - y)_+^{\alpha-1}}{(\alpha - 1)!} dy, \tag{1.10}$$

where $(t)_+ = \max\{0, t\}$. As an example, the first-order Sobolev kernel $\mathbb{K}(x, z) = \min\{x, z\}$, $x, z \in [0, 1]$, generates an RKHS of Lipschitz functions (functions with a bounded derivative) and gives $\beta = 2$. Higher-order Sobolev kernels exhibit the polynomial eigendecay condition (1.8) with larger values of the parameter β.

— exponential eigenvalue decay kernels:

$$c\exp\left(-c_1 j^2\right) \leq \mu_j \leq C\exp\left(-c_1 j^2\right), \qquad \text{for all } j = 1, 2, \ldots, \tag{1.11}$$

for strictly positive constants (c_1, c, C). Such classes include the RKHS generated by the Gaussian kernel $\mathbb{K}(x, x') = \exp\left(-\|x - x'\|^2\right)$.

An interesting consequence of Mercer's theorem 1.2.1 is in giving a relatively explicit characterization of the RKHS associated with a kernel function.

Corollary 1.2.2 (Corollary 12.26 in [114]). *Consider a kernel function satisfying the conditions of Mercer's theorem with its associated eigenfunctions* $\{\phi_j\}_{j=1}^{+\infty}$ *and non-negative eigenvalues* $\{\mu_j\}_{j=1}^{+\infty}$. *Then, it induces the following reproducing kernel Hilbert space*

$$\mathcal{H} \coloneqq \left\{ f = \sum_{j=1}^{+\infty} \beta_j \phi_j \mid \text{for some } \{\beta_j\}_{j=1}^{+\infty} \in \ell^2(\mathbb{N}) \text{ with } \sum_{j=1}^{+\infty} \frac{\beta_j^2}{\mu_j} < +\infty \right\}, \tag{1.12}$$

along with the inner product:

$$\langle f, g \rangle_{\mathcal{H}} \coloneqq \sum_{j=1}^{+\infty} \frac{\langle f, \phi_j \rangle_{L_2(\mathbb{P}_X)} \langle g, \phi_j \rangle_{L_2(\mathbb{P}_X)}}{\mu_j}, \qquad \forall f, g \in \mathcal{H}. \tag{1.13}$$

Thus, the desired RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is constructed.

1.2.3 Reproducing property

An important fact about the reproducing kernel Hilbert space is the *kernel reproducing property* [98, 114], which underlies the power of kernel methods in practice, by providing them great flexibility. In particular, it says that any positive semidefinite kernel function $\mathbb{K}$, defined on the Cartesian product space $\mathcal{X} \times \mathcal{X}$, can be used to construct a Hilbert space of functions on $\mathcal{X}$ that we denote $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$. This Hilbert space is unique and has the following property: for any $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x) \in \mathcal{H}$ and satisfies the relation

$$\langle f, \mathbb{K}(\cdot, x) \rangle_{\mathcal{H}} = f(x) \qquad \text{for all } f \in \mathcal{H}. \tag{1.14}$$

Reproducing property (1.14) allows us to think of the kernel as defining a "feature map" $x \mapsto \mathbb{K}(\cdot, x) \in \mathcal{H}$, and an inner product in $\mathcal{H}$ is reduced to kernel evaluations, meaning that $\langle \mathbb{K}(\cdot, x), \mathbb{K}(\cdot, z) \rangle_{\mathcal{H}} = \mathbb{K}(x, z)$ for all $x, z \in \mathcal{X}$. We summarize what has been said in the theorem below.

Theorem 1.2.3 (Theorem 12.11 in [114]). *Given any positive semidefinite kernel function* $\mathbb{K}$, *there is a unique Hilbert space* $\mathcal{H}$ *in which the kernel satisfies the reproducing property (1.14). It is known as the reproducing kernel Hilbert space associated with* $\mathbb{K}$.

1.3 Model selection in regression

Choosing the number of iterations of an iterative learning algorithm could be seen as a model selection task. We clarify what it means below.

Designing an estimation procedure usually requires some prior knowledge of the unknown distribution of the pair covariate-response (X, Y). Without this knowledge, choosing a proper model is one of the main obstacles for the statistician. More precisely, the aim of model selection is to construct data-driven criteria to select a model among a given list. By designing such criteria, one can consider

the so-called *nonasymptotic approach*, meaning that the size of the models is allowed to depend on the sample size n. In the case of the nonparametric regression, this approach allows to choose the models (functions) with the best approximation property, from the data. The main theoretical tool that one uses in model selection is the *concentration inequality* [36]. The central feature of the concentration inequalities is the fact that they provide deviation control of a (sum of) random variable *for any sample size* n.

Let us describe the standard procedure for model selection in the framework of nonparametric regression. To do this, we introduce a collection of models $\mathcal{S} = \{S_t,\ t \in \mathcal{T}\}$, that hereafter will be called models, indexed by a countable set $\mathcal{T}$. To each $t \in \mathcal{T}$, we associate some functional estimator f^t of f^*, relative to S_t. As an example, let us consider, for instance, the empirical norm $\|f^t - f^*\|_n^2$ to quantify the quality of the estimator f^t. Besides that, we introduce the *risk error* of the estimator f^t, which will be equal to $\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2$ ($\mathbb{E}_\varepsilon$ denotes the expectation w.r.t. the noise $\{\varepsilon_i\}_{i=1}^n$). Then, the goal will be to choose an (optimal) $\widehat{t}$ with respect to $\|f^t - f^*\|_n^2$ or $\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2$. Given the model selection procedure $\widehat{t}$, one can verify its optimality by utilizing the so-called *oracle inequalities* [47, 80, 108], meaning that $\widehat{t} \in \mathcal{T}$ could satisfy one of two (or both simultaneously) inequalities below:

—
$$\|f^{\widehat{t}} - f^*\|_n^2 \leq C_n \inf_{t\in\mathcal{T}} \mathbb{E}_\varepsilon\|f^t - f^*\|_n^2 + r_n, \tag{1.15}$$

—
$$\mathbb{E}_\varepsilon\|f^{\widehat{t}} - f^*\|_n^2 \leq C_n \inf_{t\in\mathcal{T}} \mathbb{E}_\varepsilon\|f^t - f^*\|_n^2 + r_n, \tag{1.16}$$

where Ineq. (1.15) holds with high (exponential) probability, e.g., $1 - \exp(-\sqrt{n})$. In both inequalities, constant C_n should be bounded and does not depend on the regression function f^*, and, ideally, should be close to 1 (a selection procedure with $C_n \to 1$ as $n \to +\infty$ is called *asymptotically optimal* or *efficient*). Moreover, the right hand side term r_n should be negligible (smaller) compared to $\inf_{t\in\mathcal{T}} \mathbb{E}_\varepsilon\|f^t - f^*\|_n^2$. In addition to that, notice that Ineq. (1.15) is a stronger result than Ineq. (1.16) since most often, Ineq. (1.15) could be integrated (over the noise ε), and Ineq. (1.16) will follow.

Another approach to quantify the theoretical performance of a model selection procedure is by deriving an upper bound on the risk error that matches the so-called *minimax lower bound*. More precisely, assume that there exists a set of functions Θ_f such that the performance of the estimator $f^t,\ t \in \mathcal{T}$, of f^* is measured by the *maximum risk* of this estimator on Θ_f :

$$r(f^t) \coloneqq \sup_{f\in\Theta_f} \mathbb{E}_\varepsilon\|f^t - f\|_n^2.$$

Then, the minimax lower bound, associated with Θ_f and the empirical norm $L_2(\mathbb{P}_n)$, is defined as

$$\mathcal{R}_n^* \coloneqq \inf_{\widehat{f}} \left[r(\widehat{f})\right], \tag{1.17}$$

where $\widehat{f}$ is any measurable of the data functional estimator. Note that the minimax risk $\mathcal{R}_n^*$ provides a fundamental lower bound on the performance of any estimator uniformly over the function space Θ_f. Thus, if the statistician is able to choose $\widehat{t} \in \mathcal{T}$ (via some *data-driven* statistical procedure) such that

$$\mathbb{E}_\varepsilon \|f^{\widehat{t}} - f^*\|_n^2 \leq c_u \mathcal{R}_n^*, \tag{1.18}$$

where $c_u > 1$ is a constant, then the choice $\widehat{t}$ of the model is called *minimax optimal* (w.r.t. the empirical norm $L_2(\mathbb{P}_n)$) over the set Θ_f.

Notice that Ineq. (1.15), (1.16), and (1.18) could be presented in the $L_2(\mathbb{P}_X)$ norm, e.g., $\|f^{\widehat{t}} - f^*\|_2^2$ and its expectation $\mathbb{E}\|f^{\widehat{t}} - f^*\|_2^2$.

Model selection is used in almost all statistical procedures one can imagine. For instance, it plays a crucial role in the statistical analysis of cross-validation, penalized estimators, and signal analysis (see, e.g., [80] for a thorough review of the subject).

1.4 Iterative learning algorithms

This work addresses the problem of estimating a regression function from Eq. (1.1) using iterative learning algorithms. Iterative learning algorithms are ubiquitous in machine learning, optimization, and statistics [35, 37]. From the statistical point of view, which is the main focus of this work, the central question of interest is the *statistical performance* of these iterative algorithms (see the previous section). For example, there has been a great interest in boosting-like methods [39, 58, 118, 126]. In its original and computationally flexible version, boosting seeks to minimize empirically a loss function in a greedy fashion such that, given a set of weak (base) learners, the final estimator of the regression function (1.1) is built by iteratively re-weighting a linear combination of them. Besides that, it is worth to mention different (stochastic) gradient descent methods [35, 65] that are extensively used nowadays.

Spectral filter algorithms [19, 51, 64] is a subset of the class of iterative learning algorithms. Initially, these algorithms were introduced in the inverse problem literature (the monograph [60] provides a very detailed review) for regularization of ill-posed operator (matrix) problems. The main idea of deriving such algorithms is the fact that there is a variety of estimators that behave similarly to Tikhonov (ridge) regularization. Moreover, these algorithms belong to the class of linear estimators, meaning that the estimator (evaluated on a sample) is "proportional" to the vector of the responses $Y = [Y_1, \ldots, Y_n]^\top$. In the present work, we put a particular focus on such linear estimators called *spectral filters*. One can enumerate several examples of spectral filters: spectral cut-off, Landweber iterations (corresponds to gradient descent with a constant step-size), and Tikhonov (ridge) regularization [15, 19, 60, 122]. The precise expressions of the mentioned spectral filter functions (estimators) in the case of linear Gaussian sequence model will be given in Section 1.5.3. The definition of spectral filter function in the case of reproducing kernel Hilbert space will be rigorously defined in Chapter 2 (see Eq. (2.9) and

Definition 2.2.1).

1.5 Early stopping rules

It turns out that one should know the number of iterations when using an iterative learning algorithm. This question is all the more important for two reasons. First, interrupting the learning process will provide the user with a lower computational complexity of the algorithm. Second, it has been already observed empirically [91, 92] that stopping a learning algorithm (usually, a gradient-type one) will result in a better statistical precision than waiting until some prescribed number of iterations will be executed. We remark that the latter bad behavior in the statistical learning literature is called *overfitting* [70].

1.5.1 Validation based rules

As a motivating example, Prechelt [91] has considered an artificial neural network model and proposed several empirical strategies for stopping the learning algorithm (stochastic gradient descent) that relied on splitting the initial data into two parts: one is made for training the model, the other one – for validation (prediction on this part). This validation procedure is called the Hold-out [8]. The main motivation of this strategy was its "idealized" empirical performance that we illustrate in Fig. 1.1.

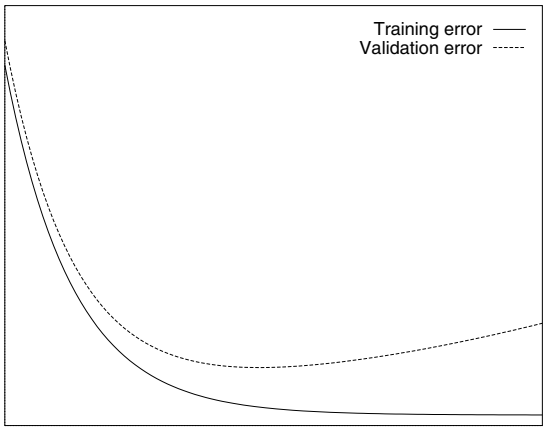


Figure 1.1 – "Idealized" training and validation error curves with an artificial neural network. Vertical: errors; horizontal: time epochs. Taken from [91].

Since for large sample sizes n, the validation error should serve as an approximation to the risk error (prediction error), as one can see in Figure 1.1, there is some number of 'time epochs' for which one achieves the minimum of the validation curve. However, the true (not "idealized") curves for the

training and validation errors are much more complicated (see Figure 2.2 in [91]), with possible local minima. In the aforementioned paper, Prechelt developed several stopping criteria and compared them on real-world data sets. In a few words, the best trade-off between the validation error and training time was achieved by the criterion based on comparing the validation error at time t and the validation error at time $t-k$, for some $k \in \mathbb{N}$.

Besides its practical evaluation, the Hold-out strategy has been proved to output minimax optimal regression function estimators in some contexts (see, e.g., [43, 45] for spectral filter algorithms in RKHS).

1.5.2 Deterministic rules

First theoretical results for the construction of early stopping rules concerned with the development of *deterministic stopping rules* [17, 39, 71, 123, 126], meaning that they depend mainly on the number of samples n. For these rules, the main focus was on either the regression function estimation in a reproducing kernel Hilbert space (RKHS) utilizing gradient descent, or boosting algorithms in regression or classification frameworks (L^2-boosting, AdaBoost, and LogitBoost) in some functional hypothesis space.

Different boosting methods.

Let us describe the framework of boosting algorithms [39, 118].

Consider a cost function $\phi : \mathbb{R} \times \mathbb{R} \mapsto [0, +\infty)$, when $\phi(y, \theta)$ denotes the cost associated with predicting θ while the true response is y. There exist three common cost functions:

— the least-squares loss $\phi(y, \theta) = 0.5(y-\theta)^2$ that yields the L^2 boosting algorithm;

— the logistic loss $\phi(y, \theta) = \ln(1 + \exp(-y\theta))$ that yields the LogitBoost algorithm;

— the exponential loss $\phi(y, \theta) = \exp(-y\theta)$ that yields the AdaBoost algorithm.

Given a loss function ϕ, one defines the population cost $f \mapsto \mathcal{L}_\phi(f)$ via

$$\mathcal{L}_\phi(f) = \mathbb{E}\left[\phi(Y, f(X))\right], \tag{1.19}$$

where $\mathbb{E}$ is the expectation w.r.t. the joint probability of (X, Y). Given some functional space $\mathcal{F}$, one minimizes the population cost, i.e.,

$$f_\phi^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{L}_\phi(f). \tag{1.20}$$

Notice that for L^2 boosting, f_ϕ^* is equal to the conditional expectation $x \mapsto \mathbb{E}[Y \mid X = x]$.

Since we do not have access to the distribution of (X, Y), the computation of f_ϕ^* is impossible. However, one can use the sample $\{X_i, Y_i\}_{i=1}^n$ and the empirical loss (error):

$$\mathcal{L}_{\phi,n}(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i, f(X_i)), \tag{1.21}$$

where the population expectation is replaced by the empirical expectation. Then, a broad class of boosting algorithms generate a sequence $\{f^t\}_{t=0}^{+\infty}$ via the updates

$$f^{t+1}(\cdot) = f^t(\cdot) - \eta_t g^t(\cdot), \qquad g^t = \underset{d \in \mathcal{F}}{\operatorname{argmax}} \langle \nabla \mathcal{L}_{\phi,n}(f^t), d(\{X_i\}_{i=1}^n) \rangle, \tag{1.22}$$

where $\{\eta_t\}_{t=0}^{+\infty}$ is a real-valued sequence of step-sizes chosen by the user, $d(\{X_i\}_{i=1}^n) = [d(X_1), \ldots, d(X_n)]^\top$, $\nabla \mathcal{L}_{\phi,n}(f) \in \mathbb{R}^n$ is the gradient taken at the vector $[f(X_1), \ldots, f(X_n)]^\top$, and $\langle h, g \rangle$ is the usual Euclidean inner product between $h, g \in \mathbb{R}^n$. Running Eq. (1.22) for an infinite number of iterations will lead to a minimizer of the empirical loss from Eq. (1.21), thus causing overfitting [118, Fig. 1].

Equipped with Eq. (1.22), Zhang et al. [126] proved the following result. Let S be a set of real-valued functions and define

$$\operatorname{span}(S) = \left\{ \sum_{j=1}^m w_j f_j \mid f_j \in S, w_j \in \mathbb{R}, \ m = 1, 2, \ldots \right\},$$

which forms a linear functional space. For all $f \in \operatorname{span}(S)$, we can define its 1-norm w.r.t. the basis of S as:

$$\|f\|_1 = \inf \left\{ \|w\|_1 \mid f = \sum_{j=1}^m w_j f_j \ ; \ f_j \in S, \ m = 1, 2, \ldots \right\}.$$

If t_n is a sequence of real numbers such that $\lim_{n\to\infty} t_n = \infty$, then under some additional assumptions on the step-size, for any $\hat{t}(n) \geq t_n$ such that $\|f^{\hat{t}(n)}\|_1 \leq \beta_n$ (β_n is some carefully chosen sequence),

$$\lim_{n\to\infty} \mathcal{L}_\phi(f^{\hat{t}(n)}) = \inf_{f \in \operatorname{span}(S)} \mathcal{L}_\phi(f), \tag{1.23}$$

where $\mathcal{L}_\phi(f)$ is the population cost of f associated with the loss function ϕ, and f^t is the output of a chosen boosting algorithm at the iteration t. Thus, one can conclude that boosting algorithms are consistent after some number of iterations $\hat{t}(n)$. Therefore, one can stop a chosen algorithm at the iteration $\hat{t}(n)$ and achieve an (asymptotically) meaningful statistical precision. Obviously, the result in Eq. (1.23) is only theoretical in nature, and $\hat{t}(n)$ is not computable in practice.

Gradient descent learning.

Yao et al. [123] focused on constructing an early stopping rule that should recover the famous bias-variance trade-off [70, Section 2.9] of the gradient descent estimator. They assumed the following: for some $s > 0$, the regression function $f^* \in T_k^s(B_R)$, where $B_R = \{f \in L_2(\mathbb{P}_X) \mid \|f\|_2 \leq R\}$ ($R > 0$), and $T_k : L_2(\mathbb{P}_X) \mapsto L_2(\mathbb{P}_X)$ is the *kernel integral operator* associated with some reproducing kernel $\mathbb{K}$. The aforementioned assumption is called a *source condition* in the statistical learning literature [44, 49, 104] (we use an assumption related to the source condition in Chapter 2 as well). One can claim that the parameter s controls the smoothness of the regression function (the bigger s – the smoother

regression function). In more detail, given the data $\{(x_i, Y_i)\}_{i=1}^n$, the following empirical error was minimized over the reproducing kernel Hilbert space $\mathcal{H}$ associated with $\mathbb{K}$:

$$\operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} (f(x_i) - Y_i)^2 \right\}. \tag{1.24}$$

Then by the reproducing property (1.14), the iterations of gradient descent $\{f^t\}_{t\in\mathbb{N}} \in \mathcal{H}$ with the step size $\{\eta_t\}_{t=0}^{+\infty}$ are defined as

$$f^{t+1}(\cdot) = f^t(\cdot) - \frac{\eta_t}{n}\sum_{i=1}^{n}(f^t(x_i) - Y_i)\mathbb{K}(\cdot, x_i), \qquad f^0 = 0. \tag{1.25}$$

Notice that Eq. (1.25) is equivalent to the boosting procedure from Eq. (1.22) with the least-squares loss (L^2 boosting).

Yao et al. [123] proved that, given some parameter $\theta \in [0, 1)$ and the step-size of gradient descent $\eta_t \asymp (t+1)^{-\theta}$ ($\asymp$ means "up to a constant factor that can depend only on the kernel"), $t \in \mathbb{N}$, for the stopping rule $t^*(n) = \left\lceil n^{\frac{1}{(2s+2)(1-\theta)}} \right\rceil$, the following holds:

$$\|f^{t^*(n)} - f^*\|_2 \leq C(\delta, \mathbb{K}, \theta) n^{-\frac{s}{2s+2}} \tag{1.26}$$

with probability at least $1 - \delta$, where constant $C(\delta, \theta, \mathbb{K})$ depends only on δ, θ, and a uniform upper bound on the kernel. Above, $\lceil x \rceil$ denotes the smallest integer greater than or equal to $x \in \mathbb{R}$.

As it was clarified in [123, Remark 2.3], the high probability upper bound from Ineq. (1.26) *does not match* the minimax lower bound, under the assumptions made. Nevertheless, when $s \to \infty$ (corresponds to very smooth functions), the upper bound matches the well-known (fast) asymptotic rate $\mathcal{O}\left(n^{-1/2}\right)$, which says that upper bound (1.26) is meaningful in some sense.

1.5.3 Minimum discrepancy principle rule

Contrary to the deterministic early stopping rules that we have considered previously, the focus of the present work is to give some theoretical (statistically optimal) and practical justifications of a *data-driven* approach that is called the *minimum discrepancy principle*. This approach was originally developed as Morozov's discrepancy principle [5, 60, 84] for solving (potentially nonlinear) inverse ill-posed operator problems. Since this principle is the main research subject in the present work, let us consider its historical part in more detail.

MDP for linear ill-posed problems.

For inverse ill-posed problems, one considers [60] linear operator equations of the form

$$Az = y, \tag{1.27}$$

where A is a bounded linear operator between some Hilbert spaces $\mathcal{Z}$ and $\mathcal{Y}$. It would be too restrictive to assume that A has its inverse A^{-1}. Nevertheless, one might still be interested in some generalized solution of Eq. (1.27), i.e., some element that solves it but in an approximate sense. To do that, minimization of the least squares solution $\|y - Az\|^2$ is considered, which results in the *normal equation* and its best-approximate solution $z^\dagger$:

$$A^*Az = A^*y, \qquad z^\dagger \coloneqq A^\dagger y, \tag{1.28}$$

where A^* is the adjoint operator of A, and $A^\dagger$ is the Moore-Penrose generalized inverse [60, 89] of A. Most iterative methods for approximating $A^\dagger y$ are based on a transformation of the normal equation (1.28) into equivalent fixed point equations of type

$$z = z + A^*(y - Az). \tag{1.29}$$

Usually, y (and z via Eq. (1.27), accordingly) is represented by its corrupted by some deterministic noise version y^σ (z^σ, respectively), where σ is the noise parameter. Then, the first natural algorithm to solve the fixed point equation (1.29) is Landweber iterations [75] (corresponds to gradient descent with a constant step-size $0 < \eta \leq \|A\|^{-2}$). Given y^σ and an initial guess $z_0^\sigma \in \mathcal{Z}$, Landweber iterations take the form

$$z_m^\sigma = z_{m-1}^\sigma + \eta A^*(y^\sigma - Az_{m-1}^\sigma), \quad m = 1, 2, \dots \tag{1.30}$$

We write z_m (corresponds to $\sigma = 0$) instead of z_m^σ when one iterates with precise data $y^\sigma = y$. Without loss of generality, we can say that $\|A\| \leq 1$, and $z_0^\sigma = 0$. With Eq. (1.30) at hand, one can obtain the following result.

Theorem 1.5.1 (Theorem 6.1 in [60]). *Define $\mathcal{D}(A^\dagger)$ as the domain of the operator $A^\dagger$, then, if $y \in \mathcal{D}(A^\dagger)$, $z_m \to A^\dagger y$ as $m \to \infty$. If $y \notin \mathcal{D}(A^\dagger)$, then $\|z_m\| \to \infty$ as $m \to \infty$.*

Thus, the sequence z_m associated with y, converges (in the norm in $\mathcal{Z}$) to the best-approximate solution $z^\dagger$. However, we cannot provide the same conclusion for the corrupted version z_m^σ.

One is able to notice that, using Eq. (1.30), we can introduce the function

$$g_m(\lambda) = \sum_{j=0}^{m-1} (1 - \lambda)^j \tag{1.31}$$

that yields

$$z_m^\sigma = g_m(A^*A)A^*y^\sigma.$$

In the formula above, we call $g_m(\lambda)$ the *spectral filter function* of Landweber iterations. As another classical example of spectral filter function, we can mention Tikhonov (ridge) regularization, parame-

terized by some $\alpha_m > 0$, as

$$g_{\alpha_m}(\lambda) = \frac{1}{\alpha_m + \lambda}, \quad \alpha_m \to 0 \ \text{ as } \ m \to \infty. \tag{1.32}$$

With this choice above, $z^{\sigma}_{\alpha_m}$ solving the perturbed normal equation $A^*Az + \alpha_m z = A^*y^{\sigma}$ is equivalent to

$$z^{\sigma}_{\alpha_m} = (A^*A + \alpha_m I)^{-1}A^*y^{\sigma} = g_{\alpha_m}(A^*A)A^*y^{\sigma}, \tag{1.33}$$

where I is the identity operator such that $I : \mathcal{Z} \mapsto \mathcal{Z},\ Iz = z$ for any $z \in \mathcal{Z}$.

Further, assume that for $y^{\sigma} \notin \mathcal{D}(A^{\dagger}), \|y^{\sigma} - y\| \leq \sigma$. This will lead to the minimum dicrepancy (MDP) stopping rule $m(\sigma, y^{\sigma}, \vartheta)$ (see [60, Section 6.1] for more details) defined as

$$\underbrace{\|y^{\sigma} - Az^{\sigma}_{m(\sigma,y^{\sigma},\vartheta)}\|}_{\text{Residuals}} \leq \vartheta\sigma, \tag{1.34}$$

with a fixed parameter $\vartheta \geq 1$ to choose by the user. We should emphasize here that the stopping rule $m(\sigma, y^{\sigma}, \vartheta)$ with a proper tuning of ϑ should provide an appropriate trade-off of the *approximation error* $z^{\dagger} - z_m$ and the *data error* $z_m - z^{\sigma}_m,\ m = 1, 2, \ldots$

MDP for statistical inverse problems.

Moving back to the statistical learning (specifically, regression) setting, the present work was inspired by [29], where Blanchard et al. studied the minimum discrepancy principle for a general spectral filter function in the following linear inverse model:

$$Y = A\zeta + \sigma\xi, \tag{1.35}$$

where the signal $\zeta \in \mathbb{R}^D, A \in \mathbb{R}^{n\times D}(D \leq n); Y, \xi \in \mathbb{R}^n$, ξ is Gaussian stochastic noise. Then, by projecting Eq. (1.35) onto the space spanned on the eigenvectors $(v_1, \ldots, v_D)$ of $(A^*A)^{1/2}$, one obtains the associated *linear Gaussian sequence model* that takes the form

$$Y_i = \lambda_i\zeta_i + \delta\varepsilon_i,\ i = 1, \ldots, D, \tag{1.36}$$

$$Y_i = \delta\varepsilon_i,\ i = D + 1, \ldots, n, \tag{1.37}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D > 0$ are the eigenvalues of $(A^*A)^{1/2}$, $\zeta_i = \langle\zeta, v_i\rangle/n,\ Y_i = \langle Y, w_i\rangle/n$, with $w_i = \sqrt{n}\frac{A^*v_i}{\|Av_i\|}$, and $\{\varepsilon_i\}_{i=1}^n$ is independent standard Gaussian noise $\mathcal{N}(0, \sigma^2)$, with the "noise level" parameter $\delta = \sigma/\sqrt{n}$.

The objective was to recover the signal $\zeta = (\zeta_i)_{1\leq i\leq D}$ from the data $(Y_i)_{1\leq i\leq n}$. To do that, the following linear spectral filter estimator was considered

$$\widehat{\zeta}^{(t)}_i = \gamma^{(t)}_i\lambda_i^{-1}Y_i,\ i = 1, \ldots, D;\ t \geq 0, \tag{1.38}$$

where $(\gamma_i^{(t)})_{i=1,\dots,D;t\geq 0}$ are called *spectral filters* that should satisfy: a) $\gamma_i^{(t)} \in [0,1],\ i=1,\dots,D$; b) $\gamma_i^{(t)}$ is a continuous non-decreasing function of t; c) $\gamma_i^{(0)}=0$, and $\gamma_i^{(t)} \to 1$ as $t \to \infty$. Typical spectral filters include:

— spectral cut-off (equivalent to the truncated singular value decomposition) with $\gamma_i^{(t)} = \mathbb{I}\,(i \leq t)$,
— Landweber iterations $\gamma_i^{(t)} = 1-(1-\lambda_i^2)^t$, corresponding to gradient descent with step-size 1,
— Tikhonov (ridge) regularization $\gamma_i^{(t)} = \lambda_i^2/(\lambda_i^2+\alpha_t)$, where $\alpha_t \to 0$ as $t \to \infty$.

One can notice that $\gamma_i^{(t)} = \lambda_i g_t(\lambda),\ i=1,\dots,D$, for the spectral filter function $g_t(\lambda)$ defined as in Eq. (1.31) or Eq. (1.32). Two important quantities for the analysis of the estimator $(\widehat{\zeta}^{(t)})_{t\geq 0}$ are its bias and variance defined as

$$B_{t,\zeta}^2 \coloneqq \|A\left(\mathbb{E}_\varepsilon\left[\widehat{\zeta}^{(t)}\right]-\zeta\right)\|^2 = \sum_{i=1}^{D}(1-\gamma_i^{(t)})^2\lambda_i^2\zeta_i^2, \tag{1.39}$$

$$V_{t,\zeta} \coloneqq \mathbb{E}_\varepsilon\left[\|A\left(\widehat{\zeta}^{(t)}-\mathbb{E}_\varepsilon\left[\widehat{\zeta}^{(t)}\right]\right)\|^2\right] = \delta^2\sum_{i=1}^{D}\left(\gamma_i^{(t)}\right)^2, \tag{1.40}$$

where we recall that $\mathbb{E}_\varepsilon$ denotes the expectation w.r.t. the noise $\{\varepsilon_i\}_{i=1}^n$. Due to the monotonicity of the spectral filter $\gamma_i^{(t)},\ i=1,\dots,D$, the bias term is a *non-increasing* function of t, whereas the variance term $V_{t,\zeta}$ is a *non-decreasing* function of t. Then, the risk (prediction) error at time t is equal to the sum of $B_{t,\zeta}^2$ and $V_{t,\zeta}$.

In the aforementioned work, the authors tried to recover the bias-variance trade-off [70, 108] of the linear estimator $(\widehat{\zeta}^{(t)})_{t\geq 0}$ (the intersection point of $B_{t,\zeta}^2$ and $V_{t,\zeta}$) through the control of the residuals (we also call this quantity the "reduced empirical risk" in the future), which is a *non-increasing* function of t and are minimized during the learning process:

$$\widetilde{R}_t \coloneqq \|Y-A\widehat{\zeta}^{(t)}\|^2 = \sum_{i=1}^{D}(1-\gamma_i^{(t)})^2Y_i^2, \qquad \forall t\geq 0, \tag{1.41}$$

as a data fidelity criterion. We illustrate the typical behavior of the quantities $B_{t,\zeta}^2$, $V_{t,\zeta}$, $B_{t,\zeta}+V_{t,\zeta}$ (the risk error), and the residuals $\widetilde{R}_t$ w.r.t. time t in Figure 1.2.

It turned out that the expectation (over the noise $\{\varepsilon_i\}_{i=1}^n$) of the residuals $\widetilde{R}_t,\ t\geq 0$, is equal to the bias term plus a deviation term related to the filter $\gamma_i^{(t)},\ i=1,\dots,D$. This way, the obstacle of not knowing the bias term (it depends on the estimated signal ζ) can be resolved by properly controlling $\mathbb{E}_\varepsilon\widetilde{R}_t$. Pushing this logic a bit further, if the residuals are close to its expectation, then it yields the stopping rule already presented for the inverse problems above:

$$\tau = \inf\{t\geq t_0 \mid \widetilde{R}_t \leq \kappa\}, \qquad \text{for some } \kappa>0 \text{ and } t_0\geq 0. \tag{1.42}$$

It appeared that to mimic the behavior of the bias-variance trade-off, κ in Eq. (1.42) should be equal to $D\delta^2$. It is worth to mention that the "starting time" t_0 was introduced artificially, which is an

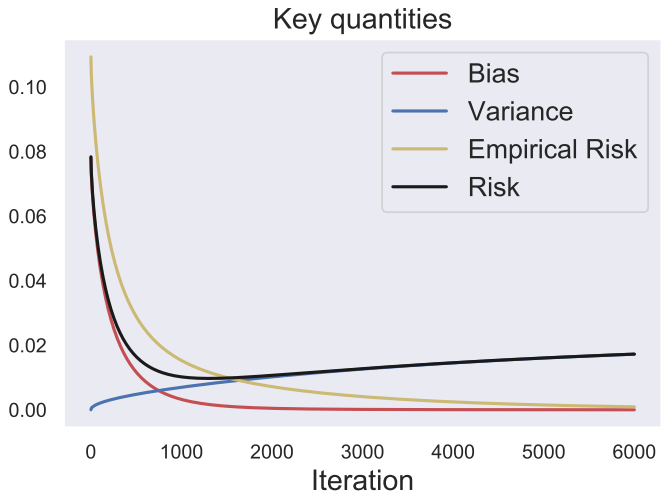


Figure 1.2 – The bias, variance, risk, and residuals ("reduced empirical risk") behavior.

inconvenience of the proposed rule. Apart from that, the stopping rule τ was proved to be optimal (in terms of an oracle inequality of type (1.16) for the prediction error) under very strong assumptions, i.e.,

— There exists a constant C_{l^1,l^2} such that for all $t \geq t_0$, we have

$$\sum_{i=1}^{D} \gamma_i^{(t)} \leq C_{l^1,l^2} \sum_{i=1}^{D} (\gamma_i^{(t)})^2.$$

— Define $t_\omega \geq t_0$ as the iteration of the bias-variance intersection of a spectral filter estimator, then $\sqrt{D} \lesssim \sum_{i=1}^{D} (\gamma_i^{(t_\omega)})^2$, where $\lesssim$ means an inequality up to a numeric constant.

Notice that these assumptions could be barely checked in practice.

Discussion and contribution of the thesis.

A lot of quite recent papers have studied stopping rules like (1.42) in different contexts. For instance, [32, 34] defined a MDP stopping rule for (kernel) conjugate gradient descent while [30, 51] considered the spectral cut-off estimator in the linear regression or linear Gaussian sequence models. In [34], it appeared that when taking into consideration conjugate gradient descent, the usual MDP stopping rule (1.34) should be modified to achieve optimal rates. That was done by means of weighting the residuals from Eq. (1.34), by applying the operator $(\alpha I + A^*A)^{-1/2}$. Moreover, Blanchard et al. [30] concluded that the stopping rule (1.34) could not produce (without additional strong assumptions) statistical optimality over Sobolev-type ellipsoids for the spectral cut-off estimator in the linear Gaussian sequence model. Stankewitz [105] corrected this sub-optimality by introducing the so-called *polynomial smoothing strategy* for the residuals. This strategy consists in weighting the residuals by utilizing the eigenvalues from Eq. (1.36) as $\{\lambda_i^\theta\}_{i=1}^D$, where θ is called the smoothing parameter. It

was shown, via an oracle inequality of type (1.16) that this strategy, for some values of the smoothing parameter, provides an optimal estimator. Another version of the smoothing strategy for the residuals involving the notion of *effective dimension* of the kernel, introduced previously by Zhang in [125], has been considered in [50].

In Chapter 2, we reexamine the early stopping rule (1.42) (without the t_0-assumption) applied to gradient descent and (Tikhonov) ridge regression (cast as an iterative algorithm) in a reproducing kernel Hilbert space. We show (via an upper bound of type (1.18) for the $L_2(\mathbb{P}_X)$ norm) that this rule provides a functional estimator that achieves the minimax-optimal rate for finite-rank kernels. After that, for infinite-rank reproducing kernels, the polynomial smoothing strategy is discussed and applied. More precisely, for reproducing kernels associated with Sobolev spaces and some *explicit* values of the smoothing parameter, we achieve the minimax-optimal rate of type (1.18) for *smoothed* MDP-based early stopped gradient descent and kernel ridge regression, in terms of the $L_2(\mathbb{P}_n)$ norm.

1.5.4 Other data-driven approaches

Another interesting approach to the problem of constructing an early stopping rule in the framework of Reproducing kernel Hilbert space is proposed in the work [92, 118]. There, the authors consider gradient descent and different boosting algorithms (L^2 boosting, Adaboost, and LogitBoost), respectively, for estimating the unknown regression function f^*. Let us describe their approach more closely and briefly at the same time.

Early stopping via localized Rademacher complexity and critical radius.

Assume that one is given some reproducing kernel Hilbert space $\mathcal{H}$ with its associated reproducing kernel $\mathbb{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Further, the kernel $\mathbb{K}$ will generate the normalized Gram (kernel) matrix $K_n = \mathbb{K}(x_i, x_j)/n,\ i, j \in \{1, \ldots, n\}$, where $\{x_i \in \mathcal{X}\}_{i=1}^n$ are given covariates.

The quantity on which their theoretical analysis relies is called the (empirical) localized Rademacher complexity [16, 82, 83, 114] of the unit ball in some functional space $\mathcal{F}$ defined as $\mathbb{B}_{\mathcal{F}}(1) := \{f \in \mathcal{F} \mid \|f\|_{\mathcal{F}} \leq 1\}$.

Definition 1.5.1. For any $\epsilon > 0$ and functional space $\mathcal{F}$, consider the empirical localized Rademacher complexity

$$\widehat{\mathcal{R}}_n(\epsilon, \mathcal{F}) = \mathbb{E}_{\mathbf{r}} \left[\sup_{\substack{f \in \mathcal{F} \\ \|f\|_n \leq \epsilon}} \left| \frac{1}{n} \sum_{i=1}^n \mathbf{r}_i f(x_i) \right| \right], \tag{1.43}$$

where $\{\mathbf{r}_i\}_{i=1}^n$ are i.i.d. Rademacher variables ($\{-1, +1\}$-random variables with equal probability $\frac{1}{2}$).

This complexity measure has become a standard tool in the modern empirical process and nonparametric regression analysis [114, Chapter 5, 13]. For a reproducing kernel Hilbert space $\mathcal{H}$, [82, 83] proved that $\widehat{\mathcal{R}}_n(\epsilon, \mathbb{B}_{\mathcal{H}}(1))$ can be upper and lower bounded (up to constant factors) by the following

quantity called the *kernel complexity function*

$$\widehat{R}_n(\epsilon, \mathcal{H}) = \left[\frac{1}{n}\sum_{i=1}^{n} \min\left\{\widehat{\mu}_i, \epsilon^2\right\}\right]^{1/2}, \tag{1.44}$$

where $(\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_n)$ are the eigenvalues of K_n. It corresponds to a rescaled sum of the empirical eigenvalues truncated at ϵ^2.

Using this measure $\widehat{R}_n(\epsilon, \mathcal{H})$, one can introduce the *critical empirical radius* $\widehat{\epsilon}_n$ to be the smallest positive solution to the following inequality

$$\frac{\widehat{R}_n(\epsilon, \mathcal{H})}{\epsilon} \leq \frac{\epsilon}{2e\sigma}. \tag{1.45}$$

Eq. (1.45) is called the *critical inequality* of $\mathbb{B}_{\mathcal{H}}(1)$. One can verify [114, Lemma 13.6] that the left-hand side is a non-increasing function of ϵ, which guarantees that $\widehat{\epsilon}_n$ exists and is unique. Then, Raskutti et al. [92] have proved that the stopping rule

$$\widehat{T}_{\mathrm{RWY}} = \operatorname{argmin}\left\{t > 0 \mid \widehat{R}_n\left(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\right) > (2e\sigma\eta t)^{-1}\right\}, \tag{1.46}$$

where η is a step-size of gradient descent, tightly estimates the bias-variance trade-off of the gradient descent estimator $(f^t)_{t>0}$ from Eq. (1.25), meaning that the following holds with probability at least $1 - c_1 \exp\left(-c_2 n \widehat{\epsilon}_n^2\right)$, where c_1 and c_2 are some positive constants:

$$\|f^{\widehat{T}_{\mathrm{RWY}}} - f^*\|_n^2 \leq 12\widehat{\epsilon}_n^2. \tag{1.47}$$

Discussion and contribution of the thesis.

First, notice that $\widehat{T}_{\mathrm{RWY}}$ only depends on the data through the design. This means that the stopping rule does not depend on the regression noise $\{\varepsilon_i\}_{i=1}^n$. One can say (see, e.g., [121, Theorem 1] for the respective lower bound in expectation) that the result (1.47) provides an upper bound that matches (up to a constant) the minimax lower bound over $\mathbb{B}_{\mathcal{H}}(1)$, where $\mathcal{H}$ is a reproducing kernel Hilbert space associated with the class of *regular kernels*. This class includes the Gaussian, Sobolev, and polynomial kernels, among others. The work [118] extended the previously mentioned strategy of the stopping rule $\widehat{T}_{\mathrm{RWY}}$ to the case of boosting learning algorithms. It was done again by estimating the bias-variance trade-off with a slightly different complexity measure that is called the *localized Gaussian complexity* of $\mathbb{B}_{\mathcal{H}}(1)$ [114, Chapter 13].

A thoughtful reader would be able to remark that the minimax optimality in both papers [92, 118] is achieved *only* over the unit ball in $\mathcal{H}$. This assumption was encoded in the definitions of these stopping rules (see, e.g., Eq. (1.46)). In Chapter 2, we remove such a strong assumption and consider a ball of an *arbitrary* radius R in $\mathcal{H}$.

After that, we prove that, for finite-rank kernels, the minimum discrepancy principle stopping rule (1.42) yields a minimax optimal function estimator over the ball of radius R in a reproducing kernel Hilbert space $\mathcal{H}$. Further, we modify the previously mentioned discrepancy principle stopping rule utilizing the polynomial smoothing strategy [105]. Then for some *explicit* values of the smoothing parameter, the modified stopping rule yields a minimax optimal functional estimator over the ball of radius R in a reproducing kernel Hilbert space $\mathcal{H}$, in particular, associated with infinite-rank Sobolev spaces.

Thus, one can say that the aforementioned (MDP-based) rules adapt to the unknown radius R, meaning that the knowledge of the radius of the ball in RKHS *is not required* for achieving statistical optimality. Moreover, these rules are *design-independent*.

1.6 Linear estimators and tuning parameter selection

Chapters 3 and 4 present a data-driven procedure to choose the tuning parameter of a linear estimator. This can be done by considering an iterative procedure (over the unknown tuning parameter) and stopping it according to some carefully designed criterion.

This section is as follows. First, we define the *linear estimator* [7, 70, 116] and give some famous examples of it. Second, we provide several widely-used strategies for tuning the parameter of a linear estimator. Third, we list the contributions made in the thesis concerning the described framework.

1.6.1 Linear estimators description

Assume that one has a model selection set $\Lambda \coloneqq \{\lambda_1, \dots, \lambda_S\}$ for some $S \in \mathbb{N}$. A linear estimator $F^\lambda \coloneqq [f^\lambda(x_1), \dots, f^\lambda(x_n)]^\top$ of the regression function from Eq. (1.1) could be seen as

$$F^\lambda = A_\lambda Y, \qquad \lambda \in \Lambda, \tag{1.48}$$

where A_λ is an $n \times n$ matrix and $Y = [Y_1, \dots, Y_n]^\top$, λ is a (smoothing) parameter to choose (learn/tune).

There are several well-known examples of the linear estimator. Let us enumerate them.

k-nearest neighbor regression [26, 67]. Assume that one is given a similarity measure $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and k is a strictly positive integer, then, from n covariates $x_1, \dots, x_n$, one can find k nearest neighbors of x_i, $i = 1, \dots, n$, e.g., find a set J_i of k points, which are among the k closest to x_i according to d. We can build an $n \times n$ matrix A_k of nearest neighbors, which is equal to $1/k$ for all pairs (i, j) such that $j \in J_i$ for all $i \in \{1, \dots, n\}$, and equal to zero otherwise. Usually, one chooses the Euclidean metrics for d.

Nadaraya-Watson regression [85, 116]. Assume that one is given a 'window function' $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$, and we build the $n \times n$ matrix W of pairwise evaluations. Then, the corresponding matrix

A_h from Eq. (1.48) is equal to $A_h = WD^{-1}$, where $D = \operatorname{diag}(W\mathbf{1})$ ($\mathbf{1}$ is the $n \times n$ unit matrix) is the diagonal matrix of row sums. Usually, one chooses the matrix W to be $W_{ij} = \exp\left(-\|x_i - x_j\|^2/h\right)$, $h > 0$, where x_i and x_j, $i, j \in \{1, \ldots, n\}$, are the input data points, and h is the bandwidth to tune.

Variable selection in regression [102]. Consider the standard nonparametric regression model $Y_i = f^*(x_i) + \varepsilon_i$, where $\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ is a full-rank fixed design matrix with $d \leq n$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$, is i.i.d. Gaussian observation noise. In the variable selection setting, one would like to select a subset $J \subseteq \{1, \ldots, d\}$ such that non-relevant variables of $\mathbf{X}$ will be discarded. Denote X_J as the matrix of size $n \times |J|$ composed of the columns of $\mathbf{X}$ indexed by J. Consider then the linear predictor $F^J = A_J Y$, where the matrix $A_J = X_J(X_J^\top X_J)^{-1} X_J^\top$. It is worth to mention that the tuning parameter here could be the cardinality $|J|$ of the chosen subset J (not the whole subset), and for the matrix A_J, $\operatorname{tr}(A_J) = |J|$. We consider this estimator in more detail in Chapter 4.

Pinsker filters [90]. In this case, the smoothing matrix from Eq. (1.48) is equal to $A_{\omega,k} = \operatorname{diag}\left\{(1 - (k^\alpha/\omega))_+, \; k = 1, \ldots, n\right\}$ for some parameters $\alpha, \omega > 0$, where $x_+ = \max(x, 0)$.

The matrix A_λ plays a crucial role in the performance evaluation of the linear estimators described above. More precisely, one can define the so-called "modal's degree of freedom" [70, Chapter 7]

$$\mathrm{df}(\lambda) = \operatorname{tr}(A_\lambda),$$

which measures the complexity of the learning model (1.48). This way, for most of the cases, the variance of the linear estimator (1.48) will be a non-decreasing function w.r.t. $\mathrm{df}(\lambda)$ (see Fig 1.3). In particular, for the k-NN regression estimator, the variance term is proportional to the degree of freedom (see Chapter 3).

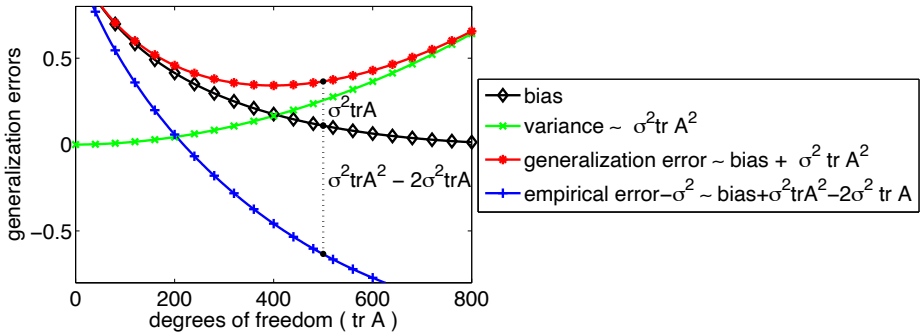


Figure 1.3 – The bias, variance, empirical risk, and generalization (prediction) error; taken from [6].

1.6.2 Strategies to tune the parameter

Tuning parameter selection in linear estimators is an old problem. To the best of our knowledge, the first results for this model selection problem were concerned with the Forward Selection and Backward Elimination procedures [70, Chapter 3] for variable selection in linear regression. After that, a seminal breakthrough for this problem was achieved by developing Akaike's AIC [2] and Mallows' C_p [79] criteria.

Mallow's C_p criterion [70, 76, 79].

For example, Mallows considered the standard linear regression model $Y = \mathbf{X}\theta + \varepsilon \in \mathbb{R}^n$ with the idea to propose an unbiased estimator for the risk error $\mathbb{E}_\varepsilon[\|\mathbf{X}(\widehat{\theta}_J - \theta)\|^2]$ using the empirical error $\|Y - \mathbf{X}\widehat{\theta}_J\|^2$, where $\widehat{\theta}_J$ is an estimator of θ, based on the variable selection set $J \subseteq \{1, \dots, d\}$. This strategy will result in the following generalized criterion for choosing the parameter λ in any linear estimator of type $A_\lambda Y$ (in this case, often called C_L):

$$C_p(\lambda) = \frac{\|Y - A_\lambda Y\|^2}{n} + 2\frac{\mathrm{df}(\lambda)}{n}\widehat{\sigma}^2, \quad \lambda \in \Lambda, \tag{1.49}$$

where $\widehat{\sigma}^2$ is an estimator of the noise variance σ^2 obtained from a low-bias model. Using this criterion, we adjust the training error by a factor proportional to the "degree of freedom". In practice, the rule for selecting the "best" candidate in Λ is the minimization of $C_p(\lambda)$.

AIC criterion.

The Akaike information criterion [2, 70] is a similar (but a more general) estimate of the risk error, where a log-likelihood loss function is used. In the case of linear estimators with Gaussian noise, maximum likelihood and least-squares are the same things. This gives the AIC criterion as

$$\mathrm{AIC}(\lambda) = \frac{1}{n\widehat{\sigma}^2}\left(\|Y - A_\lambda Y\|^2 + 2\mathrm{df}(\lambda)\widehat{\sigma}^2\right), \quad \lambda \in \Lambda, \tag{1.50}$$

where $\widehat{\sigma}^2$ is the noise variance estimator from Eq. (1.49). Notice that in the mentioned case, $\mathrm{AIC}(\lambda)$ and $C_p(\lambda)$ produce the same model selection procedure.

Both C_p and AIC criteria have been criticized in the literature, especially for the constant 2 in their definitions. This is why some authors proposed corrections to these criteria [100, 103, 120]. Nevertheless, AIC, C_p, and other related penalized model selection procedures have been proved to satisfy oracle inequalities (1.16) in some frameworks [27] when $|\Lambda| \leq Cn^\alpha$, for some constants $C, \alpha \geq 0$. Furthermore, the proposed criteria are still widely used in practice [107].

Data-splting strategies [8, 63, 117].

Another approach for choosing the tuning parameter is based on a *data splitting strategy*, meaning that a part of data (the training sample) is used for training a learning algorithm, and the remaining part of data (the test sample) is used for the evaluation of the performance of the algorithm. If data

is distributed in an i.i.d. manner, then the error of the algorithm on the test sample can serve as an approximation of the true risk. One can enumerate several data-splitting procedures (see [8] for a thorough review). The most simple one is called the *Hold-out validation strategy* that relies on a single split of data. An example of the so-called exhaustive data splitting could be the *leave-p-out* strategy, with $p \in \{1, \dots, n-1\}$ (n is the sample size), where every possible subset of p data points is hold-out of the sample and used to validate (to estimate the true risk error). Notice that the case $p = 1$ corresponds to the *leave-one-out* strategy. Often, considering $\binom{n}{p}$ training samples is computationally exhaustive and only the case $p = 1$ is implementable in practice.

Generalized cross-validation [42, 53, 70].

A rotation-invariant version of the leave-one-out strategy called the *generalized cross-validation* was proposed by [53]. The goal was the same – to estimate the risk of a linear estimator of type $A_\lambda Y$ (A_λ is an $n \times n$ matrix) as follows.

$$\mathrm{GCV}(\lambda) = \frac{\|Y - A_\lambda Y\|^2}{n(1 - n^{-1}\mathrm{tr}(A_\lambda))}, \quad \lambda \in \Lambda. \tag{1.51}$$

The practitioner who wants to apply the GCV strategy should, first, construct a tuning parameter set Λ and after that choose $\widehat{\lambda}$ that minimizes the $\mathrm{GCV}(\lambda)$ criterion. The asymptotic optimality of $\mathrm{GCV}(\widehat{\lambda})$ for several linear estimators, meaning that $\|f^{\widehat{\lambda}} - f^*\|_n^2 / \inf_{\lambda \in \Lambda} \|f^\lambda - f^*\|_n^2 \to 1$ in probability, is established in [76]. However, there are examples [77] for which GCV is not asymptotically optimal for ridge (Tikhonov) regularization. In order to get asymptotic optimality, one should have a special condition on a tail sum of the eigenvalues of the matrix $\mathbf{X}^\top \mathbf{X}$, where $\mathbf{X}$ is the design matrix.

V–fold cross-validation.

By far, the most usable model selection procedure (due to its relatively mild computational cost for small V) is the so-called V-fold cross-validation [8, 63, 70]. This procedure splits the data into V roughly equal-sized parts; after that, for each $v = 1, \dots, V$, for the v^{th} part we fit the model to the other $V - 1$ parts of the data and calculate the prediction error of the fitted model on the v^{th} part. More formally, let $\kappa : \{1, \dots, n\} \mapsto \{1, \dots, V\}$ be an indexing function that indicates the partition to which observation i is allocated by the splitting randomization. Denote $f^\lambda_{-v}(x)$ as the fitted linear estimator with the tuning parameter λ, computed with the v^{th} part removed, then

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f^\lambda_{-\kappa(i)}(x_i))^2, \quad \lambda \in \Lambda. \tag{1.52}$$

Then, one should minimize the criterion from Eq. (1.52) over some grid of values of $\lambda \in \Lambda$. Note that V-fold cross-validation with $V = n$ is equivalent to the leave-one-out procedure. An interesting question that could be asked is how to choose V? It is often suggested [70] that V between 5 and 10 is optimal since the statistical performance does not increase a lot for larger values of V, whereas the

leave-one-out ($V = n$) could suffer from high variance.

1.6.3 Contribution of the thesis

In *all* aforementioned model selection procedures, the user should compute a functional estimator f^λ (and one of the selection criteria) over some grid of values of $\lambda \in \Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_S\}$. In practice, it is generally computationally demanding, especially when the cardinality of Λ is large. For this reason, in this work, we propose applying a special procedure to choose λ without computing all these estimators. The strategy will rely on the fact that one can arrange the grid Λ such that, by computing *iteratively* f^λ, the variance term of f^λ will not decrease at each iteration. Thus, this term will stay monotonic as a function of λ. To mention one example of such a situation, consider the k-NN regression estimator (hence, $\lambda = k$). Then, it is known [26, 67] that the variance term for this estimator is equal to σ^2/k, where σ^2 is the noise variance in Eq. (1.1). Therefore, decreasing iteratively the value of $k \in \{1, \ldots, n\}$ will result in increasing the variance term. Eventually, one has to stop this process since, if it is not stopped, the variance term will be large, and there would be no hope to get optimality.

Keeping in mind the strategy described above, as it was in the case of spectral filter estimators, the user should control the *empirical risk* that shows how well f^λ fits Y:

$$R_\lambda = \|(I_n - A_\lambda)Y\|_n^2, \quad \lambda \in \Lambda. \tag{1.53}$$

Intuitively, if we iterate a learning algorithm in such a way that the variance term at λ of the estimator f^λ is non-decreasing, then R_λ should be approximately a non-increasing function of λ (see Figure 1.3 for an illustration). However, there is no monotonicity, as it was in the case of spectral filter estimators. Then, one needs to define a threshold applying to R_λ so that one would stop the iterations if R_λ crosses this threshold. The most natural way to do it is to detect the first iteration $\widehat{\lambda}$ at which $R_{\widehat{\lambda}} \approx \sigma^2$, where σ^2 is the noise variance (and $\approx$ means an approximate equality). The mentioned approach has been already explored in the statistical (and not only) literature and is called the *minimum discrepancy principle* (MDP) (see, e.g., Section 1.5 for a discussion about MDP applied to spectral filter algorithms). The precise expressions for the threshold parameter with different linear estimators will be mentioned in Chapter 3 and Chapter 4.

Let us now enumerate our main contributions. In Chapter 3, we prove that the model selection procedure based on the minimum discrepancy principle provides a minimax optimal estimator (in the sense of Ineq. (1.18)) for choosing k in the k-NN regression function estimator while reducing the computational time compared to, for instance, AIC or Mallow's C_p criteria. It holds over any class of functions for which the minimax lower bound (1.17) is slower than $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$. This rate is the price to pay for knowing nothing about the bias term. This result only requires one a mild assumption on the regression function from Eq. (1.1): it should be bounded. Besides that, in Chapter 4, we provide less

technical arguments illustrating the optimality of MDP applied to other linear estimators such as the Nadaraya-Watson regression and variable selection estimators in the nonparametric regression model. More precisely, we carry out an extensive simulation study of the minimum discrepancy stopping rule applied to the tuning parameter selection problem with these linear estimators. What we can say is that, in most of the cases, the proposed strategy performs comparably well to other typically used ones in practice, among which the Hold-out method described above, Mallows' C_p (1.49), and generalized cross-validation (1.51).

Chapter 2

Early stopping and polynomial smoothing

Abstract

In this chapter, we study the problem of early stopping for iterative learning algorithms in reproducing kernel Hilbert space (RKHS) in the nonparametric regression framework. In particular, we work with the gradient descent and (iterative) kernel ridge regression algorithms. We present a *data-driven* rule to perform early stopping without a validation set that is based on the so-called minimum discrepancy principle. This method enjoys only one assumption on the regression function: it belongs to a reproducing kernel Hilbert space (RKHS). The proposed rule is proved to be minimax optimal over different types of kernel spaces, including finite-rank and Sobolev smoothness classes. The proof is derived from the fixed-point analysis of the localized Rademacher complexities, which is a standard technique for obtaining optimal rates in the nonparametric regression literature. In addition to that, we present simulations results on artificial datasets that show the comparable performance of the designed rule with respect to other stopping rules such as the one determined by $V-$fold cross-validation.

2.1 Introduction

The present chapter is concerned with nonparametric regression by means of a reproducing kernel Hilbert space (RKHS) associated with a reproducing kernel [11, 66, 98, 112]. There is a large amount of literature on the application of kernel machines in many areas of science and engineering [98, 101, 124], which is out of the main scope here.

A family of linear estimators called *spectral filter estimators* [15, 19, 60, 122] can be seen as particular instances of iterative learning algorithms. This family includes two famous examples: gradient descent and iterative (Tikhonov) ridge regression. In several papers, it was observed empirically and proved theoretically that these two algorithms are closely related [3, 4, 61, 92, 122]. For example, [3] showed that in the linear regression model, under the calibration $t = 1/\lambda$, where t is the time parameter in gradient descent and λ the tuning parameter in ridge regression, the risk error of gradient descent could not be much higher than that of ridge regression. It gives some intuition of why the idea of implicit regularization could work.

Early stopping rule (ESR) is a form of regularization that consists in choosing when to stop an iterative algorithm based on some design criterion. Its main idea is lowering the computational complexity of an iterative algorithm while preserving statistical optimality. This approach is quite old and initially was developed for Landweber iterations to solve ill-posed matrix problems in the 1970s [60, 113]. The next wave of interest in this topic was in the 1990s and has been applied to neural network parameters learning with stochastic gradient descent [46, 91]. For instance, [91] suggested some heuristics that rely on monitoring the train and validation errors for stopping the learning process, and gave some consistent simulation findings. Nevertheless, until the 2000s there was a lack of theoretical understanding of this phenomenon. Recent papers provided some insights for the connection between early stopping and boosting methods [17, 39, 118, 126], gradient descent, and Tikhonov regularization in reproducing kernel Hilbert space (RKHS) [19, 92, 122]. For instance, [39] established the first optimal in-sample convergence rate of L^2-boosting with early stopping. Raskutti et al. [92] provided a result on a stopping rule that achieves the minimax-optimal rate for kernelized gradient descent and ridge regression over different smoothness classes. This work established an important connection between the localized Rademacher complexities [16, 72, 114], that characterizes the size of the explored function space, and early stopping. The main drawback of the result is that one needs to know the RKHS-norm of the regression function or its tight upper bound to apply this early stopping rule in practice. Besides that, this rule is design-dependent, which limits its practical application as well. The subsequent work [118] showed how to control early stopping optimality via the localized Gaussian complexities in RKHS for different boosting algorithms (L^2-boosting, LogitBoost, and AdaBoost). Another theoretical result for a not data-driven ESR was built by [31], where Blanchard et al. proved a minimax optimal (in the $L_2(\mathbb{P}_X)$ out-of-sample norm) stopping rule for conjugate gradient descent in the nonparametric regression setting. Angles et al. [4] proposed a different approach, focusing on both time/memory computational savings combining early stopping with Nyström subsampling technique.

Some stopping rules that could be applied in practice were provided by [28, 30, 105] and developed on the so-called *minimum discrepancy principle* [31, 34, 60, 69]. This principle consists in monitoring the empirical risk and determining the first iteration at which a given learning algorithm starts to fit the noise. In the papers mentioned, the authors considered spectral filter estimators such as gradient descent, Tikhonov (ridge) regularization, and spectral cut-off regression for the linear Gaussian sequence model and derived several oracle inequalities for the proposed ESR. The main deficiency of the works [28, 30, 105] is that the authors dealt only with the linear Gaussian sequence model, and the minimax optimality result was restricted to the spectral cut-off estimator. It is worth to mention that [105] introduced the so-called *polynomial smoothing* strategy to achieve adaptivity of the minimum discrepancy principle ESR over Sobolev balls for the spectral cut-off estimator. More recently, Celisse and Wahl [49] studied a minimum discrepancy principle stopping rule and its modified version, where they provided the range of values of the regression function regularity, for which these stopping rules are optimal for different spectral filter estimators in RKHS.

Contribution. Hence, to the best of our knowledge, there is no *fully data-driven* stopping rule for gradient descent or ridge regression in RKHS that does not use a validation set, not depend on the parameters of the model, such as the RKHS-norm of the regression function, and explains why it is statistically optimal. Here, we combine techniques from [28], [92], and [105] to construct such an ESR. Our analysis relies on the bias and variance trade-off of an estimator, and we try to catch the iteration of their intersection by means of the *minimum discrepancy principle* [28, 34, 49] and the *localized Rademacher complexities* [16, 72, 82, 114]. In particular, for the kernels with infinite rank, we propose using a special technique [34, 105] for the empirical risk to reduce its variance. Further, we introduce new notions of *smoothed empirical Rademacher complexity* and *smoothed critical radius* in order to achieve minimax optimality bounds for the functional estimator based on the proposed rule. This can be done by solving the associated fixed-point equation. It implies that the bounds in our analysis cannot be improved (up to numeric constants). It is important to note that in the present chapter, we establish an important connection between a smoothed version of the *statistical dimension* of the kernel matrix, introduced by [121] for randomized projections in kernel ridge regression, with early stopping (see Section 2.4.3 for more details). We show also how to estimate the noise variance of the regression model, specifically, for the class of polynomial eigenvalue decay kernels. In the meanwhile, we provide experimental results on artificial data indicating the consistent performance of the proposed rules.

Outline of the chapter. The organization of the chapter is as follows. In Section 2.2, we introduce the background on the nonparametric regression and reproducing kernel Hilbert space. There, we explain the updates of two spectral filter iterative algorithms: gradient descent and (iterative) kernel ridge regression that will be studied. In Section 2.3, we clarify how to compute our first early stopping rule for finite-rank kernels and provide an oracle-type inequality (Theorem 2.3.1), and an upper bound for the risk error of this stopping rule with fixed covariates (Corollary 2.3.2). After that, we present a similar upper bound for the risk error with random covariates (Theorem 2.3.4) that is proven to be minimax-rate optimal. By contrast, Section 2.4 is devoted to the development of a new stopping rule for infinite-rank kernels based on the *polynomial smoothing* [34, 105] strategy. There, Theorem 2.4.1 shows, under some quite general assumptions on the eigenvalues of the kernel matrix, a high probability upper bound for the performance of this stopping rule measured in the $L_2(\mathbb{P}_n)$ in-sample norm. In particular, this upper bound leads to minimax optimality over Sobolev smoothness classes. In Section 2.5, we compare our stopping rules to other rules, such as methods using hold-out data and $V-$fold cross-validation. After that, we propose using a strategy for the estimation of the variance σ^2 of the regression model. Section 2.6 summarizes the content of the chapter and describes some perspectives. Supplementary and more technical proofs are deferred to Appendix.

2.2 Nonparametric regression and reproducing kernel framework

2.2.1 Probabilistic model and notation

The context of the present work is that of nonparametric regression, where an i.i.d. sample $\{(x_i, y_i),\ i = 1, \ldots, n\}$ of cardinality n is given, with $x_i \in \mathcal{X}$ (feature space) and $y_i \in \mathbb{R}$. The goal is to estimate the regression function $f^* : \mathcal{X} \to \mathbb{R}$ from the model

$$y_i = f^*(x_i) + \overline{\varepsilon}_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where the error variables $\overline{\varepsilon}_i$ are i.i.d. zero-mean Gaussian random variables $\mathcal{N}(0, \sigma^2)$, with $\sigma > 0$. In all what follows (except for Section 2.5, where results of empirical experiments are reported), the values of σ^2 is assumed to be known, as in [92] and [118].

Along the chapter, calculations are mainly derived in the *fixed-design* context, where the $\{x_i\}_{i=1}^n$ are assumed to be fixed, and only the error variables $\{\overline{\varepsilon}_i\}_{i=1}^n$ are random. In this context, the performance of any estimator $\widehat{f}$ of the regression function f^* is measured in terms of the so-called *empirical norm*, that is, the $L_2(\mathbb{P}_n)$ norm defined by

$$\|\widehat{f} - f^*\|_n^2 \coloneqq \frac{1}{n} \sum_{i=1}^n \Big[\widehat{f}(x_i) - f^*(x_i)\Big]^2,$$

where $\|h\|_n \coloneqq \sqrt{1/n \sum_{i=1}^n h(x_i)^2}$ for any bounded function h over $\mathcal{X}$, and $\langle \cdot, \cdot \rangle_n$ denotes the related inner-product defined by $\langle h_1, h_2 \rangle_n \coloneqq 1/n \sum_{i=1}^n h_1(x_i) h_2(x_i)$, for any functions h_1 and h_2 bounded over $\mathcal{X}$. In this context, $\mathbb{P}_\varepsilon$ and $\mathbb{E}_\varepsilon$ respectively denote the probability and expectation with respect to the $\{\overline{\varepsilon}_i\}_{i=1}^n$.

By contrast, Section 2.3.1 discusses some extensions of the previous results to the *random design* context, where both the covariates $\{x_i\}_{i=1}^n$ and responses $\{y_i\}_{i=1}^n$ are random variables. In this random design context, the performance of an estimator $\widehat{f}$ of f^* is measured in terms of the $L_2(\mathbb{P}_X)$-norm defined by

$$\|\widehat{f} - f^*\|_2^2 \coloneqq \mathbb{E}_X\Big[(\widehat{f}(X) - f^*(X))^2\Big],$$

where $\mathbb{P}_X$ denotes the probability distribution of the $\{x_i\}_{i=1}^n$. In what follows, $\mathbb{P}$ and $\mathbb{E}$ respectively state for the probability and expectation with respect to the couples $\{(x_i, y_i)\}_{i=1}^n$.

Notation Throughout the chapter, $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ are the usual Euclidean norm and inner product in $\mathbb{R}^n$. We shall write $a_n \lesssim b_n$ whenever $a_n \leq C b_n$ for some numeric constant $C > 0$ for all $n \geq 1$. $a_n \gtrsim b_n$ whenever $a_n \geq C b_n$ for some numeric constant $C > 0$ and all $n \geq 1$. Similarly, $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \gtrsim a_n$. $a \wedge b$ means $\min\{a, b\}$, and $a \vee b$ signifies $\max\{a, b\}$. $[M] \equiv \{1, \ldots, M\}$ for any $M \in \mathbb{N}$. For $a \geq 0$, we denote by $\lfloor a \rfloor$ the largest natural number that is smaller than or equal to a. We denote by $\lceil a \rceil$ the smallest natural number that is greater than or equal to a. Throughout the

chapter, we use the notation $c, c_1, C, \widetilde{c}, \widetilde{C}, \ldots$ to show that the numeric constants $c, c_1, C, \widetilde{c}, \widetilde{C}, \ldots$ do not depend on the parameters considered. The values of the constants may change from line to line.

2.2.2 Statistical model and assumptions

Reproducing Kernel Hilbert Space (RKHS)

Let us start by introducing a reproducing kernel Hilbert space (RKHS) denoted by $\mathcal{H}$ [11, 23]. Such an RKHS $\mathcal{H}$ is a class of functions associated with a *reproducing kernel* $\mathbb{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and endowed with an inner-product denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and satisfying $\langle \mathbb{K}_x, \mathbb{K}_y \rangle_{\mathcal{H}} = \mathbb{K}(x, y)$ for all $x, y \in \mathcal{X}$. Each function within $\mathcal{H}$ admits a representation as an element of $L_2(\mathbb{P}_X)$, which justifies the slight abuse when writing $\mathcal{H} \subset L_2(\mathbb{P}_X)$ (see [54] and [49, Assumption 3]).

Assuming the RKHS $\mathcal{H}$ is separable, Mercer's theorem [98] guarantees that the kernel can be expanded as

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_k(x'), \quad \forall x, x' \in \mathcal{X},$$

where $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ and $\{\phi_k\}_{k=1}^{\infty}$ are, respectively, the eigenvalues and corresponding eigenfunctions of the kernel integral operator T_k given by

$$T_k(f)(x) = \int_{\mathcal{X}} \mathbb{K}(x, u) f(u) d\mathbb{P}_X(u), \quad \forall f \in \mathcal{H}, \ x \in \mathcal{X}. \tag{2.2}$$

It is then known that the family $\{\phi_k\}_{k=1}^{\infty}$ is an orthonormal basis of $L_2(\mathbb{P}_X)$, while $\{\sqrt{\mu_k}\phi_k\}_{k=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}$. Then, any function $f \in \mathcal{H} \subset L_2(\mathbb{P}_X)$ can be expanded as

$$f = \sum_{k=1}^{\infty} \sqrt{\mu_k} \theta_k \phi_k,$$

where the coefficients $\{\theta_k\}_{k=1}^{\infty}$ are given by

$$\theta_k = \langle f, \sqrt{\mu_k}\phi_k \rangle_{\mathcal{H}} = \frac{1}{\sqrt{\mu_k}} \langle f, \phi_k \rangle_{L_2(\mathbb{P}_X)} = \int_{\mathcal{X}} \frac{f(x)\phi_k(x)}{\sqrt{\mu_k}} d\mathbb{P}_X(x). \tag{2.3}$$

Therefore, each functions $f, g \in \mathcal{H}$ can be represented by the respective sequences $\{a_k\}_{k=1}^{\infty}, \{b_k\}_{k=1}^{\infty} \in \ell_2(\mathbb{N})$ such that

$$f = \sum_{k=1}^{+\infty} a_k \phi_k \quad \text{and} \quad g = \sum_{k=1}^{+\infty} b_k \phi_k,$$

with the inner-product in the Hilbert space $\mathcal{H}$ given by

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \frac{a_k b_k}{\mu_k}.$$

This leads to the following representation of $\mathcal{H}$ as an ellipsoid

$$\mathcal{H} = \left\{ f = \sum_{k=1}^{+\infty} a_k \phi_k, \quad \sum_{k=1}^{+\infty} a_k^2 < +\infty \text{ and } \sum_{k=1}^{+\infty} \frac{a_k^2}{\mu_k} < +\infty \right\}.$$

Main assumptions

From the initial model given by Eq. (2.1), we make the following assumption.

Assumption 1 (Statistical model)**.** *Let $\mathbb{K}(\cdot,\cdot)$ denote a reproducing kernel as defined above and $\mathcal{H}$ the induced separable RKHS. Then, there exists a constant $R > 0$ such that the n-sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X}^n \times \mathbb{R}^n$ satisfies the statistical model*

$$y_i = f^*(x_i) + \overline{\varepsilon}_i, \quad \textit{with} \quad f^* \in \mathbb{B}_{\mathcal{H}}(R) = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}, \tag{2.4}$$

where the $\{\overline{\varepsilon}_i\}_{i=1}^n$ are i.i.d. Gaussian random variables with $\mathbb{E}[\overline{\varepsilon}_i \mid x_i] = 0$ and $\mathbb{V}[\overline{\varepsilon}_i \mid x_i] = \sigma^2$.

The model from Assumption 1 can be vectorized as

$$Y = [y_1, ..., y_n]^\top = F^* + \overline{\varepsilon} \in \mathbb{R}^n, \tag{2.5}$$

where $F^* = [f^*(x_1), \ldots, f^*(x_n)]^\top$ and $\overline{\varepsilon} = [\overline{\varepsilon}_1, \ldots, \overline{\varepsilon}_n]^\top$, which turns to be useful all along the chapter. Let us emphasize that Assumption 1 encapsulates a (mild) smoothness assumption about f^* encoded by the specification of the reproducing kernel $\mathbb{K}(\cdot,\cdot)$. For instance, this affects the convergence rates one can achieve [93]. More precisely, from the kernel operator T_k (2.2), that is self-adjoint and trace-class, the smoothness of f^* can be quantified by means of a so-called *source condition* expressed as

$$f^* = T_k^s u \quad \text{with} \quad u \in L_2(\mathbb{P}_X), \ \|u\|_2 \leq \rho, \tag{2.6}$$

where $s > 0$ and $\rho > 0$ are constants. For instance, assuming $s \geq \frac{1}{2}$ is equivalent to requiring $f^* \in \mathcal{H}$. See also [49, Assumption 3] for a deeper discussion about the source condition.

Examples of celebrated reproducing kernels that are used in practice include the Gaussian RBF kernel [9, Section 3.2], the Sobolev kernel [92], polynomial kernels of degree d [121], ... For more examples, see [62, 98, 112].

In the present chapter, we make a boundness assumption on the reproducing kernel $\mathbb{K}(\cdot,\cdot)$.

Assumption 2. *Let us assume that the reproducing kernel $\mathbb{K}(\cdot,\cdot)$ is uniformly bounded on its support, meaning that there exists a constant $B > 0$ such that*

$$\sup_{x \in \mathcal{X}} \Big[\mathbb{K}(x,x) \Big] = \sup_{x \in \mathcal{X}} ||\mathbb{K}_x||_{\mathcal{H}}^2 \leq B.$$

Moreover, in what follows, we assume that $B = 1$ without loss of generality.

Assumption 2 holds true for many kernels. On the one hand, it is fulfilled with an unbounded domain $\mathcal{X}$ for a bounded kernel (e.g. the Gaussian, Laplace kernels). On the other hand, it amounts to assume that the domain $\mathcal{X}$ is bounded with an unbounded kernel such as the polynomial or Sobolev kernels [98]. Let us also mention that Assumptions 1 and 2 (combined with the reproducing property) imply that f^* is uniformly bounded since

$$\|f^*\|_\infty = \sup_{x\in\mathcal{X}} |\langle f^*, \mathbb{K}_x\rangle_{\mathcal{H}}| \leq \|f^*\|_{\mathcal{H}} \sup_{x\in\mathcal{X}} \|\mathbb{K}_x\|_{\mathcal{H}} \leq R. \tag{2.7}$$

Considering now the Gram matrix $K = \{\mathbb{K}(x_i, x_j)\}_{1\leq i,j\leq n}$, the related *normalized Gram matrix* $K_n = \{\mathbb{K}(x_i, x_j)/n\}_{1\leq i,j\leq n}$ turns out to be symmetric and positive semidefinite. This entails the existence of the empirical eigenvalues $\widehat{\mu}_1, \ldots, \widehat{\mu}_n$ (respectively, the eigenvectors $\widehat{u}_1, \ldots, \widehat{u}_n$) such that $K_n\widehat{u}_i = \widehat{\mu}_i \cdot \widehat{u}_i$ for all $i \in [n]$. Let us further assume that the rank of K_n satisfies $\mathrm{rk}(K_n) = r \leq n$ with

$$\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \ldots \geq \widehat{\mu}_r > 0 = \widehat{\mu}_{r+1} = \ldots = \widehat{\mu}_n = 0.$$

Remark that Assumption 2 implies $0 \leq \max(\widehat{\mu}_1, \mu_1) \leq 1$.

For technical convenience, it turns out to be useful rephrasing the model (2.5) by using the SVD of the normalized Gram matrix K_n. This leads to the new (rotated) model

$$Z_i = \langle \widehat{u}_i, Y\rangle = G_i^* + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.8}$$

where $G_i^* = \langle \widehat{u}_i, F^*\rangle$, and $\varepsilon_i = \langle \widehat{u}_i, \bar{\varepsilon}\rangle$ is a zero-mean Gaussian random variable with variance σ^2.

2.2.3 Spectral filter algorithms

Spectral filter algorithms were first introduced for solving ill-posed inverse problems with deterministic noise [60]. Among others, one typical example of such an algorithm is the gradient descent algorithm (that is named L^2-boosting [39] as well). They were more recently brought to the supervised learning community, for instance, by [19, 43, 64, 122]. For estimating the vector F^* from Eq. (2.5) in the fixed-design context, such a spectral filter algorithm is a linear estimator, which can be expressed as

$$F^\lambda := \left(f^\lambda(x_1), \ldots, f^\lambda(x_n)\right)^\top = g_\lambda(K_n)K_nY, \tag{2.9}$$

where $g_\lambda : [0, 1] \to \mathbb{R}$ is called the *spectral filter function* and is defined as follows.

Definition 2.2.1 (see, e.g., [64])**.** $\lambda \mapsto g_\lambda$ is called an admissible spectral filter function if it is continuous, non-increasing, and obeys the next four conditions:

1. There exists $\overline{B} > 0$ such that $\sup_{0<\xi\leq 1} |g_\lambda(\xi)| \leq \frac{\overline{B}}{\lambda}, \quad \forall\lambda \in [0, +\infty)$.

2. For all $\xi \in (0,1]$, $\lim_{\lambda \to 0}[\xi g_\lambda(\xi)] = 1$.
3. There exists $\overline{D} > 0$ such that $\sup_{0<\xi\leq 1} |\xi g_\lambda(\xi)| \leq \overline{D}, \quad \forall \lambda \in [0, +\infty)$.
4. There exists $\bar{\nu} > 0$ called the *qualification* of g_λ, and a constant $C_\nu > 0$ independent of λ such that

$$\sup_{0<\xi\leq 1} |1 - \xi g_\lambda(\xi)|\xi^\nu \leq C_\nu \lambda^\nu, \quad \forall\, 0 < \nu \leq \bar{\nu}. \tag{2.10}$$

The choice $g_\lambda(\xi) = \frac{1}{\xi+\lambda}$, which corresponds to the kernel ridge estimator with the regularization parameter $\lambda > 0$, is an admissible spectral filter function with $\overline{B} = \overline{D} = 1$, where qualification Ineq. (2.10) holds with $C_\nu = 1$ for $0 < \nu \leq 1 = \bar{\nu}$ (see [28, 49] for other possible choices).

From the model expressed in the empirical eigenvectors basis (2.8), the resulting spectral filter estimator (2.9) can be expressed as

$$G_i^{\lambda_t} = \langle \widehat{u}_i, F^{\lambda_t} \rangle = \gamma_i^{(t)} Z_i, \quad \forall i = 1, \ldots, n, \tag{2.11}$$

where $t \mapsto \lambda_t > 0$ is a decreasing function mapping t to a regularization parameter value at time t, and $t \mapsto \gamma_i^{(t)}$ is defined by

$$\gamma_i^{(t)} = \widehat{\mu}_i g_{\lambda_t}(\widehat{\mu}_i), \quad \forall i = 1, \ldots, n.$$

From Definition 2.2.1, it can be proved that $\gamma_i^{(t)}$ is a non-decreasing function of t, $\gamma_i^{(0)} = 0$, and $\lim_{t\to\infty} \gamma_i^{(t)} = 1$. Moreover, $\widehat{\mu}_i = 0$ implies $\gamma_i^{(t)} = 0$ as it is the case for kernels with a finite rank, that is, when $\mathrm{rk}(K_n) = r < n$.

Thanks to the remark above, we define the following convenient notations: $f^t := f^{\lambda_t}$ (for the functions) and $F^t := F^{\lambda_t}$ (for the vectors), with a continuous iteration (time) $t > 0$.

In what follows, we introduce an assumption on $\gamma_i^{(t)}$ function that will play a crucial role in our analysis.

Assumption 3.

$$c \min\{1, \eta t \widehat{\mu}_i\} \leq \gamma_i^{(t)} \leq \min\{1, \eta t \widehat{\mu}_i\},$$

for any $i = 1, \ldots, n$, *some positive constants* $c \in (0, 1)$, *and* $\eta > 0$.

Let us mention two famous examples of spectral filter estimators that satisfy Assumption 3 with $c = 1/2$ (see Lemma 2.7.2 in Appendix). These examples will be further studied in the present chapter.

— Gradient descent (GD) with a constant step-size $0 < \eta \leq 1/\widehat{\mu}_1$:

$$\gamma_i^{(t)} = 1 - (1 - \eta\widehat{\mu}_i)^t, \quad \forall t > 0,\ \forall i = 1, \ldots, n. \tag{2.12}$$

Note that GD satisfies the qualification condition (2.10) with arbitrary $\bar{\nu} > 0$ (see e.g. [19] for more discussion on the qualification). The constant step-size η can be replaced by any non-increasing sequence $\{\eta_t\}_{t=0}^{+\infty}$ satisfying [92]

— $(\widehat{\mu}_1)^{-1} \geq \eta_t \geq \eta_{t+1} \geq \ldots$, for $t = 0, 1, \ldots$,
— $\sum_{s=0}^{t-1} \eta_s \to +\infty$ as $t \to +\infty$.

— Kernel ridge regression (KRR) with the regularization parameter $\lambda_t = 1/(\eta t)$ with $\eta > 0$:

$$\gamma_i^{(t)} = \frac{\widehat{\mu}_i}{\widehat{\mu}_i + \lambda_t}, \quad \forall t > 0, \ \forall i = 1, \ldots, n. \tag{2.13}$$

The linear parameterization $\lambda = 1/(\eta t)$ is chosen for theoretical convenience and could be replaced by any alternative choice, such as the exponential parameterization $\lambda = 1/(e^{\eta t} - 1)$.

We refer interested readers, for instance, to [92, Sections 4.1 and 4.4] for the derivation of the $\gamma_i^{(t)}$ expressions. The expressions of the two above examples have been derived from $F^0 = [f^0(x_1), \ldots, f^0(x_n)]^\top = [0, \ldots, 0]^\top$ as an initialization condition without loss of generality.

2.2.4 Reference stopping rule and oracle-type inequality

From a set of iterations $t \in \mathcal{T} := \{0, \ldots, T\}$ for an iterative learning algorithm (like the spectral filter described in Section 2.2.3), the present goal is to design $\widehat{t} = \widehat{t}(\{x_i, y_i\}_{i=1}^n)$ from the data $\{x_i, y_i\}_{i=1}^n$ such that the functional estimator $f^{\widehat{t}}$ is as close as possible to the optimal one among $\mathcal{T}$.

Numerous classical model selection procedures for choosing $\widehat{t}$ already exist, e.g. the (generalized) cross validation [111], AIC and BIC criteria [1, 100], the unbiased risk estimation [47], or Lepski's balancing principle [81]. Their main drawback in the present context is that they require the practitioner to calculate all the estimators $\{f^t, \ t \in \mathcal{T}\}$, in a first step, and then choose the optimal estimator among the candidates in a second step, which can be high computationally demanding.

By contrast, early stopping is a less time-consuming approach. It is based on observing one estimator at each iteration $t \in \mathcal{T}$ and deciding to stop the learning process according to some criterion. Its aim is to reduce the computational cost, induced by this selection procedure while preserving the statistical optimality properties of the output estimator.

The prediction error (risk) of an estimator f^t at iteration t is split into a bias and a variance term as

$$R(t) = \mathbb{E}_\varepsilon \|f^t - f^*\|_n^2 = \|\mathbb{E}_\varepsilon f^t - f^*\|_n^2 + \mathbb{E}_\varepsilon \|f^t - \mathbb{E}_\varepsilon f^t\|_n^2 = B^2(t) + V(t),$$

with

$$B^2(t) = \frac{1}{n} \sum_{i=1}^n (1 - \gamma_i^{(t)})^2 (G_i^*)^2, \qquad V(t) = \frac{\sigma^2}{n} \sum_{i=1}^n (\gamma_i^{(t)})^2. \tag{2.14}$$

From the properties of Definition 2.2.1, the bias term is a non-increasing function of t converging to zero, while the variance term is a non-decreasing function of t converging to $\frac{r\sigma^2}{n}$ ($\mathrm{rk}(K_n) = r$). Since minimizing the risk as a function of t cannot be achieved, the empirical risk R_t (that measures the

size of the residuals) is introduced with the notation of Eq. (2.8).

$$R_t = \frac{1}{n}\sum_{i=1}^{n}(1-\gamma_i^{(t)})^2 Z_i^2 = \frac{1}{n}\sum_{i=1}^{r}(1-\gamma_i^{(t)})^2 Z_i^2 + \frac{1}{n}\sum_{i=r+1}^{n} Z_i^2, \tag{2.15}$$

This is a non-increasing function of t, which measures how well an estimator f^t fits the data (or equivalently, how much information is still contained within the residuals).

An illustration of the typical behavior of the risk, empirical risk, bias, and variance is displayed by Figure 2.1. The risk achieves its (global) minimum at $t \approx 1000$. Making additional iterations will eventually lead to the waste of the computational resources and worsen the statistical performance, which empirically justifies the need for a data-driven early stopping rule.

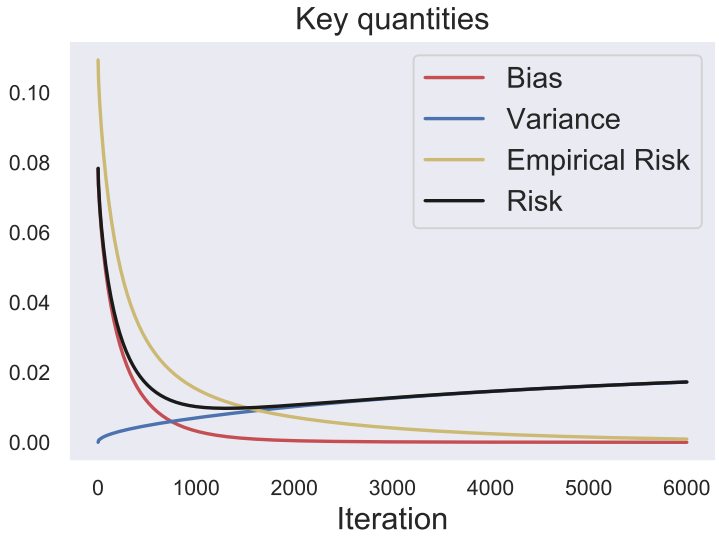


Figure 2.1 – The bias, variance, risk, and empirical risk behavior.

Let us now introduce our "reference stopping rule". This stopping rule balances the bias and variance described above, which is a common strategy for model selection in the nonparametric statistics literature since it usually yields a minimax optimal estimator (see, e.g. [108]). This reference stopping rule is defined as the first time the bias term becomes smaller than or equal to the variance term, that is,

$$t^b = \inf\{t > 0 \mid B^2(t) \leq V(t)\}. \tag{2.16}$$

This is a purely theoretical stopping rule since it strongly depends on unknown quantities. However, its main interest lies in the way it compares with the global optimum performance, that is, with the oracle performance. This is the purpose of the next lemma.

Lemma 2.2.1. *Under the monotonicity of the bias and variance terms,*

$$\mathbb{E}_\varepsilon\|f^{t^b} - f^*\|_n^2 \leq 2 \inf_{t>0}\Big[\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2\Big].$$

Proof of Lemma 2.2.1. The proof is quite simple and can be deduced from [28, p.8]. For any $t > 0$,

$$B^2(t) + V(t) \geq \min\{B^2(t^b), V(t^b)\} = \frac{1}{2}\Big[B^2(t^b) + V(t^b)\Big] = \frac{1}{2}\mathbb{E}_\varepsilon\|f^{t^b} - f^*\|_n^2.$$

To finish the proof, it is sufficient to take $t = \operatorname{argmin}_{t>0}\Big[\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2\Big]$. ■

This lemma provides a fundamental result that guarantees the optimality of t^b for an iterative estimator, for which the bias is a non-increasing function of t, and the variance is a non-decreasing function of t. It also implies that the risk of any spectral filter estimator computed at t^b cannot be higher than 2 times the risk of the oracle rule. This is the main reason for considering t^b as a reference stopping rule in our analysis. It is also worth mentioning that even if we knew $B^2(t)$ for all $t \leq t_1$ for some $t_1 > 0$, the bias could still suddenly drop after at time $t_2 > t_1$. Stopping at t_1 could then result in a much worse performance than stopping at time t_2, where the bias term is zero. This remark suggests that recovering the oracle performance cannot be achieved in full generality in the present framework, where one has access to a limited number of "observations" of the risk curve. This is why the balancing stopping rule t^b plays the role of a reference stopping rule – its performance can nevertheless be linked with the one of the oracle stopping rule.

Our main concern is formulating a data-driven stopping rule (a mapping from the data $\{(x_i, y_i)\}_{i=1}^n$ to positive time $\widehat{t}$) so that the prediction errors $\mathbb{E}_\varepsilon\|f^{\widehat{t}} - f^*\|_n^2$ or, equivalently, $\mathbb{E}\|f^{\widehat{t}} - f^*\|_2^2$ are as small as possible. A classical tool commonly used in model selection for quantifying the performance of a procedure is the oracle-type inequality [47, 72, 108, 114]. In the fixed design context, an oracle inequality (in expectation) can be formulated as follows

$$\mathbb{E}_\varepsilon\|f^{\widehat{t}} - f^*\|_n^2 \leq C_n \inf_{t\in\mathcal{T}}\Big[\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2\Big] + r_n, \tag{2.17}$$

where the bounded constant $C_n \geq 1$ on the right hand side can depend on various parameters of the problem (except f^*). The main term $\inf_{t\in\mathcal{T}}\Big[\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2\Big]$ is the best possible performance any estimator among $\{f^t,\ t \in \mathcal{T}\}$ can achieve. Ideally, for the oracle inequality to be meaningful, the last term r_n on the right hand side should be negligible compared to the oracle performance.

2.2.5 Localized empirical Rademacher complexity

The analysis of the forthcoming early stopping rules involves the use of a model complexity measure known as the *localized empirical Rademacher complexity* [16, 72, 114].

Definition 2.2.2. For any given $\epsilon > 0$ and function class $\mathcal{F}$, consider the localized empirical Rademacher complexity

$$\widehat{\mathcal{R}}_n(\epsilon, \mathcal{F}) = \mathbb{E}_{\mathbf{r}} \left[\sup_{\substack{f \in \mathcal{F} \\ \|f\|_n \leq \epsilon R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{r}_i f(x_i) \right| \right], \tag{2.18}$$

where $\{\mathbf{r}_i\}_{i=1}^n$ are i.i.d. Rademacher variables ($\{-1, +1\}$-random variables with equal probability $\frac{1}{2}$).

Usually, the localized empirical Rademacher complexity is defined for $R = 1$, but due to the scaling factor of $\|f^*\|_{\mathcal{H}}$, one needs to consider the radius ϵR within the supremum.

Along the analysis, more explicit lower and upper bounds on the above localized empirical Rademacher complexity have to be derived. This is the purpose of introducing the so-called *kernel complexity function* [82, 83] that is proved to be of the same size (up to numeric constants) as the localized empirical Rademacher complexity of $\mathcal{F} = \mathbb{B}_{\mathcal{H}}(R)$, that is,

$$\widehat{\mathcal{R}}_n(\epsilon, \mathcal{H}) = R \left[\frac{1}{n} \sum_{j=1}^{r} \min\{\epsilon^2, \widehat{\mu}_j\} \right]^{1/2}. \tag{2.19}$$

It corresponds to a rescaled sum of the empirical eigenvalues truncated at ϵ^2.

For a given RKHS $\mathcal{H}$ and noise level σ, let us finally define the *empirical critical radius* $\widehat{\epsilon}_n$ as the smallest positive value ϵ such that

$$\frac{\widehat{R}_n(\epsilon, \mathcal{H})}{\epsilon R} \leq \frac{2\epsilon R}{\sigma}. \tag{2.20}$$

There is an extensive literature on this empirical critical equation and the related empirical critical radius [16, 82, 92], and it is out of the scope of the chapter providing an exhaustive review on this topic. Nevertheless, it has been proved that $\widehat{\epsilon}_n$ does always exist and is unique. The constant 2 in Ineq. (2.20) is for theoretical convenience only.

2.3 Data-driven early stopping rule and minimum discrepancy principle

Let us start by recalling that the expression of the empirical risk in Eq. (2.15) gives that the empirical risk is a non-increasing function of t (as illustrated by Fig. 2.1 as well). This is consistent with the intuition that the amount of available information within the residuals decreases as the number of iterations grows. If there exists an iteration t such that $f^t \approx f^*$, then the empirical risk is approximately equal to σ^2 (level of noise), that is,

$$\mathbb{E}_\varepsilon R_t = \mathbb{E}_\varepsilon \left[\|F^t - Y\|_n^2 \right] \approx \mathbb{E}_\varepsilon \left[\|F^* - Y\|_n^2 \right] = \mathbb{E}_\varepsilon \left[\|\varepsilon\|_n^2 \right] = \sigma^2. \tag{2.21}$$

Additional iterations would result in fitting to noise (overfitting). Introducing, moreover, the reduced empirical risk $\widetilde{R}_t,\ t>0$, and recalling that r denotes the rank of the Gram matrix, it comes

$$\mathbb{E}_\varepsilon R_t = \mathbb{E}_\varepsilon\Big[\frac{1}{n}\sum_{i=1}^{n}(1-\gamma_i^{(t)})^2 Z_i^2\Big] = \underbrace{\mathbb{E}_\varepsilon\Big[\frac{1}{n}\sum_{i=1}^{r}(1-\gamma_i^{(t)})^2 Z_i^2\Big]}_{:=\widetilde{R}_t} + \frac{n-r}{n}\sigma^2 \overset{\text{(i)}}{\approx} \sigma^2, \tag{2.22}$$

where (i) is due to Eq. (2.21). This heuristic argument gives rise to a first deterministic stopping rule t^* involving the reduced empirical risk and given by

$$t^* = \inf\left\{t>0 \mid \mathbb{E}_\varepsilon \widetilde{R}_t \le \frac{r\sigma^2}{n}\right\}. \tag{2.23}$$

Since t^* is *not achievable* in practice, an estimator of t^* is given by the data-driven stopping rule τ based on the so-called minimum discrepancy principle (MDP)

$$\tau = \inf\left\{t>0 \mid \widetilde{R}_t \le \frac{r\sigma^2}{n}\right\}. \tag{2.24}$$

The existing literature considering the MDP stopping rule usually defines τ by the event $\{R_t \le \sigma^2\}$ [28, 31, 34, 60, 69, 105]. On the one hand, with a full-rank kernel ($r=n$), the reduced empirical risk $\widetilde{R}_t$ is equal to the classical empirical risk, leading then to the same stopping rule. On the other hand, with a finite-rank kernel ($r \ll n$), using the reduced empirical risk and the event $\{\widetilde{R}_t \le \frac{r\sigma^2}{n}\}$ rather than the empirical risk and $\{R_t \le \sigma^2\}$ should lead to a less variable stopping rule. From a practical perspective, the knowledge of the rank of the Gram matrix (which is exploited by the reduced empirical risk, unlike the classical empirical risk) avoids estimating the last $n-r$ components of the vector G^*, which are already known to be zero (see Appendix 2.7 for more details).

Intuitively, if the empirical risk is close to its expectation, then τ should be optimal in some sense. Therefore, the main theoretical analysis will concern quantifying how close τ and t^* are to each other. It appeared in practice that, if the model is quite simple, e.g. the kernel is of finite rank, or the variance σ^2 is low compared to the signal f^*, τ is close to t^* and τ performs well. As soon as the model becomes complex, e.g. an infinite-rank kernel, or the variance σ^2 is high compared to the signal f^*, τ, as a random variable, has high variance that should be reduced. Of course, the smoothness of the regression function should play a role too. This not rigorous statement will be further developed in Section 2.3.2.

2.3.1 Finite-rank kernels

Fixed-design framework

Let us start by discussing our results with the case of RKHS of finite-rank kernels with rank $r < n : \widehat{\mu}_i = 0,\ i > r$, and $\mu_i = 0,\ i > r$. Examples that include these kernels are the linear kernel $\mathbb{K}(x_1, x_2) = x_1^\top x_2$ and the polynomial kernel of degree $d \in \mathbb{N}$ $\mathbb{K}(x_1, x_2) = (1 + x_1^\top x_2)^d$. It is easy to show that the polynomial kernel is of finite rank at most $d + 1$, meaning that the kernel matrix K_n has at most $\min\{d + 1, n\}$ nonzero eigenvalues.

The following theorem applies to any functional sequence $\{f^t\}_{t=0}^\infty$ generated by (2.11) and initialized at $f^0 = 0$. The main part of the proof of this result consists of properly upper bounding $\mathbb{E}_\varepsilon|\mathbb{E}_\varepsilon\widetilde{R}_{t^*} - \widetilde{R}_{t^*}|$ and follows the same trend of [28, Proposition 3.1].

Theorem 2.3.1. *Under Assumptions 1 and 2, given the stopping rule (2.24),*

$$\mathbb{E}_\varepsilon\|f^\tau - f^*\|_n^2 \leq 2(1 + \theta^{-1})\mathbb{E}_\varepsilon\|f^{t^*} - f^*\|_n^2 + 2(\sqrt{3} + \theta)\frac{\sqrt{r}\sigma^2}{n}, \tag{2.25}$$

for any strictly positive θ.

Proof of Theorem 2.3.1. In this proof, we will use the following inequalities: for any $a, b \geq 0$: $(a-b)^2 \leq |a^2 - b^2|$, and $2ab \leq \theta a^2 + \frac{1}{\theta}b^2$ for $\forall\theta > 0$.

Let us first proof the subsequent oracle-type inequality for the difference between f^τ and f^{t^*}. Consider

$$\begin{aligned}
\|f^{t^*} - f^\tau\|_n^2 &= \frac{1}{n}\sum_{i=1}^r \left(\gamma_i^{(t^*)} - \gamma_i^{(\tau)}\right)^2 Z_i^2 \leq \frac{1}{n}\sum_{i=1}^r |(1 - \gamma_i^{(t^*)})^2 - (1 - \gamma_i^{(\tau)})^2| Z_i^2 \\
&= (\widetilde{R}_{t^*} - \widetilde{R}_\tau)\mathbb{I}\{\tau \geq t^*\} + (\widetilde{R}_\tau - \widetilde{R}_{t^*})\mathbb{I}\{\tau < t^*\} \\
&\leq (\widetilde{R}_{t^*} - \mathbb{E}_\varepsilon\widetilde{R}_{t^*})\mathbb{I}\{\tau \geq t^*\} + (\mathbb{E}_\varepsilon\widetilde{R}_{t^*} - \widetilde{R}_{t^*})\mathbb{I}\{\tau < t^*\} \\
&\leq |\widetilde{R}_{t^*} - \mathbb{E}_\varepsilon\widetilde{R}_{t^*}|.
\end{aligned}$$

From the definition of $\widetilde{R}_t$ (2.22), one notices that

$$|\widetilde{R}_{t^*} - \mathbb{E}_\varepsilon\widetilde{R}_{t^*}| = \left|\sum_{i=1}^r (1 - \gamma_i^{(t^*)})^2\Big[\frac{1}{n}(\varepsilon_i^2 - \sigma^2) + \frac{2}{n}\varepsilon_i G_i^*\Big]\right|.$$

From $\mathbb{E}_\varepsilon|X(\varepsilon)| \leq \sqrt{\mathrm{var}_\varepsilon X(\varepsilon)}$ for $X(\varepsilon)$ centered, and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, and $\mathbb{E}_\varepsilon(\varepsilon^4) \leq$

$3\sigma^4$, it comes

$$\begin{aligned}\mathbb{E}_\varepsilon|\widetilde{R}_{t^*} - \mathbb{E}_\varepsilon\widetilde{R}_{t^*}| &\leq \sqrt{\frac{2\sigma^2}{n^2}\sum_{i=1}^{r}(1-\gamma_i^{(t^*)})^4\left[\frac{3}{2}\sigma^2 + 2(G_i^*)^2\right]} \\ &\leq \sqrt{\frac{3\sigma^4}{n^2}\sum_{i=1}^{r}(1-\gamma_i^{(t^*)})^2} + \sqrt{\frac{4\sigma^2}{n^2}\sum_{i=1}^{r}(1-\gamma_i^{(t^*)})^2(G_i^*)^2} \\ &\leq \frac{\sqrt{3}\sigma^2\sqrt{r}}{n} + \theta\frac{\sigma^2}{n} + \theta^{-1}B^2(t^*) \\ &\leq \theta^{-1}B^2(t^*) + (\sqrt{3}+\theta)\frac{\sqrt{r}\sigma^2}{n}.\end{aligned}$$

Applying the inequalities $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \geq 0$, and $B^2(t^*) \leq \mathbb{E}_\varepsilon\|f^{t^*} - f^*\|_n^2$, we arrive at

$$\begin{aligned}&\mathbb{E}_\varepsilon\|f^\tau - f^*\|_n^2 \\ &\leq 2\mathbb{E}_\varepsilon\|f^{t^*} - f^*\|_n^2 + 2\mathbb{E}_\varepsilon\|f^\tau - f^{t^*}\|_n^2 \\ &\leq 2(1+\theta^{-1})\mathbb{E}_\varepsilon\|f^{t^*} - f^*\|_n^2 + 2(\sqrt{3}+\theta)\frac{\sqrt{r}\sigma^2}{n}.\end{aligned}$$

The claim is proved. ■

First of all, it is worth noting that the risk of the estimator f^{t^*} is proved to be *optimal* for gradient descent and kernel ridge regression no matter the kernel we use (see Appendix 2.9 for the proof), so it remains to focus on the remainder term on the right-hand side in Ineq. (2.25). Theorem 2.3.1 applies to any reproducing kernel, but one remarks that for infinite-rank kernels, $r \asymp n$ and we achieve only the rate $\mathcal{O}(1/\sqrt{n})$. This rate is suboptimal since, for instance, an RKHS with polynomial eigenvalue decay kernels (will be considered in the next subsection) has the minimax-optimal rate for the risk error of the order $\mathcal{O}\left(n^{-\frac{\beta}{\beta+1}}\right)$ with $\beta > 1$. Therefore, the oracle-type inequality (2.25) could be useful only for finite-rank kernels due to the fast $\mathcal{O}(\sqrt{r}/n)$ rate of the remainder term.

Notice that to make artificially the term $\mathcal{O}\left(\sqrt{r}/n\right)$ a remainder one (even for the cases corresponding to infinite-rank kernels), [28, 30] introduced in the definitions of their stopping rules a restriction on the "starting time" t_0. However, in the work mentioned, this restriction incurred the price of possibility to miss the designed time τ. For instance, [28] took t_0 as the first time at which the variance becomes of the order $\frac{\sqrt{r}\sigma^2}{n}(\sqrt{D}\delta^2$ in their notations). Besides that, [30] developed an additional procedure, built on standard model selection criteria such as the AIC-criterion, for the spectral cut-off estimator to recover the "missing" stopping rule and achieve adaptivity over Sobolev-type ellipsoids. In our work, we removed such a strong assumption.

As a corollary of Theorem 2.3.1, one can prove that f^τ provides a minimax estimator of f^* over the ball of radius R.

Corollary 2.3.2. *Under Assumptions 1, 2, 3, if a kernel has finite rank r, then*

$$\mathbb{E}_\varepsilon \|f^\tau - f^*\|_n^2 \le c_u R^2 \widehat{\epsilon}_n^2, \tag{2.26}$$

where constant c_u is numeric.

Proof of Corollary 2.3.2. From Theorem 2.3.1 and Lemma 2.9.1 in Appendix,

$$\mathbb{E}_\varepsilon \|f^\tau - f^*\|_n^2 \le 16(1+\theta^{-1}) R^2 \widehat{\epsilon}_n^2 + 2(\sqrt{3}+\theta)\frac{\sqrt{r}\sigma^2}{n}. \tag{2.27}$$

Further, from Lemma 2.7.4 in Appendix, $\widehat{\epsilon}_n^2 = c\frac{r\sigma^2}{nR^2}$ with a positive numeric constant c, and it implies that

$$\mathbb{E}_\varepsilon \|f^\tau - f^*\|_n^2 \le \left[16(1+\theta^{-1}) + \frac{2(\sqrt{3}+\theta)}{c}\right] R^2 \widehat{\epsilon}_n^2. \tag{2.28}$$

■

Note that the critical radius $\widehat{\epsilon}_n$ cannot be arbitrary small, since it should satisfy Ineq. (2.20). As it will be clarified later, the squared empirical critical radius is essentially optimal.

Random-design framework

We would like to transfer the minimax optimality bound for the estimator f^τ from the empirical $L_2(\mathbb{P}_n)$ norm to the $L_2(\mathbb{P}_X)$ in-sample norm by means of the so-called localized population Rademacher complexity. This complexity measure became a standard tool in empirical processes and nonparametric regression [16, 72, 92, 114].

For any kernel function class studied in the chapter, we consider the localized Rademacher complexity that can be seen as a population counterpart of the empirical Rademacher complexity (2.19) introduced earlier:

$$\overline{\mathcal{R}}_n(\epsilon, \mathcal{H}) = R\left[\frac{1}{n}\sum_{i=1}^{\infty} \min\{\mu_i, \epsilon^2\}\right]^{1/2}. \tag{2.29}$$

Using the localized population Rademacher complexity, we define its *population critical radius* $\epsilon_n > 0$ to be the smallest positive solution ϵ that satisfies the inequality

$$\frac{\overline{\mathcal{R}}_n(\epsilon, \mathcal{H})}{\epsilon R} \le \frac{2\epsilon R}{\sigma}. \tag{2.30}$$

In contrast to the empirical critical radius $\widehat{\epsilon}_n$, this quantity is not data-dependent since it is specified by the population eigenvalues of the kernel operator T_k underlying the RKHS.

Recall the definition of the population critical radius (2.30), then the following result provides a fundamental lemma on the transfer between the $L_2(\mathbb{P}_n)$ and $L_2(\mathbb{P}_X)$ functional norms. In what follows, we assume that $\mathcal{H}$ is a star-shaped function class, meaning that for any $f \in \mathcal{H}$ and scalar $\omega \in [0, 1]$,

the function ωf belongs to $\mathcal{H}$. The assumption on $\mathcal{H}$ being star-shape holds if f is assumed to lie in the $\mathcal{H}$-norm ball of an *arbitrary* finite radius.

Lemma 2.3.3. *[114, Theorem 14.1] Assume a star-shaped kernel function class $\mathcal{H}$ and Assumption 2 of the bounded kernel. Let ϵ_n be as in Ineq. (2.30), then for any $f \in \mathbb{B}_{\mathcal{H}}(cR)$, where $c > 1$ is a numeric constant, and $h \geq \epsilon_n$, one has*

$$\left| \|f\|_n^2 - \|f\|_2^2 \right| \leq \frac{1}{2}\|f\|_2^2 + c_1 R^2 h^2 \tag{2.31}$$

with probability at least $1 - c_2 e^{-c_3 \frac{nh^2R^2}{\sigma^2}}$, for some positive numeric constants c_1, c_2 and c_3.

We deduce from Lemma 2.3.3 that, with probability at least $1 - c_2 e^{-c_3 \frac{nh^2R^2}{\sigma^2}}$,

$$\frac{1}{2}\|f\|_2^2 - c_1 R^2 h^2 \leq \|f\|_n^2 \leq \frac{3}{2}\|f\|_2^2 + c_1 R^2 h^2.$$

The previous lemma means the following. If we are able to proof that for some $t > 0$, $\|f^t - f^*\|_{\mathcal{H}}^2 \leq cR$ with high probability, for a positive numeric constant c, then we can directly change the optimality result in terms of $\|f^t - f^*\|_n^2$ to the optimality result in terms of the $L_2(\mathbb{P}_X)$ norm $\|f^t - f^*\|_2^2$, losing only $c_1 R^2 h^2 \asymp R^2 \epsilon_n^2$ by choosing $h = \epsilon_n$.

Equipped with the localized Rademacher complexity (2.29), we can state the optimality theorem for finite-rank kernels and any functional sequence $\{f^t\}_{t=0}^{\infty}$ generated by (2.11) and initialized at $f^0 = 0$.

Theorem 2.3.4. *Under Assumptions 1, 2, and 3, given the stopping rule (2.24), there is a numeric constant $\widetilde{c}_u$ so that for finite-rank kernels with rank r:*

$$\mathbb{E}\|f^\tau - f^*\|_2^2 \leq \widetilde{c}_u \frac{r\sigma^2}{n}. \tag{2.32}$$

Proof intuition. The full proof is deferred to Section 2.12. Its main ingredient is Lemma 2.14.2 in Appendix that states the following: $\|f^t\|_{\mathcal{H}} \leq \sqrt{7}R$ for any $t \leq \bar{t}_\epsilon$, where $\bar{t}_\epsilon = \inf\left\{t > 0 \mid B^2(t) = \frac{\sigma^2}{2n}\sum_{i=1}^r \gamma_i^{(t)}\right\}$, with high probability. With this argument, we can apply the triangular inequality and Lemma 2.3.3, if $\tau \leq \bar{t}_\epsilon$ w.h.p.

Remark. Theorem 2.3.4 provides a rate for the $L_2(\mathbb{P}_X)$ norm that matches up to a constant the minimax bound (see e.g. [93, Theorem 2(a)] with $s = d = 1$) when f^* belongs to the $\mathcal{H}$-norm ball of a fixed radius R, thus not improvable in general. A similar bound for finite-rank kernels was achieved in [92, Corollary 4].

We summarize our findings in the following corollary.

Corollary 2.3.5. *Under Assumptions 1, 2, 3, and a finite-rank kernel, early stopping rule τ satisfies*

$$\mathbb{E}\|f^{\tau} - f^*\|_2^2 \asymp \inf_{\widehat{f}} \sup_{\|f^*\|_{\mathcal{H}} \leq R} \mathbb{E}\|\widehat{f} - f^*\|_2^2, \tag{2.33}$$

where the infimum is taken over all measurable functions of the input data.

2.3.2 Practical behavior of τ with infinite-rank kernels

A typical example of RKHS that produces a "smooth" infinite-rank kernel is the k^{th}-order Sobolev spaces for some fixed integer $k \geq 1$ with Lebesgue measure on a bounded domain. We consider Sobolev spaces that consist of functions that have k^{th}-order weak derivatives $f^{(k)}$ being Lebesgue integrable, and $f^{(0)} = f^{(1)}(0) = \ldots = f^{(k-1)}(0) = 0$. It is worth to mention that for such classes, the eigenvalues of the Gram matrix $\widehat{\mu}_i \asymp i^{-\beta},\ i \in [r]$. Another example of kernels with this decay condition for the eigenvalues is the Laplace kernel $\mathbb{K}(x_1, x_2) = e^{-|x_1 - x_2|},\ x_1, x_2 \in \mathbb{R}$ (see [98, p.402]).

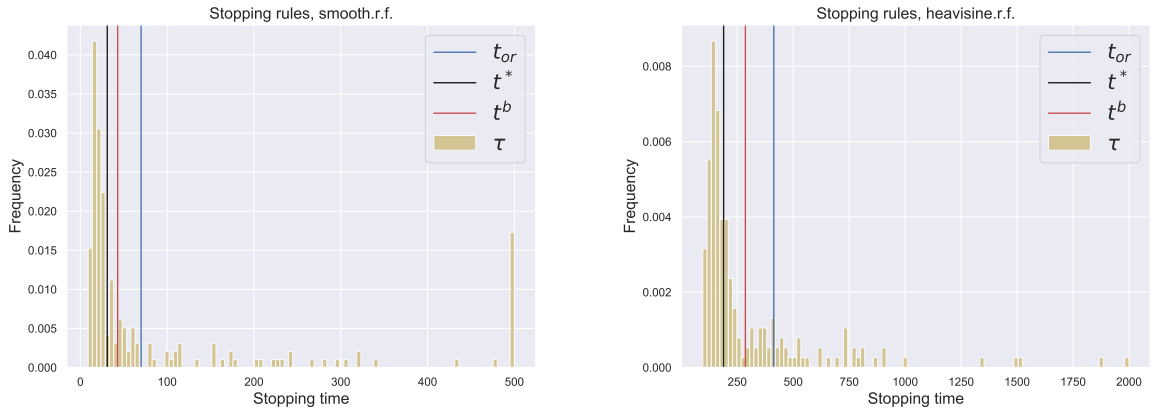
Firstly, let us now illustrate a practical behavior of ESR (2.24) (its histogram) for gradient descent (2.11) with step-size $\eta = 1/(1.2\widehat{\mu}_1)$ for the one dimensional Sobolev kernel $\mathbb{K}(x_1, x_2) = \min\{x_1, x_2\}$ that generates the reproducing space

$$\mathcal{H} = \left\{ f : [0,1] \to \mathbb{R} \mid f(0) = 0, \int_0^1 (f'(x))^2 dx < \infty \right\}. \tag{2.34}$$

We deal with the model (2.1) with two regression functions: the smooth piece-wise linear $f^*(x) = |x - 1/2| - 1/2$ and nonsmooth heavisine $f^*(x) = 0.093\ [4\sin(4\pi x) - \text{sign}(x - 0.3) - \text{sign}(0.72 - x)]$ functions. The design points are random $x_i \overset{\text{i.i.d.}}{\sim} \mathbb{U}[0,1]$. The number of observations is $n = 200$. For both functions, $\|f^*\|_n \approx 0.28$, and we set up a middle difficulty noise level $\sigma = 0.15$. The number of repetitions is $N = 200$.

In panel (a) of Figure 2.2, we detect that our stopping rule τ has high variance. This could be explained by the variability of τ around its proxy version t^* or the variability of the empirical risk R_t around its expectation at t^*. To understand this phenomenon, we move back to Theorem 2.3.1 and notice that the remainder term there vanishes at the fast rate $\mathcal{O}(\sqrt{r}/n)$ when the kernel rank is fixed. If the kernel is not of finite rank, as a consequence, the worst-case rate is $\mathcal{O}(1/\sqrt{n})$ and we could not guarantee that we get a true remainder term at the end. Thus, high variance comes from a large remainder term. Moreover, it has been shown in [30] that the term $\mathcal{O}(\sqrt{r}/n)$ is unavoidable for the spectral cut-off algorithm (in their notation it corresponds to $\sqrt{D}\delta^2$, where $\delta^2 = \frac{\sigma^2}{n}$).

If we change the signal f^* from the smooth to the nonsmooth one, the regression function does not belong anymore to $\mathcal{H}$ defined in (2.34). In this case (panel (b) in Figure 2.2), stopping rule τ performs much better than for the previous regression function. A conclusion one can make is that for the smooth functions in $\mathcal{H}$, one needs to reduce the variance of the empirical risk. In order to do that and to get a stable early stopping rule that will be close to t^*, we propose using a special smoothing



(a) (b)

Figure 2.2 – Histogram of τ vs t^b vs t^* vs $t_{\text{or}} := \underset{t>0}{\text{argmin}}\left[\mathbb{E}_\varepsilon\|f^t - f^*\|_n^2\right]$ for kernel gradient descent with the step-size $\eta = 1/(1.2\widehat{\mu}_1)$ for the piece-wise linear $f^*(x) = |x - 1/2| - 1/2$ (panel (a)), and heavisine $f^*(x) = 0.093\ [4\ \sin(4\pi x) - \text{sign}(x - 0.3) - \text{sign}(0.72 - x)]$ (panel (b)) regression functions, and the first-order Sobolev kernel $\mathbb{K}(x_1, x_2) = \min\{x_1, x_2\}$.

technique for the empirical risk.

2.4 Polynomial smoothing

As it was mentioned earlier, the main issue of poor behavior of the stopping rule τ for "smooth" infinite-rank kernels is the variability of the empirical risk around its expectation. We would like to reduce this variability. To grasp additional intuition of this variability, consider the expectation of the empirical risk $\mathbb{E}_\varepsilon R_t \approx \frac{1}{n}\sum_{i=1}^r (1 - \gamma_i^{(t)})^2 (G_i^*)^2$ and the fact that there exist components $i \in [r]$, for which $(G_i^*)^2 \leq \varepsilon_i^2$, then one can conclude that there is no hope to apply the early stopping rule τ with this type of kernels. That would be extremely difficult to recover the part of the regression function associated with these components since we observe pure noise. Our goal then is to reduce the number of these components and, by doing that, to reduce the variance of the empirical risk. A solution that we propose is to smooth the empirical risk utilizing the eigenvalues of the normalized Gram matrix.

2.4.1 Polynomial smoothing and minimum discrepancy principle rule

We start by defining the squared α-norm as $\|f\|_{n,\alpha}^2 := \langle K_n^\alpha F, F\rangle_n$ for all $F = [f(x_1), \ldots, f(x_n)]^\top \in \mathbb{R}^n$, from which we also introduce the smoothed risk, bias, and variance of a spectral filter estimator as

$$R_\alpha(t) = \mathbb{E}_\varepsilon\|f^t - f^*\|_{n,\alpha}^2 = \|\mathbb{E}_\varepsilon f^t - f^*\|_{n,\alpha}^2 + \mathbb{E}_\varepsilon\|f^t - \mathbb{E}_\varepsilon f^t\|_{n,\alpha}^2 = B_\alpha^2(t) + V_\alpha(t),$$

with

$$B_\alpha^2(t) = \frac{1}{n}\sum_{i=1}^{r} \widehat{\mu}_i^\alpha (1-\gamma_i^{(t)})^2 (G_i^*)^2, \qquad V_\alpha(t) = \frac{\sigma^2}{n}\sum_{i=1}^{r} \widehat{\mu}_i^\alpha (\gamma_i^{(t)})^2. \tag{2.35}$$

The smoothed empirical risk is

$$R_{\alpha,t} = \|F^t - Y\|_{n,\alpha}^2 = \|G^t - Z\|_{n,\alpha}^2 = \frac{1}{n}\sum_{i=1}^{r} \widehat{\mu}_i^\alpha (1-\gamma_i^{(t)})^2 Z_i^2, \qquad \text{for } t>0. \tag{2.36}$$

Recall that the kernel is bounded by $B=1$, thus $\widehat{\mu}_i \le 1$ for all $i=1,\dots,r$, then the smoothed bias $B_\alpha^2(t)$ and smoothed variance $V_\alpha(t)$ are smaller their non-smoothed counterparts.

Analogously to the heuristic derivation leading to the stopping rule (2.24), the new stopping rule is based on the discrepancy principle applied to the $\alpha-$smoothed empirical risk, that is,

$$\tau_\alpha = \inf\left\{ t>0 \mid R_{\alpha,t} \le \sigma^2 \frac{\operatorname{tr}(K_n^\alpha)}{n} \right\}, \tag{2.37}$$

where $\sigma^2 \operatorname{tr}(K_n^\alpha)/n = \sigma^2 \sum_{i=1}^r \widehat{\mu}_i^\alpha / n$ is the natural counterpart of $r\sigma^2/n$ in the case of an infinite-rank kernel and the $\alpha-$norm.

Since there is no straightforward connection between τ_α and the former reference stopping rule $t^b = \inf\{t>0 \mid B^2(t) \le V(t)\}$, we need to introduce a new reference one for the theoretical analysis of the behavior of τ_α. We first define a new smoothed reference stopping rule (which balances between the smoothed bias and variance)

$$t_\alpha^b = \inf\left\{ t>0 \mid B_\alpha^2(t) \le V_\alpha(t) \right\}, \tag{2.38}$$

and also the analogue of (2.23) with the $\alpha-$norm:

$$t_\alpha^* = \inf\left\{ t>0 \mid \mathbb{E}_\varepsilon R_{\alpha,t} \le \frac{\sigma^2}{n}\sum_{i=1}^{r} \widehat{\mu}_i^\alpha \right\}. \tag{2.39}$$

2.4.2 Related work

The idea of smoothing the empirical risk (the residuals) is not new in the literature. For instance, [31, 33, 34] discussed various smoothing strategies applied to (kernelized) conjugate gradient descent, and [49] considered spectral regularization with spectral filter estimators. More closely related to the present work, [105] studied a statistical performance improvement allowed by polynomial smoothing of the residuals (as we do here), but restricted to the spectral cut-off estimator.

[33, 34] considered the following statistical inverse problem: $z = Ax + \sigma\zeta$, where A is a self-adjoint operator and ζ is Gaussian noise. In their case, for the purpose of achieving optimal rates, the usual discrepancy principle rule $\|Ax_m - z\| \le \vartheta\delta$ (m is an iteration number, ϑ is a parameter) was modified and took the form $\|\rho_\lambda(A)(Ax_m - z)\| \le \vartheta\delta$, where $\rho_\lambda(t) = \frac{1}{\sqrt{t+\lambda}}$ and δ is the normalized variance of

Gaussian noise.

In [31], the minimum discrepancy principle was modified as well to the following: each iteration m of conjugate gradient descent was represented by a vector $\widehat{\alpha}_m = K_n^{\dagger} Y$, $K_n^{\dagger}$ is the pseudo-inverse of the normalized Gram matrix, and the learning process was stopped if $\|Y - K_n \widehat{\alpha}_m\|_{K_n} < \Omega$, for some positive Ω, where $\|\alpha\|_{K_n}^2 = \langle \alpha, K_n \alpha \rangle$. Thus, this method corresponds (up to a threshold) to the stopping rule (2.37) with $\alpha = 1$.

In the work [105], the authors concentrated on the inverse problem $Y = A\xi + \delta W$ and its corresponding Gaussian vector observation model $Y_i = \tilde{\mu}_i \xi_i + \delta \varepsilon_i,\ i \in [r]$, where $\{\tilde{\mu}_i\}_{i=1}^r$ are the singular values of the linear bounded operator A and $\{\varepsilon_i\}_{i=1}^r$ are Gaussian noise variables. They recovered the signal $\{\xi_i\}_{i=1}^r$ by a cut-off estimator of the form $\widehat{\xi}_i^{(t)} = \mathbb{I}\{i \leq t\} \tilde{\mu}_i^{-1} Y_i,\ i \in [r]$. The minimum discrepancy principle in this case was $\|(AA^{\top})^{\alpha/2}(Y - A\widehat{\xi}^{(t)})\|^2 \leq \kappa$ for some positive κ. They found out that if the smoothing parameter α lies in the interval $[\frac{1}{4p}, \frac{1}{2p})$, where p is the polynomial decay of the singular values $\{\tilde{\mu}_i\}_{i=1}^r$, then the cut-off estimator is adaptive to Sobolev ellipsoids. Therefore, our work could be considered as an extension of the work [105] to generalize the polynomial smoothing strategy to more complex filter estimators such as gradient descent and (Tikhonov) ridge regression in the reproducing kernel framework.

2.4.3 Optimality result (fixed-design)

To take into account in our analysis the fact that we use the $\alpha-$norm, we define a modified version of the localized empirical Rademacher complexity that we call the *smoothed empirical Rademacher complexity*. The derivation of the next expression is deferred to Appendix 2.13.

Definition 2.4.1. The smoothed empirical Rademacher complexity of $\mathbb{B}_{\mathcal{H}}(R)$ is defined as

$$\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) = R \sqrt{\frac{1}{n} \sum_{i=1}^{r} \widehat{\mu}_i^{\alpha} \min\{\widehat{\mu}_i, \epsilon^2\}}, \tag{2.40}$$

where $\alpha \in [0, 1]$ and $\{\widehat{\mu}_i\}_{i=1}^r$ are the eigenvalues of the Gram matrix K_n.

This new definition leads to the next updated smoothed version of the critical inequality and its related empirical critical radius.

Definition 2.4.2. Define the *smoothed empirical critical radius* $\widehat{\epsilon}_{n,\alpha}$ as the smallest positive solution $\epsilon > 0$ to the following fixed-point inequality

$$\frac{\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H})}{\epsilon R} \leq \frac{2R}{\sigma} \epsilon^{1+\alpha}. \tag{2.41}$$

Appendix 2.14 establishes that the smoothed empirical critical radius $\widehat{\epsilon}_{n,\alpha}$ does exist, is unique and achieves the equality in Ineq. (2.41).

We pursue the analogy a bit further by defining the *smoothed statistical dimension* as

$$d_{n,\alpha} := \min\Big\{j \in [r] \mid \widehat{\mu}_j \leq \widehat{\epsilon}^2_{n,\alpha}\Big\}, \tag{2.42}$$

and $d_{n,\alpha} = r$ if no such index does exist. Combined with (2.40), this implies that

$$\widehat{\mathcal{R}}^2_{n,\alpha}(\widehat{\epsilon}_{n,\alpha}, \mathcal{H}) \geq \frac{\sum_{j=1}^{d_{n,\alpha}} \widehat{\mu}_j^\alpha}{n} R^2 \widehat{\epsilon}^2_{n,\alpha}, \quad \text{and} \quad \widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha} \geq \frac{\sigma^2 \sum_{j=1}^{d_{n,\alpha}} \widehat{\mu}_j^\alpha}{4R^2 n}, \tag{2.43}$$

where the second statement results from Ineq. (2.41). Let us emphasize that [121] already introduced the so-called *statistical dimension* (corresponds to $\alpha = 0$ in our notations). It appeared that the statistical dimension provides an upper bound on the minimax optimal dimension of randomized projections for kernel ridge regression (see [121, Theorem 2, Corollary 1]).

In our case, $d_{n,\alpha}$ can be seen as a (α-smooth) version of the statistical dimension. One motivation is that this notion turns out to be useful in the derivation of minimax rates. In particular, this can be achieved by using the following assumptions that involve this quantity.

Assumption 4. *There exists a numeric $\mathcal{A} > 0$ such that for all $\alpha \in [0, 1]$,*

$$\sum_{i=d_{n,\alpha}+1}^{r} \widehat{\mu}_i \leq \mathcal{A} d_{n,\alpha} \widehat{\epsilon}^2_{n,\alpha}. \tag{2.44}$$

This assumption will further make the transfer from the smooth critical inequality (2.41) to its non-smooth version (2.20). Indeed, under Assumption 4, if ϵ satisfies Ineq. (2.41), then it satisfies Ineq. (2.20) as well, where constant 2 on the right-hand side is replaced by $2\sqrt{1+\mathcal{A}}$ (see Lemma 2.14.3 in Appendix 2.14). Although there are reproducing kernels for which Assumption 4 does not hold, for most of them, it holds [121], including all the examples in the present chapter. We detail one of them below.

Example 4 (β-polynomial eigenvalue decay)**.** *Let us assume that the eigenvalues of the normalized Gram matrix satisfy that there exist numeric constants $0 < c \leq C$ such that*

$$ci^{-\beta} \leq \widehat{\mu}_i \leq Ci^{-\beta}, \quad i = 1, \ldots, r, \tag{2.45}$$

for some $\beta > 1$. Instances of kernels in this class are mentioned at the beginning of Section 2.3.2. Then, Assumption 4 holds true with $\mathcal{A} = \frac{C}{c}\frac{1}{\beta-1}$.

Another key property for the smoothing to yield optimal results is that the value of α has to be large enough to control the tail sum of the smoothed eigenvalues by the corresponding cumulative sum, which is the purpose of the assumption below.

Assumption 5. *There exists* $\Upsilon = [\alpha_0, 1]$, $\alpha_0 \geq 0$, *such that, for all* $\alpha \in \Upsilon$,

$$\sum_{i=d_{n,\alpha}+1}^{r} \widehat{\mu}_i^{2\alpha} \leq \mathcal{M} \sum_{i=1}^{d_{n,\alpha}} \widehat{\mu}_i^{2\alpha}, \tag{2.46}$$

where $\mathcal{M} \geq 1$ *denotes a numeric constant.*

Let us remark that controlling the tail sum of the empirical eigenvalues has been already made, for example, by [18] (effective rank) and more recently by [49, Assumption 6]. Let us also mention that Assumption 5 does not imply Assumption 4 holds.

We enumerate several classical examples, for which this assumption holds.

Example 5 (β-polynomial eigenvalue decay kernels (2.45))**.** *For the polynomial eigenvalue-decay kernels, Assumption 5 holds with*

$$\mathcal{M} = 2\Big(\frac{C}{c}\Big)^2 \quad \text{and} \quad 1 \geq \alpha \geq \frac{1}{\beta+1} = \alpha_0. \tag{2.47}$$

Example 6 (γ-exponential eigenvalue-decay kernels)**.** *Let us assume that the eigenvalues of the normalized Gram matrix satisfy that there exist numeric constants* $0 < c \leq C$ *and a constant* $\gamma > 0$ *such that*

$$ce^{-i^\gamma} \leq \widehat{\mu}_i \leq Ce^{-i^\gamma}.$$

Instances of kernels within this class include the Gaussian kernel with respect to the Lebesgue measure on the real line (with $\gamma = 2$*) or on a compact domain (with* $\gamma = 1$*) (up to* log *factor in the exponent). Then, Assumption 5 holds with*

$$\mathcal{M} = \Big(\frac{C}{c}\Big)^2 \frac{\int_0^\infty e^{-y^\gamma} dy}{\int_{(2\alpha_0)^{1/\gamma}}^{2(2\alpha_0)^{1/\gamma}} e^{-y^\gamma} dy} \quad \text{and} \quad \alpha \in [\alpha_0, 1], \quad \textit{for} \quad \alpha_0 > 0.$$

For *any reproducing kernel* satisfying the above assumptions, the next theorem provides a high probability bound on the performance of f^{τ_α} (measured in terms of the $L_2(\mathbb{P}_n)$ norm), which depends on the smoothed empirical critical radius.

Theorem 2.4.1 (Upper bound)**.** *Under Assumptions 1, 2, 3, 4, and 5, given the stopping rule (2.37):*

$$\|f^{\tau_\alpha} - f^*\|_n^2 \leq c_u R^2 \widehat{\epsilon}_{n,\alpha}^2 \tag{2.48}$$

with probability at least $1 - 5\exp\Big[-c_1 \frac{R^2}{\sigma^2} n \widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$*, for some positive constants* c_1 *and* c_u*, where* c_1 *depends only on* $\mathcal{M}$*,* c_u *depends only on* $\mathcal{A}$*.*

Moreover,

$$\mathbb{E}_\varepsilon \|f^{\tau_\alpha} - f^*\|_n^2 \leq CR^2\widehat{\epsilon}_{n,\alpha}^2 + 6\max\{\sigma^2, R^2\}\exp\left[-c_3\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\right] \tag{2.49}$$

for constant C only depending on $\mathcal{A}$, constant c_3 only depending on $\mathcal{M}$.

First of all, Theorem 2.4.1 is established in the fixed-design framework, and Ineq. (2.49) is a direct consequence of the high probability bound (2.48). The main message is that the final performance of the estimator f^{τ_α} is controlled by the smoothed critical radius $\widehat{\epsilon}_{n,\alpha}$. From the existing literature on the empirical critical radius [92, 93, 114, 121], it is already known that the non-smoothed version $\widehat{\epsilon}_n^2$ is the typical quantity that leads to minimax rates in the RKHS (see also Theorem 2.4.2 below). In particular, tight upper bounds on $\widehat{\epsilon}_n^2$ can be computed from a priori information about the RKHS, e.g., the decay rate of the empirical/population eigenvalues. However, the behavior of $\widehat{\epsilon}_{n,\alpha}^2$ with respect to n is likely to depend on α, as emphasized by the notation. Intuitively, this suggests that there could exist a range of values of α, for which $\widehat{\epsilon}_{n,\alpha}^2$ is of the same order as (or faster than) $\widehat{\epsilon}_n^2$, leading therefore to optimal rates. But there could also exist ranges of values of α, where this does not hold, leading to suboptimal rates.

Another striking aspect of Ineq. (2.49) is related to the additional terms involving the exponential function in Ineq. (2.49). As far as (2.48) is a statement with "high probability", this term is expected to converge to 0 at a rate depending on $n\widehat{\epsilon}_{n,\alpha}^2$. Therefore, the final convergence rate as well as the fact that this term is (or not) negligible will depend on α.

Sketch of proof of Theorem 2.4.1. The complete proof is given in Appendix 2.10 and starts from splitting the risk error $\|f^{\tau_\alpha} - f^*\|_n^2$ into two parts:

$$2B^2(\tau_\alpha) + 2v(\tau_\alpha), \tag{2.50}$$

where $v(t) \coloneqq \frac{1}{n}\sum_{i=1}^n (\gamma_i^{(t)})^2\varepsilon_i^2$ is called the stochastic part of the variance at iteration t.

The key ingredients of the proof are the next two deviation inequalities.

$$i)\ \mathbb{P}_\varepsilon\left(\tau_\alpha > \bar{t}_{\epsilon,\alpha}\right) \leq 2\ \exp\left[-c_1\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\right],$$
$$ii)\ \mathbb{P}_\varepsilon\left(\tau_\alpha < \widetilde{t}_{\epsilon,\alpha}\right) \leq 2\ \exp\left[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\right],$$

where $\bar{t}_{\epsilon,\alpha}$ and $\widetilde{t}_{\epsilon,\alpha}$ are some properly chosen upper and lower bounds of t_α^*.

Since it can be shown that $\eta\widetilde{t}_{\epsilon,\alpha} \asymp \eta\bar{t}_{\epsilon,\alpha} \asymp (\widehat{\epsilon}_{n,\alpha}^2)^{-1}$, these two inequalities show that τ_α stays of the optimal order $(\widehat{\epsilon}_{n,\alpha}^2)^{-1}$ with high probability. After that, it is sufficient to upper bound each term in (2.50), and the claim follows.

■

The purpose of the following result is to give more insight into the understanding of Theorem 2.4.1 regarding the influence of the different terms in the convergence rate.

Theorem 2.4.2 (Lower bound from Theorem 1 in [121]). *If Assumption 4 holds true with $\alpha = 0$, then for any estimator $\widetilde{f}$ of $f^* \in \mathbb{B}_{\mathcal{H}}(R)$ satisfying the nonparametric model defined in Eq. (2.1), we get*

$$\mathbb{E}_\varepsilon \|\widetilde{f} - f^*\|_n^2 \geq c_l R^2 \widehat{\epsilon}_n^2,$$

for some numeric constant c_l that only depends on $\mathcal{A}$ from Assumption 4.

Firstly, Theorem 2.4.2 has been proved in [121] with $R = 1$, and a simple rescaling argument provides the above statement, so we do not reproduce the proof here. Secondly, Theorem 2.4.2 applies to any kernel as long as Assumption 4 is fulfilled with $\alpha = 0$, which is in particular true for the reproducing kernels from Theorem 2.4.1. Therefore, the fastest achievable rate by an estimator of f^* is $\widehat{\epsilon}_n^2$. As a consequence, as far as there exist values of α such that $\widehat{\epsilon}_{n,\alpha}^2$ is at most as large as $\widehat{\epsilon}_n^2$, the estimator f^{τ_α} is optimal.

2.4.4 Consequences for β-polynomial eigenvalue-decay kernels

The leading idea in the present section is identifying values of α, for which the bound (2.48) from Theorem 2.4.1 scales as $R^2\widehat{\epsilon}_n^2$.

Let us recall the definition of a polynomial decay kernel from (2.45):

$$ci^{-\beta} \leq \widehat{\mu}_i \leq Ci^{-\beta}, \ i \in [r], \quad \text{for some } \beta > 1 \text{ and numeric constants } c, C > 0.$$

One typical example of the reproducing kernel satisfying this condition is the Sobolev kernel on $[0,1] \times [0,1]$ given by $\mathbb{K}(x,x') = \min\{x,x'\}$, with $\beta = 2$ [92]. The corresponding RKHS is the first-order Sobolev class, that is, the class of functions that are almost everywhere differentiable with the derivative in $L_2[0,1]$.

Lemma 2.4.3. *Assume there exists $\beta > 1$ such that the β-polynomial decay assumption from (2.45) holds. Then there exist numeric constants $c_1, c_2 > 0$ such that, for $\alpha < 1/\beta$, one has*

$$c_1\widehat{\epsilon}_n^2 \leq \widehat{\epsilon}_{n,\alpha}^2 \leq c_2\widehat{\epsilon}_n^2 \asymp \left[\frac{\sigma^2}{2R^2 n}\right]^{\frac{\beta}{\beta+1}}.$$

The proof of Lemma 2.4.3, which can be derived from combining Lemmas 2.7.4 and 2.7.5 from Appendix 2.7, is not reproduced here.

Therefore, if $\alpha\beta < 1$, then $\widehat{\epsilon}^2_{n,\alpha} \asymp \widehat{\epsilon}^2_n \asymp \left[\frac{\sigma^2}{2R^2 n}\right]^{\frac{\beta}{\beta+1}}$. Let us now recall from (2.47) that Assumption 5 holds true for $\alpha \geq (\beta+1)^{-1}$. All these arguments lead us to the next result, which establishes the minimax optimality of τ_α with any kernel satisfying the β-polynomial eigenvalue-decay assumption, as long as $\alpha \in [\frac{1}{\beta+1}, \frac{1}{\beta})$.

Corollary 2.4.4. *Under Assumptions 1, 2, 3, and the β-polynomial eigenvalue decay (2.45), for any $\alpha \in [\frac{1}{\beta+1}, \frac{1}{\beta})$, the early stopping rule τ_α satisfies*

$$\mathbb{E}_\varepsilon \|f^{\tau_\alpha} - f^*\|_n^2 \asymp \inf_{\widehat{f}} \sup_{\|f^*\|_{\mathcal{H}} \leq R} \mathbb{E}_\varepsilon \|\widehat{f} - f^*\|_n^2, \tag{2.51}$$

where the infimum is taken over all measurable functions of the input data.

Corollary 2.4.4 establishes an optimality result in the fixed-design framework since, as long as $(\beta+1)^{-1} \leq \alpha < \beta^{-1}$, the upper bound matches the lower bound up to multiplicative constants. Moreover, this property holds uniformly with respect to $\beta > 1$ provided the value of α is chosen appropriately. An interesting feature of this bound is that the optimal value of α only depends on the (polynomial) decay rate of the empirical eigenvalues of the normalized Gram matrix. This suggests that any effective estimator of the unknown parameter β could be plugged into the above (fixed-design) result and would lead to an optimal rate. Note that [105] has recently emphasized a similar trade-off ($(\beta+1)^{-1} \leq \alpha < \beta^{-1}$) for the smoothing parameter α (polynomial smoothing), considering the spectral cut-off estimator in the Gaussian sequence model. Regarding convergence rates, Corollary 2.4.4 combined with Lemma 2.4.3 suggests that the convergence rate of the expected (fixed-design) risk is of the order $\mathcal{O}\left(n^{-\frac{\beta}{\beta+1}}\right)$. This is the same as the already known one in nonparametric regression in the random design framework [92, 106], which is known to be minimax optimal as long as f^* belongs to the RKHS $\mathcal{H}$.

2.5 Empirical comparison with existing stopping rules

The present section aims at illustrating the practical behavior of several stopping rules discussed in the chapter as well as making a comparison with existing alternative stopping rules.

2.5.1 Stopping rules involved

The empirical comparison is carried out between the stopping rules τ (2.24) and τ_α with $\alpha \in [\frac{1}{\beta+1}, \frac{1}{\beta})$ (2.37), and four alternative stopping rules that are briefly described in what follows. For the sake of comparison, most of them correspond to early stopping rules already considered in [92].

Hold-out stopping rule

We consider a procedure built on the hold-out idea [8]. The data $\{(x_i, y_i)\}_{i=1}^n$ are split into two parts: the training sample $S_{\text{train}} = (x_{\text{train}}, y_{\text{train}})$ and the test sample $S_{\text{test}} = (x_{\text{test}}, y_{\text{test}})$ so that the training sample and test sample represent a half of the whole dataset. We train the learning algorithm for $t = 0, 1, \ldots$ and estimate the risk, for each t, by $R_{\text{ho}}(f^t) = \frac{1}{n} \sum_{i \in S_{\text{test}}} ((\widehat{y}_{\text{test}})_i - y_i)^2$, where $(\widehat{y}_{\text{test}})_i$ denotes the output of the algorithm trained at iteration t on S_{train} and evaluated at the point x_i of the test sample. The final stopping rule is defined as

$$\widehat{\mathrm{T}}_{\text{HO}} = \operatorname{argmin}\Big\{ t \in \mathbb{N} \mid R_{\text{ho}}(f^{t+1}) > R_{\text{ho}}(f^t) \Big\} - 1. \tag{2.52}$$

Although it does not completely use the data for training (loss of information), the hold-out strategy has been proved to output minimax optimal estimators in various contexts (see, for instance, [43, 45] with Sobolev spaces and $\beta \leq 2$).

V-fold stopping rule

The observations $\{(x_i, y_i)\}_{i=1}^n$ are randomly split into $V = 4$ equal sized blocks. At each round (among the V ones), $V - 1$ blocks are devoted to training $S_{\text{train}} = (x_{\text{train}}, y_{\text{train}})$, and the remaining one serves for the test sample $S_{\text{test}} = (x_{\text{test}}, y_{\text{test}})$. At each iteration $t = 0, 1, \ldots$, the risk is estimated by $R_{\text{VFCV}}(f^t) = \frac{1}{V-1} \sum_{j=1}^{V-1} \frac{1}{n/V} \sum_{i \in S_{\text{test}}(j)} ((\widehat{y}_{\text{test}})_i - y_i)^2$, where $\widehat{y}_{\text{test}}$ was described for the hold-out stopping rule. The final stopping rule is

$$\widehat{\mathrm{T}}_{\text{VFCV}} = \operatorname{argmin}\{ t \in \mathbb{N} \mid R_{\text{VFCV}}(f^{t+1}) > R_{\text{VFCV}}(f^t) \} - 1. \tag{2.53}$$

V-fold cross-validation is widely used in practice since, on the one hand, it is more computationally tractable than other splitting-based methods such as leave-one-out or leave-p-out (see the survey [8]), and on the other hand, it enjoys better statistical performance than the hold-out (lower variability).

Raskutti-Wainwright-Yu stopping rule (from [92])

The use of this stopping rule heavily relies on the assumption that $\|f^*\|_{\mathcal{H}}^2$ is known, which is a strong requirement in practice. It controls the bias-variance trade-off by using upper bounds on the bias and variance terms. The latter involves the localized empirical Rademacher complexity $\widehat{\mathcal{R}}_n\left(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\right)$. Similarly to t^b, it stops as soon as (upper bound of) the bias term becomes smaller than (upper bound on) the variance term, which leads to

$$\widehat{\mathrm{T}}_{\text{RWY}} = \operatorname{argmin}\Big\{ t \in \mathbb{N} \mid \widehat{\mathcal{R}}_n\Big(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\Big) > (2e\sigma\eta t)^{-1} \Big\} - 1. \tag{2.54}$$

Theoretical minimum discrepancy-based stopping rule t^*

The fourth stopping rule is the one introduced in (2.23). It relies on the minimum discrepancy principle and involves the (theoretical) expected empirical risk $\mathbb{E}_\varepsilon R_t$:

$$t^* = \inf\left\{t > 0 \mid \mathbb{E}_\varepsilon R_t \leq \sigma^2\right\}.$$

This stopping rule is introduced for comparison purposes only since it cannot be computed in practice. This rule is proved to be optimal (see Appendix 2.9) for *any bounded reproducing kernel* so that it could serve as a reference in the present empirical comparison.

Oracle stopping rule

The "oracle" stopping rule is defined as the first time the risk curve starts to increase.

$$t_{\text{or}} = \operatorname{argmin}\{t \in \mathbb{N} \mid \mathbb{E}_\varepsilon \|f^{t+1} - f^*\|_n^2 > \mathbb{E}_\varepsilon \|f^t - f^*\|_n^2\} - 1. \tag{2.55}$$

In situations where only one global minimum does exist for the risk, this rule coincides with the global minimum location. Its formulation reflects the realistic constraint that we do not have access to the whole risk curve (unlike in the classical model selection setup).

2.5.2 Simulation design

Artificial data are generated according to the regression model $y_j = f^*(x_j) + \varepsilon_j$, $j = 1, \dots, n$, where $\varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, with equidistant $x_j = j/n$, $j = 1, \dots, n$, and $\sigma = 0.15$. The same experiments have been also carried out with $x_j \sim \mathbb{U}[0,1]$ (not reported here) without any change regarding the conclusions. The sample size n varies from 40 to 400.

The gradient descent algorithm (2.11) has been used with the step-size $\eta = (1.2\,\widehat{\mu}_1)^{-1}$ and initialization $F^0 = [0, \dots, 0]^\top$.

The present comparison involves two regression functions with the same $L_2(\mathbb{P}_n)$ norms of the signal $\|f^*\|_n \approx 0.28$: (i) a piecewise linear function called "smooth" $f^*(x) = |x - 1/2| - 1/2$, and (ii) a "sinus" function $f^*(x) = 0.9\,\sin(8\pi x)x^2$. An illustration of the corresponding curves is displayed in Figure 2.3.

To ease the comparison, the piecewise linear regression function was set up as in [92, Figure 3].

The case of finite-rank kernels is addressed in Section 2.5.3 with the so-called polynomial kernel of degree 3 defined by $\mathbb{K}(x_1, x_2) = (1 + x_1^\top x_2)^3$ on the unit square $[0,1] \times [0,1]$. By contrast, Section 2.5.3 tackles the polynomial decay kernels with the first-order Sobolev kernel $\mathbb{K}(x_1, x_2) = \min\{x_1, x_2\}$ on the unit square $[0,1] \times [0,1]$.

The performance of the early stopping rules is measured in terms of the $L_2(\mathbb{P}_n)$ squared norm $\|f^t - f^*\|_n^2$ averaged over $N = 100$ independent trials.

For our simulations, we use a variance estimation method that is described in Section 2.5.4. This method is asymptotically unbiased, which is sufficient for our purposes.

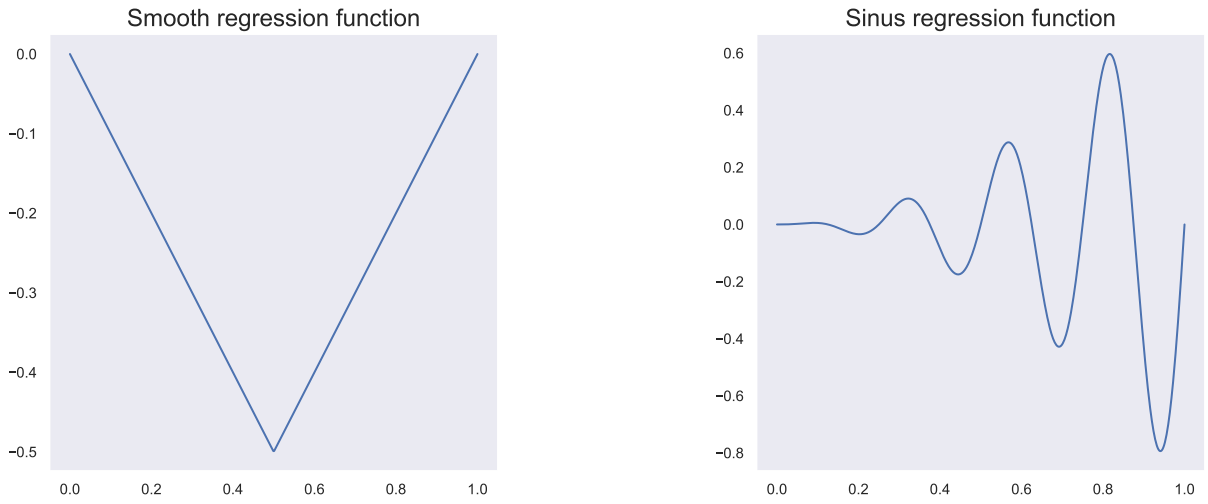


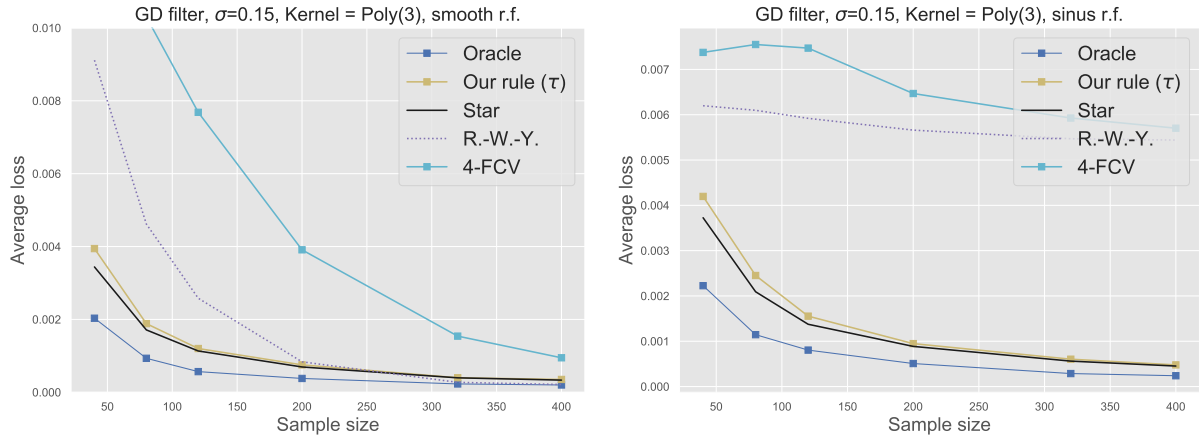
Figure 2.3 – "Smooth" and "sinus" regression functions

2.5.3 Results of the simulation experiments

Finite-rank kernels

Figure 2.4 displays the (averaged) $L_2(\mathbb{P}_n)$ norm error of the oracle stopping rule (2.55), our stopping rule τ (2.24), t^* (2.23), the minimax optimal stopping rule $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ (2.54), and the 4-fold cross validation stopping rule $\widehat{\mathrm{T}}_{\mathrm{VFCV}}$ (2.53) versus the sample size. Figure 2.4a shows the results for the piece-wise linear regression function whereas Figure 2.4b corresponds to the "sinus" regression function.

All the curves decrease as n grows. From these graphs, the overall worst performance is achieved by $\widehat{\mathrm{T}}_{\mathrm{VFCV}}$, especially with a small sample size, which can be due to the additional randomness induced by the preliminary random splitting with $4-FCV$. By contrast, the minimum discrepancy-based stopping rules (τ and t^*) exhibit the best performances compared to the results of $\widehat{\mathrm{T}}_{\mathrm{VFCV}}$ and $\widehat{\mathrm{T}}_{\mathrm{RWY}}$. The averaged mean-squared error of τ is getting closer to the one of t^* as the number of samples n increases, which was expected from the theory and also intuitively, since τ has been introduced as an estimator of t^*. From Figure 2.4a, $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ is less accurate for small sample sizes but improves a lot as n grows up to achieving a performance similar to that of τ. This can result from the fact that $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ is built from upper bounds on the bias and variance terms, which are likely to be looser with a small sample size but achieves an optimal convergence rate as n increases. In Figure 2.4b, the reason why τ exhibits (strongly) better results than $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ owes to the main assumption on the regression function, namely that $\|f^*\|_{\mathcal{H}} \leq 1$. It could be violated for the "sinus" function.



(a) (b)

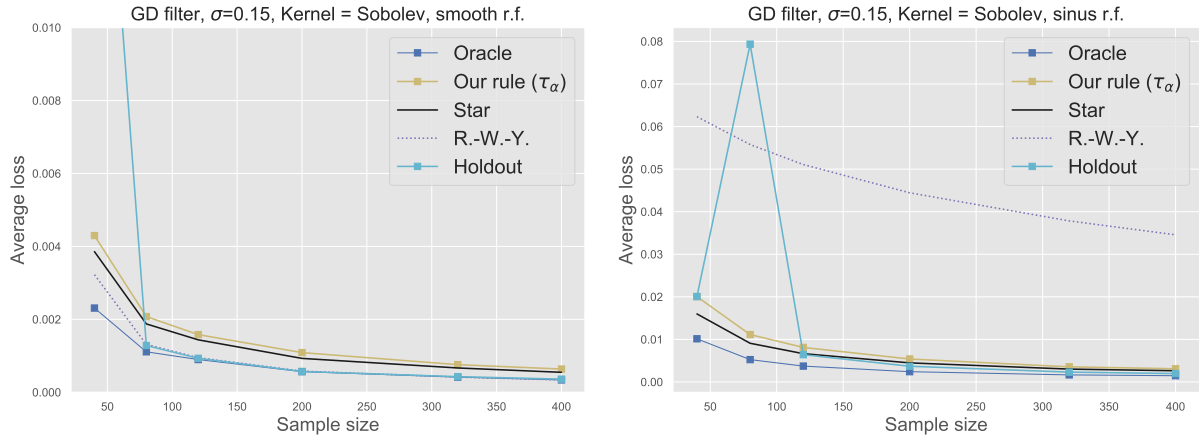
Figure 2.4 – Kernel gradient descent with the step-size $\eta = 1/(1.2\widehat{\mu}_1)$ and polynomial kernel $\mathbb{K}(x_1, x_2) = (1 + x_1^\top x_2)^3$, $x_1, x_2 \in [0, 1]$, for the estimation of two noised regression functions from Figure 2.3: smooth $f^*(x) = |x - 1/2| - 1/2$ for the panel (a) and "sinus" $f^*(x) = 0.9\sin(8\pi x)x^2$ for the panel (b), with the equidistant covariates $x_j = j/n$. Each curve corresponds to the $L_2(\mathbb{P}_n)$ squared norm error for the stopping rules (2.55), (2.23), (2.54), (2.53), (2.24), averaged over 100 independent trials, versus the sample size $n = \{40, 80, 120, 200, 320, 400\}$.

Polynomial eigenvalue decay kernels

Figure 2.5 displays the resulting (averaged over 100 repetitions) $L_2(\mathbb{P}_n)$ error of τ_α (with $\alpha = (\beta+1)^{-1} = 0.33$) (2.37), $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ (2.54), t^* (2.23), and $\widehat{\mathrm{T}}_{\mathrm{HO}}$ (2.52) versus the sample size n. Figure 2.5a shows that all stopping rules seem to work equivalently well, although there is a slight advantage for $\widehat{\mathrm{T}}_{\mathrm{HO}}$ and $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ compared to t^* and τ_α. However, as n grows to $n = 400$, the performances of all stopping rules become very close to each others. Let us mention that the true value of β is not known in these experiments. Therefore, the value $(\beta + 1)^{-1} = 0.33$ has been estimated from the decay of the empirical eigenvalue of the normalized Gram matrix. This can explain why the performance of τ_α remains worse than that of $\widehat{\mathrm{T}}_{\mathrm{RWY}}$.

The story described by Figure 2.5b is somewhat different. The first striking remark is that $\widehat{\mathrm{T}}_{\mathrm{RWY}}$ completely fails on this example, which still stems from the (unsatisfied) constraint on the $\mathcal{H}$-norm of f^*. However, the best performance is still achieved by the Hold-out stopping rule, although τ_α and t^* remain very close to the latter. The fact that t^* remains close to the oracle stopping rule (without any need for smoothing) supports the idea that the minimum discrepancy is a reliable principle for designing an effective stopping rule. The deficiency of τ (by contrast to τ_α) then results from the variability of the empirical risk, which does not remain close enough to its expectation. This bad behavior is then balanced by introducing the polynomial smoothing at level α within the definition of τ_α, which enjoys close to optimal practical performances.

Let us also mention that $\widehat{\mathrm{T}}_{\mathrm{HO}}$ exhibit some variability, in particular, with small sample sizes as it



(a) (b)

Figure 2.5 – Kernel gradient descent (2.11) with the step-size $\eta = 1/(1.2\widehat{\mu}_1)$ and Sobolev kernel $\mathbb{K}(x_1, x_2) = \min\{x_1, x_2\}$, $x_1, x_2 \in [0, 1]$, for the estimation of two noised regression functions from Figure 2.3: smooth $f^*(x) = |x - 1/2| - 1/2$ for the panel (a) and "sinus" $f^*(x) = 0.9\sin(8\pi x)x^2$ for the panel (b), with the equidistant covariates $x_j = j/n$. Each curve corresponds to the $L_2(\mathbb{P}_n)$ squared norm error for the stopping rules (2.55), (2.23), (2.54), (2.52), (2.37) with $\alpha = 0.33$, averaged over 100 independent trials, versus the sample size $n = \{40, 80, 120, 200, 320, 400\}$.

is illustrated by Figures 2.5a and 2.5b.

The overall conclusion is that the smoothed minimum discrepancy-based stopping rule τ_α leads to almost optimal performances provided $\alpha = (\beta + 1)^{-1}$, where β quantifies the polynomial decay of the empirical eigenvalues of the normalized Gram matrix.

2.5.4 Estimation of variance and decay rate for polynomial eigenvalue decay kernels

The purpose of the present section is to describe two strategies for estimating: (i) the decay rate of the empirical eigenvalues of the normalized Gram matrix, and (ii) the variance parameter σ^2.

Polynomial decay parameter estimation

From the polynomial decay assumption (2.45), one easily derives upper and lower bounds for β as

$$\frac{\log(\widehat{\mu}_i/\widehat{\mu}_{i+1}) - \log(C/c)}{\log(1 + 1/i)} \leq \beta \leq \frac{\log(\widehat{\mu}_i/\widehat{\mu}_{i+1}) + \log(C/c)}{\log(1 + 1/i)}.$$

The difference between these upper and lower bounds is equal to $\frac{2\log(C/c)}{\log(1+1/i)}$, which is minimized for $i = 1$. Then, the best precision on the estimated value of β is reached with $i = 1$, which yields the

estimator

$$\widehat{\beta} = \frac{\log(\widehat{\mu}_1/\widehat{\mu}_2)}{\log 2}. \tag{2.56}$$

Note that this estimator $\widehat{\beta}$ from (2.56) is not rigorously grounded but only serves as a rough choice in our simulation experiments (see Section 2.5.3).

Variance parameter estimation

There is a bunch of suggestions for variance estimation with linear smoothers; see, e.g., Section 5.6 in the book [116]. In our simulation experiments, two cases are distinguished: the situation, where the reproducing kernel has finite rank r, and the situation, where the empirical eigenvalues of the normalized Gram matrix exhibit a polynomial decay. In both cases, an asymptotically unbiased estimator of σ^2 is designed.

Finite-rank kernel. With such a finite-rank kernel, one estimates the noise from the coordinates $\{Z_i\}_{i=r+1}^n$ corresponding to the situation, where $G_i^* = 0,\ i > r$ (see Lemma 2.7.1 in Appendix 2.7). Actually, these coordinates (which are pure noise) are exploited to build an easy-to-compute estimator of σ^2, that is,

$$\widehat{\sigma}^2 = \frac{\sum_{i=n-r+1}^n Z_i^2}{n-r}. \tag{2.57}$$

Polynomial decay kernel. If the empirical eigenvalues of K_n satisfy the polynomial eigenvalue decay assumption (2.45), we suggest overly-smoothing the residuals by choosing $\alpha = 1$, which intuitively results in reducing by a large amount the variability of the corresponding smoothed empirical risk around its expectation, that is, $\mathbb{E}_\varepsilon R_{1,t} \approx R_{1,t}$.

Therefore, the smoothed empirical risk can be approximated by $R_{1,t} \approx B_1^2(t) + \frac{\sigma^2}{n}\sum_{i=1}^r (1-\gamma_i^{(t)})^2$, and

$$\sigma^2 \approx \frac{R_{1,t} - B_1^2(t)}{\frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i (1-\gamma_i^{(t)})^2}.$$

Using furthermore that $B_1^2(t) \to 0$ as t increases to $+\infty$, the final choice is

$$\widehat{\sigma}^2 = \frac{R_{1,t}}{\frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i (1-\gamma_i^{(t)})^2}.$$

Following the above heuristic argument, let us emphasize that $\widehat{\sigma}^2$ is likely to be an upper bound on the true variance σ^2 since the (non-negative) bias is lower bounded by 0. Nevertheless, the next result justifies this choice.

Lemma 2.5.1. *Under the polynomial eigenvalue decay assumption* (2.45)*, any value of t satisfying $t \cdot \eta\widehat{\epsilon}_n^2 \to +\infty$ as $n \to +\infty$ yields that $\widehat{\sigma}^2 = \frac{R_{1,t}}{\frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i (1-\gamma_i^{(t)})^2}$ is an asymptotically unbiased estimator of σ^2.*

A sketch of the proof of Lemma 4.22 is given in Appendix 2.15. Based on this lemma, we suggest taking $t = T$, where T is the maximum number of iterations allowed to execute due to computational constraints. Notice that as long as we access to closed-form expressions of the estimator, there is no need to compute all estimators for t between $1 \leq t \leq T$. The final estimator of σ^2 used in the experiments of Section 2.5.3 is given by

$$\widehat{\sigma}^2 = \frac{R_{1,T}}{\frac{1}{n}\sum_{i=1}^{r} \widehat{\mu}_i (1 - \gamma_i^{(T)})^2}. \tag{2.58}$$

2.6 Conclusion

In this chapter, we described spectral filter estimators (gradient descent, kernel ridge regression) for the nonparametric regression function estimation in RKHS. Two new data-driven early stopping rules τ (2.24) and τ_α (2.37) for these iterative algorithms are designed. In more detail, we show that for the infinite-rank reproducing kernels, τ has high variance due to the variability of the empirical risk around its expectation, and we proposed a way to reduce this variability by means of smoothing the empirical $L_2(\mathbb{P}_n)$ norm (and, as a consequence, the empirical risk) by the eigenvalues of the normalized kernel matrix. We demonstrate in Corollaries 2.3.5 and 2.4.4 that our stopping rules τ and τ_α yield minimax-optimal rates, in particular, for finite-rank kernel classes and Sobolev spaces. It is worth to mention that computing our stopping rules (for a general reproducing kernel) requires *only* the estimation of the variance σ^2 and computing $(\widehat{\mu}_1, \ldots, \widehat{\mu}_r)$. Theoretical results are confirmed empirically: τ and τ_α with the smoothing parameter $\alpha = (\beta + 1)^{-1}$, where β is the polynomial decay rate of the eigenvalues of the normalized Gram matrix, perform favorably in comparison with stopping rules, based on hold-out data, and 4-fold cross-validation.

There are various open questions that could be tackled after our results. A deficiency of our strategy is that the construction of τ and τ_α used the assumption that the regression function belongs to a known RKHS, which restricts (mildly) the smoothness of the regression function. We would like to understand how our results could be extended to other loss functions besides the squared loss (for example, in the classification framework), as it was done in [118]. Another research direction would be to use early stopping with fast approximation techniques for kernels [4, 96] to avoid calculation of all eigenvalues of the normalized Gram matrix that can be prohibited for large-scale problems.

Appendix

First, we provide a plan for Appendix to facilitate the reading.

In Appendix 2.7 we state some results that are repeatedly used all along Appendix. Most of them are already known in the literature.

Appendix 2.8 establishes an upper bound on the α-smoothed bias term and provides a deviation inequality for the variance term. These two results will be used throughout Appendix.

In Appendix 2.9 we state an auxiliary lemma of minimax optimality of the stopping rule t^* from Eq. (2.23). This lemma is used in the proof of Corollary 2.3.2.

The main goal of Appendix 2.10 is to provide auxiliary results for the proof of Theorem 2.4.1:

— Lemma 2.10.1 $\longrightarrow$ decomposition of the risk error $\|f^{\tau_\alpha} - f^*\|_n^2$ that involves the following quantities:

$$B^2(\tau_\alpha) \text{ and } v(\tau_\alpha),$$

where $v(t) = \frac{1}{n}\sum_{i=1}^n (\gamma_i^{(\tau_\alpha)})^2 \varepsilon_i^2$ is the stochastic part of the variance at time t from Eq. (2.50).

— Lemma 2.10.2 $\longrightarrow$ the (right) deviation inequality for the stopping rule τ_α.

— Lemma 2.10.3 $\longrightarrow$ the (left) deviation inequality for the stopping rule τ_α.

After that, Lemma 2.10.2 and Lemma 2.10.3 will be used in the following.

— Lemma 2.10.4 will use Lemma 2.10.2 to upper bound $v(\tau_\alpha)$ with high probability.

— Lemma 2.10.5 will use Lemma 2.10.3 to upper bound $B^2(\tau_\alpha)$ with high probability.

Further, we prove Theorem 2.4.1 in Appendix 2.11 by combining all the results from Appendix 2.10.

In Appendix 2.12, one can find the proof of Theorem 2.3.4. To be precise, in this proof, we are able to set $\alpha = 0$ and use the same arguments as in Appendix 2.11. This is the reason why Appendix 2.12 follows Appendix 2.11.

Appendix 2.13 establishes an explicit expression for the smoothed Rademacher compexity $\widehat{R}_{n,\alpha}(\epsilon, \mathcal{H})$.

We collect all the remaining auxiliary lemmas in Appendix 2.14. A sketch of the proof of Lemma 2.5.1 is in Appendix 2.15.

2.7 Useful results

In this section, we present several auxiliary lemmas that are repeatedly used along the chapter.

The first one provides a result showing that we have some coordinates of G^* equal to zero when we transform the initial model (2.5) to its rotated version (2.8).

Lemma 2.7.1. *[92,* Section 4.1.1*] If $f^* \in \mathcal{H}$ with a bounded kernel $\mathbb{K}$ and Gram matrix $K = \mathbb{K}(x_i, x_j),\ i,j = 1,\dots,n$, such that $rk(K) = r \le n$, then*

$$G_i^* = \langle \widehat{u}_i, F^* \rangle = 0 \text{ when } i > r. \tag{2.59}$$

The following auxiliary lemma plays a crucial role in all the proofs. It provides a sharp control of the spectral filter function defined in Eq. (2.11).

Lemma 2.7.2. *[92,* Lemma 8 and Section 4.1.1*] For any bounded kernel, with $\gamma_i^{(t)}$ corresponding to gradient descent or kernel ridge regression, for every $t > 0$,*

$$\frac{1}{2}\min\{1, \eta t \widehat{\mu}_i\} \le \gamma_i^{(t)} \le \min\{1, \eta t \widehat{\mu}_i\}, \quad i = 1,\dots,n; \tag{2.60}$$

$$\frac{1}{n}\sum_{i=1}^{r} \frac{(G_i^*)^2}{\widehat{\mu}_i} \le R^2 \quad \text{and} \quad B^2(t) \le \frac{R^2}{\eta t}. \tag{2.61}$$

Lemma 2.7.3 establishes the magnitude of the population critical radius ϵ_n for different kernel spaces.

Lemma 2.7.3. *[92,* Section 4.3*] Recall the definitions of the localized population Rademacher complexity (2.29) and its population critical radius ϵ_n (2.30), then*

— *for finite-rank kernels with rank r:*

$$\epsilon_n^2 = c_1 \frac{r\sigma^2}{nR^2}$$

for a positive numeric constant $c_1 > 0$.

— *for polynomial eigenvalue decay kernels $\mu_i \le C_\mu i^{-\beta},\ i = 1, 2, \dots$:*

$$\epsilon_n^2 \asymp \left[\frac{\sigma^2}{2R^2 n}\left[1 + \sqrt{\frac{C_\mu}{\beta - 1}}\right]^2\right]^{\frac{\beta}{\beta+1}}. \tag{2.62}$$

Lemma 2.7.4 establishes the magnitude of the empirical critical radius $\widehat{\epsilon}_n$ for different kernel spaces.

Lemma 2.7.4. *Recall the definitions of the empirical localized Rademacher complexity (2.19) and its critical radius (2.20). Then,*

— *for finite-rank kernels with rank r:*

$$\widehat{\epsilon}_n^2 = c\frac{\sigma^2 r}{nR^2} \quad \text{for a positive numeric constant } c.$$

— *for polynomial eigenvalue decay kernels (2.45) with the eigenvalue decay $\beta > 1$:*

$$\widehat{\epsilon}_n^2 \asymp \left[1 + \sqrt{\frac{C}{\beta - 1}}\right]^{\frac{2\beta}{\beta+1}} \left[\frac{\sigma^2}{2nR^2}\right]^{\frac{\beta}{\beta+1}}.$$

Proof of Lemma 2.7.4. The bounds for finite-rank and polynomial eigendecay kernels could be derived in the same manner as in the proof of Lemma 2.7.3, using the upper bound on the eigenvalues $\widehat{\mu}_i \leq Ci^{-\beta}, i = 1, \ldots, r$. ■

The following result shows the magnitude of the smoothed critical radius, defined in Ineq. (2.41), for polynomial eigenvalue decay kernels.

Lemma 2.7.5. *Under the assumption $\widehat{\mu}_i \leq Ci^{-\beta},\ i \in [r]$, for $\alpha\beta < 1$, one has*

$$\widehat{\epsilon}_{n,\alpha}^2 \asymp \left[\sqrt{\frac{C^\alpha}{1-\alpha\beta}} + \sqrt{\frac{C^{1+\alpha}}{\beta(1+\alpha)-1}}\right]^{\frac{2\beta}{\beta+1}} \left[\frac{\sigma^2}{2R^2 n}\right]^{\frac{\beta}{\beta+1}}.$$

Proof of Lemma 2.7.5. For every $M \in (0, r]$ and $\alpha\beta < 1$, we have

$$\begin{aligned}
\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) &\leq R\sqrt{\frac{1}{n}}\sqrt{\sum_{j=1}^{r} \min\{Cj^{-\beta}, \epsilon^2\} C^\alpha j^{-\beta\alpha}} \\
&\leq R\sqrt{\frac{C^\alpha}{n}}\sqrt{\sum_{j=1}^{\lfloor M \rfloor} j^{-\beta\alpha}\epsilon} + R\sqrt{\frac{C^{1+\alpha}}{n}}\sqrt{\sum_{j=\lceil M \rceil}^{n} j^{-\beta-\beta\alpha}} \\
&\leq R\sqrt{\frac{C^\alpha}{1-\alpha\beta}\frac{M^{1-\alpha\beta}}{n}}\epsilon + R\sqrt{\frac{C^{1+\alpha}}{n}}\sqrt{\frac{1}{\beta(1+\alpha)-1}\frac{1}{M^{\beta(1+\alpha)-1}}}.
\end{aligned}$$

Set $M = \epsilon^{-2/\beta}$ that implies $\sqrt{M^{1-\alpha\beta}}\epsilon = \epsilon^{1-\frac{1-\alpha\beta}{\beta}}$, and

$$\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) \leq R\left[\sqrt{\frac{C^\alpha}{1-\alpha\beta}} + \sqrt{\frac{C^{1+\alpha}}{\beta(1+\alpha)-1}}\right]\epsilon^{1-\frac{1-\alpha\beta}{\beta}}\frac{1}{\sqrt{n}}.$$

Therefore, the smoothed critical inequality $\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) \leq \frac{2R^2}{\sigma}\epsilon^{2+\alpha}$ is satisfied for

$$\widehat{\epsilon}_{n,\alpha}^2 \asymp \left[\sqrt{\frac{C^\alpha}{1-\alpha\beta}} + \sqrt{\frac{C^{1+\alpha}}{\beta(1+\alpha)-1}}\right]^{\frac{2\beta}{\beta+1}} \left[\frac{\sigma^2}{2R^2 n}\right]^{\frac{\beta}{\beta+1}}.$$

■

In order to transfer the $L_2(\mathbb{P}_n)$ norm into the $L_2(\mathbb{P}_X)$ norm, we need to relate the empirical critical radius $\widehat{\epsilon}_n$ with its population counterpart ϵ_n. It is achieved by the following result.

Lemma 2.7.6. *There are numeric constants $c_1, c_2, c_3, c_4 > 0$ such that $c_1\epsilon_n \leq \widehat{\epsilon}_n \leq c_2\epsilon_n$ with probability at least $1 - c_3 \exp\Big(-c_4\frac{R^2}{\sigma^2}n\epsilon_n^2\Big)$.*

Sketch of the proof of Lemma 2.7.6. The claim follows from known results on empirical processes and RKHS (see, e.g., Theorem 14.1 and the discussion afterwards in [114]). ■

2.8 Handling the smoothed bias and variance

2.8.1 Upper bound on the smoothed bias

The first lemma provides an upper bound on the smoothed bias term.

Lemma 2.8.1. *Under Assumptions 1, 2,*

$$B_\alpha^2(t) \leq \frac{R^2}{(\eta t)^{1+\alpha}}, \ \alpha \in [0,1]. \tag{2.63}$$

Proof of Lemma 2.8.1. For any $t > 0$,

$$\begin{aligned} B_\alpha^2(t) = \frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha (1-\gamma_i^{(t)})^2 (G_i^*)^2 &\leq \frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha (1-\gamma_i^{(t)})^{1+\alpha} (G_i^*)^2 \\ &\overset{\text{(i)}}{\leq} \frac{1}{n(\eta t)^{1+\alpha}} \sum_{i=1}^r \frac{(G_i^*)^2}{\widehat{\mu}_i} \\ &\overset{\text{(ii)}}{\leq} \frac{R^2}{(\eta t)^{1+\alpha}}. \end{aligned}$$

(i) is true thanks to the qualification condition (2.10) with $\overline{\nu} = 1$, (ii) is due to the bounds in (2.61). ■

2.8.2 Deviation inequality for the variance term

In this subsection, we recall one concentration result from [92, Section 4.1.2].

For any $t > 0$, define the matrix $Q_t \coloneqq \mathrm{diag}\{(\gamma_j^{(t)})^2, \ j \in [r]\}$, then one concludes that $V(t) = \mathbb{E}_\varepsilon[v(t)] = \frac{\sigma^2}{n}\mathrm{tr}[Q_t]$. After that, since $\gamma_i^{(t)} \leq \min\{1, \eta t \widehat{\mu}_i\}$ for $i \in [r]$,

$$V(t) = \frac{\sigma^2}{n}\mathrm{tr}[Q_t] \leq \frac{\sigma^2}{n}\sum_{j=1}^r \min\{1, \eta t \widehat{\mu}_i\} = \frac{\sigma^2 \eta t}{R^2}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\right). \tag{2.64}$$

Consider a random variable of the form $\widetilde{Q}_n = \sum_{i,j=1}^n a_{ij} Z_i Z_j$, where $\{Z_i\}_{i=1}^n$ are zero-mean Gaussian r.v. with parameter σ. Then, [95] proved that

$$\mathbb{P}_\varepsilon\left(|\widetilde{Q}_n - \mathbb{E}_\varepsilon[\widetilde{Q}_n]| \geq \delta\right) \leq 2 \ \exp\left[-c\min\left\{\frac{\delta}{\sigma^2\|A\|_{\mathrm{op}}}, \frac{\delta^2}{\sigma^4\|A\|_F^2}\right\}\right], \ \forall \delta > 0, \tag{2.65}$$

where $\|A\|_{\mathrm{op}}$ and $\|A\|_F$ are the operator and Frobenius norms of the matrix $A = \{a_{ij}\}_{i,j=1}^n$, respectively.

Applying Ineq. (2.65) with $A = \frac{1}{n}Q_t$, $Z_i = \varepsilon_i,\ i \in [r]$, yields $\widetilde{Q}_n = v(t)$, and

$$\begin{aligned} \|A\|_{\mathrm{op}} &\leq \frac{1}{n}, \\ \|A\|_F^2 = \frac{1}{n^2}\mathrm{tr}[Q_t^2] \leq \frac{1}{n^2}\mathrm{tr}[Q_t] &\leq \frac{\eta t}{nR^2}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\right). \end{aligned} \tag{2.66}$$

Consequently, for any $t > 0$ and $\delta > 0$, one gets

$$\mathbb{P}_\varepsilon\Big(|v(t) - V(t)| \geq \delta\Big) \leq 2\ \exp\left[-\frac{cn\delta}{\sigma^2}\min\left\{1, \frac{R^2\delta}{\sigma^2\eta t\widehat{\mathcal{R}}_n^2(\frac{1}{\sqrt{\eta t}}, \mathcal{H})}\right\}\right]. \tag{2.67}$$

Let us first transfer the critical inequality (2.20) from ϵ to t.

Definition 2.8.1. Set $\epsilon = \frac{1}{\sqrt{\eta t}}$ in Eq. (2.20) and let us define $\widehat{t}_\epsilon$ as the largest positive solution to the following fixed-point equation

$$\frac{\sigma^2\eta t}{R^2}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\right) \leq \frac{4R^2}{\eta t}. \tag{2.68}$$

Note that in Lemma 2.7.4, the empirical critical radius $\widehat{\epsilon}_n = \frac{1}{\sqrt{\eta\widehat{t}_\epsilon}}$, and such a point $\widehat{t}_\epsilon$ exists since $\widehat{\epsilon}_n$ exists and is unique [16, 82, 92]. Moreover, $\widehat{t}_\epsilon$ provides the equality in Ineq. (2.68).

2.9 Auxiliary lemma for finite-rank kernels

Remark that at $t = t^* : B^2(t) = \frac{2\sigma^2}{n}\sum_{i=1}^r \gamma_i^{(t)} - V(t) \geq \frac{\sigma^2}{n}\sum_{i=1}^r \gamma_i^{(t)}$. Thus, due to the construction of $\widehat{t}_\epsilon$ ($\widehat{t}_\epsilon$ is the point of intersection of an upper bound on the bias and a lower bound on $\frac{\sigma^2}{2n}\sum_{i=1}^r \gamma_i^{(t)}$) and monotonicity (in t) of all the terms involved, we get $t^* \leq \widehat{t}_\epsilon$.

Lemma 2.9.1. *Recall the definition of stopping rule t^* (2.23). Under Assumptions 1, 2, and 3, for the gradient descent/kernel ridge regression filter, the following holds for any reproducing kernel:*

$$\mathbb{E}_\varepsilon\|f^{t^*} - f^*\|_n^2 \leq 8R^2\widehat{\epsilon}_n^2.$$

Proof of Lemma 2.9.1. Let us define a proxy version of the variance term: $\widetilde{V}(t) \coloneqq \frac{\sigma^2}{n}\sum_{i=1}^r \gamma_i^{(t)}$. Further, for all $t > 0$,

$$\mathbb{E}_\varepsilon R_t = B^2(t) + \frac{\sigma^2}{n}\sum_{i=1}^n (1 - \gamma_i^{(t)})^2. \tag{2.69}$$

From the fact that $\mathbb{E}_\varepsilon R_{t^*} = \sigma^2$,

$$\mathbb{E}_\varepsilon\|f^{t^*} - f^*\|_n^2 = B^2(t^*) + V(t^*) = 2\widetilde{V}(t^*). \tag{2.70}$$

Therefore, in order to prove the lemma, our goal is to get an upper bound on $\widetilde{V}(t^*)$.

Since the function $\eta t\widehat{\mathcal{R}}_n^2(\frac{1}{\sqrt{\eta t}}, \mathcal{H})$ is monotonic in t (see, for example, Lemma 2.14.1), and $t^* \le \widehat{t}_\epsilon$, we conclude that

$$\widetilde{V}(t^*) \le \frac{\sigma^2\eta t^*}{R^2}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta t^*}}, \mathcal{H}\right) \le \frac{\sigma^2\eta \widehat{t}_\epsilon}{R^2}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta \widehat{t}_\epsilon}}, \mathcal{H}\right) = 4R^2\widehat{\epsilon}_n^2.$$

■

2.10 Proofs for polynomial smoothing

In the proofs, we will need three additional definitions below.

Definition 2.10.1. In Definition 2.4.2, set $\epsilon = \frac{1}{\sqrt{\eta t}}$, then the smoothed critical inequality (2.41) is equivalent to

$$\frac{\sigma^2\eta t}{4}\widehat{\mathcal{R}}_{n,\alpha}^2\Big(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\Big) \le \frac{R^4}{(\eta t)^{1+\alpha}}. \tag{2.71}$$

Due to Lemma 2.14.1, the left-hand side of (2.71) is non-decreasing in t, and the right-hand side is non-increasing in t.

Definition 2.10.2. For any $\alpha \in [0, 1]$, define the stopping rule $\widehat{t}_{\epsilon,\alpha}$ such that

$$\widehat{\epsilon}_{n,\alpha}^2 = \frac{1}{\eta \widehat{t}_{\epsilon,\alpha}}, \tag{2.72}$$

then Ineq. (2.71) becomes the equality at $t = \widehat{t}_{\epsilon,\alpha}$ thanks to the monotonicity and continuity of both terms in the inequality.

Further, we define the stopping rules $\widetilde{t}_{\epsilon,\alpha}$ and $\overline{t}_{\epsilon,\alpha}$ that will serve as a lower bound and an upper bound on t^*_α for all $\alpha \in [0, 1]$.

Definition 2.10.3. Define the smoothed proxy variance $\widetilde{V}_\alpha(t) \coloneqq \frac{\sigma^2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \gamma_i^{(t)}$ and the following stopping rules

$$\begin{aligned} \overline{t}_{\epsilon,\alpha} &= \inf\{t > 0 \mid B_\alpha^2(t) = \frac{1}{2}\widetilde{V}_\alpha(t)\}, \\ \widetilde{t}_{\epsilon,\alpha} &= \inf\{t > 0 \mid B_\alpha^2(t) = 3\widetilde{V}_\alpha(t)\}. \end{aligned} \tag{2.73}$$

Notice that at $t = \widetilde{t}_{\epsilon,\alpha}$:

$$\frac{6R^2}{(\eta t)^{1+\alpha}} \ge \frac{R^2}{(\eta t)^{1+\alpha}} \ge B_\alpha^2(t) = 3\widetilde{V}_\alpha(t) \ge \frac{3}{2}\frac{\sigma^2}{R^2}\eta t\widehat{\mathcal{R}}_{n,\alpha}^2\Big(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\Big).$$

At $t = \bar{t}_{\epsilon,\alpha}$:

$$\frac{R^2}{(\eta t)^{1+\alpha}} \geq B_\alpha^2(t) = \frac{1}{2}\widetilde{V}_\alpha(t) \geq \frac{\sigma^2 \eta t}{4R^2}\widehat{\mathcal{R}}_{n,\alpha}^2\Big(\frac{1}{\sqrt{\eta t}}, \mathcal{H}\Big).$$

Thus, $\widetilde{t}_{\epsilon,\alpha}$ and $\bar{t}_{\epsilon,\alpha}$ satisfy the smoothed critical inequality (2.71). Moreover, $\widehat{t}_{\epsilon,\alpha}$ is always greater than or equal to $\bar{t}_{\epsilon,\alpha}$ and $\widetilde{t}_{\epsilon,\alpha}$ since $\widehat{t}_{\epsilon,\alpha}$ is the largest value satisfying Ineq. (2.71). As a consequence of Lemma 2.14.1, one has

$$\frac{1}{\eta\widetilde{t}_{\epsilon,\alpha}} \asymp \frac{1}{\eta\bar{t}_{\epsilon,\alpha}} \asymp \frac{1}{\eta\widehat{t}_{\epsilon,\alpha}} = \widehat{\epsilon}_{n,\alpha}^2.$$

Assume for simplicity that

$$\bar{\epsilon}_{n,\alpha}^2 \coloneqq \frac{1}{\eta\bar{t}_{\epsilon,\alpha}} = c'\frac{1}{\eta\widehat{t}_{\epsilon,\alpha}} = c'\widehat{\epsilon}_{n,\alpha}^2, \quad \text{and}$$
$$\widetilde{\epsilon}_{n,\alpha}^2 \coloneqq \frac{1}{\eta\widetilde{t}_{\epsilon,\alpha}} = c''\frac{1}{\eta\widehat{t}_{\epsilon,\alpha}} = c''\widehat{\epsilon}_{n,\alpha}^2$$

for some positive numeric constants $c', c'' \geq 1$, due to the fact that $\widehat{t}_{\epsilon,\alpha} \geq \bar{t}_{\epsilon,\alpha}$, and $\widehat{t}_{\epsilon,\alpha} \geq \widetilde{t}_{\epsilon,\alpha}$.

The following lemma decomposes the risk error into two parts that will be further analyzed in subsequent Lemmas 2.10.4, 2.10.5.

Lemma 2.10.1. *Recall the definition of τ_α (2.37), then*

$$\|f^{\tau_\alpha} - f^*\|_n^2 \leq 2B^2(\tau_\alpha) + 2v(\tau_\alpha),$$

where $v(t) = \frac{1}{n}\sum_{i=1}^n (\gamma_i^{(t)})^2\varepsilon_i^2,\ t > 0$, is the stochastic part of the variance.

Proof of Lemma 2.10.1. Recall Definition 2.2.1 of the spectral filter function $g_{\lambda_t}(\xi) \equiv g_t(\xi)$.

Let us define the noise vector $\varepsilon \coloneqq [\varepsilon_1, ..., \varepsilon_n]^\top$ and, for each $t > 0$, two vectors that correspond to the bias and variance parts, respectively:

$$\tilde{b}^2(t) \coloneqq (g_t(K_n)K_n - I)F^*,$$
$$\tilde{v}(t) \coloneqq g_t(K_n)K_n\varepsilon.$$

This gives the following expressions for the stochastic part of the variance and bias:

$$v(t) = \langle\tilde{v}(t), \tilde{v}(t)\rangle_n, \quad B^2(t) = \langle\tilde{b}^2(t), \tilde{b}^2(t)\rangle_n. \tag{2.74}$$

General expression for the $L_2(\mathbb{P}_n)$ norm error at τ_α takes the form

$$\|f^{\tau_\alpha} - f^*\|_n^2 = B^2(\tau_\alpha) + v(\tau_\alpha) + 2\langle\tilde{b}^2(\tau_\alpha), \tilde{v}(\tau_\alpha)\rangle_n. \tag{2.75}$$

Therefore, by applying the inequality $2\,|\langle x, y\rangle_n| \leq \|x\|_n^2 + \|y\|_n^2$ for any $x, y \in \mathbb{R}^n$ and (2.74), we obtain

$$\|f^{\tau_\alpha} - f^*\|_n^2 \leq 2B^2(\tau_\alpha) + 2v(\tau_\alpha). \tag{2.76}$$

■

2.10.1 Two deviation inequalities for τ_α

This is the first deviation inequality for τ_α that will be further used in Lemma 2.10.4 to control the variance term.

Lemma 2.10.2. *Recall Definition 2.10.3 of $\bar{t}_{\epsilon,\alpha}$, then under Assumptions 1, 2, 3, 5,*

$$\mathbb{P}_\varepsilon\Big(\tau_\alpha > \bar{t}_{\epsilon,\alpha}\Big) \leq 2\ \exp\left[-c_1\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\right],$$

where positive constant c_1 depends only on $\mathcal{M}$.

Proof of Lemma 2.10.2. Set $\kappa_\alpha \coloneqq \sigma^2 \mathrm{tr}K_n^\alpha/n$, then due to the monotonicity of the smoothed empirical risk, for all $t \geq t_\alpha^*$:

$$\mathbb{P}_\varepsilon\left(\tau_\alpha > t\right) = \mathbb{P}_\varepsilon\left(R_{\alpha,t} - \mathbb{E}_\varepsilon R_{\alpha,t} > \kappa_\alpha - \mathbb{E}_\varepsilon R_{\alpha,t}\right).$$

Consider

$$R_{\alpha,t} - \mathbb{E}_\varepsilon R_{\alpha,t} = \underbrace{\frac{\sigma^2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha(1-\gamma_i^{(t)})^2\left(\frac{\varepsilon_i^2}{\sigma^2}-1\right)}_{\Sigma_1} + \underbrace{\frac{2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha(1-\gamma_i^{(t)})^2 G_i^*\varepsilon_i}_{\Sigma_2}. \tag{2.77}$$

Define

$$\Delta_{t,\alpha} \coloneqq \kappa_\alpha - \mathbb{E}_\varepsilon R_{\alpha,t} = -B_\alpha^2(t) - V_\alpha(t) + 2\widetilde{V}_\alpha(t),$$

where $\widetilde{V}_\alpha(t) = \frac{\sigma^2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha\gamma_i^{(t)}$.

Further, set $t = \bar{t}_{\epsilon,\alpha}$ and recall that $\eta\bar{t}_{\epsilon,\alpha} = \frac{\eta\widehat{t}_{\epsilon,\alpha}}{c'}$ for $c' \geq 1$. This implies

$$\begin{aligned}
\Delta_{\bar{t}_{\epsilon,\alpha},\alpha} \geq \frac{1}{2}\widetilde{V}_\alpha(\bar{t}_{\epsilon,\alpha}) &\geq \frac{\sigma^2}{4n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \min\left\{1, \frac{\eta\widehat{t}_{\epsilon,\alpha}}{c'}\widehat{\mu}_i\right\} \\
&= \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{4nc'}\sum_{i=1}^r \widehat{\mu}_i^\alpha \min\left\{\frac{c'}{\eta\widehat{t}_{\epsilon,\alpha}}, \widehat{\mu}_i\right\} \\
&\geq \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{4c'R^2}\widehat{\mathcal{R}}_{n,\alpha}^2\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\right) \\
&= \frac{R^2}{c'}\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}.
\end{aligned}$$

Then, by standard concentration results on linear and quadratic sums of Gaussian random variables (see, e.g., [28, Lemma 6.1]),

$$\mathbb{P}_\varepsilon\left(\Sigma_1 > \frac{\Delta_{\bar{t}_{\epsilon,\alpha},\alpha}}{2}\right) \leq \exp\left[-\frac{\Delta^2_{\bar{t}_{\epsilon,\alpha},\alpha}}{16(\|a(\bar{t}_{\epsilon,\alpha})\|^2 + \frac{\Delta_{\bar{t}_{\epsilon,\alpha},\alpha}}{2}\|a(\bar{t}_{\epsilon,\alpha})\|_\infty)}\right], \tag{2.78}$$

$$\mathbb{P}_\varepsilon\left(\Sigma_2 > \frac{\Delta_{\bar{t}_{\epsilon,\alpha},\alpha}}{2}\right) \leq \exp\left[-\frac{n\Delta^2_{\bar{t}_{\epsilon,\alpha},\alpha}}{32\sigma^2 B^2_\alpha(\bar{t}_{\epsilon,\alpha})}\right], \tag{2.79}$$

where $a_i(\bar{t}_{\epsilon,\alpha}) = \frac{\sigma^2}{n}\widehat{\mu}_i^\alpha(1-\gamma_i^{(\bar{t}_{\epsilon,\alpha})})^2,\ i \in [r]$.

In what follows, we simplify the bounds above.

Firstly, recall that $B = 1$, which implies $\widehat{\mu}_1 \leq 1$, $\|a(\bar{t}_{\epsilon,\alpha})\|_\infty = \max_{i\in[r]}|a_i(\bar{t}_{\epsilon,\alpha})| \leq \frac{\sigma^2}{n}$, and

$$\begin{aligned}\frac{1}{2}\Delta_{\bar{t}_{\epsilon,\alpha},\alpha} \leq \frac{3}{4}\widetilde{V}_\alpha(\bar{t}_{\epsilon,\alpha}) \leq \frac{3}{4}\widetilde{V}_\alpha(\widehat{t}_{\epsilon,\alpha}) &\leq \frac{3}{4R^2}\sigma^2\eta\widehat{t}_{\epsilon,\alpha}\widehat{\mathcal{R}}^2_{n,\alpha}\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\right)\\ &= 3R^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}.\end{aligned}$$

Secondly, we will upper bound the Euclidean norm of $a(\bar{t}_{\epsilon,\alpha})$. Recall Assumption 5, the definition of the smoothed statistical dimension $d_{n,\alpha} = \min\{j \in [r] \mid \widehat{\mu}_j \leq \widehat{\epsilon}^2_{n,\alpha}\}$, and Ineq. (2.43): $\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha} \geq \frac{\sigma^2\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha}{4R^2 n}$, which implies that

$$\begin{aligned}\|a(\bar{t}_{\epsilon,\alpha})\|^2 = \frac{\sigma^4}{n^2}\sum_{i=1}^r \widehat{\mu}_i^{2\alpha}(1-\gamma_i^{(\bar{t}_{\epsilon,\alpha})})^4 &\leq \frac{\sigma^4}{n^2}\left[\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha + \sum_{i=d_{n,\alpha}+1}^r \widehat{\mu}_i^{2\alpha}\right]\\ &\leq \frac{\sigma^4}{n^2}\left[\frac{4nR^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}}{\sigma^2} + \mathcal{M}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha\right]\\ &\leq \frac{4\sigma^2}{n}\left(1+\mathcal{M}\right)R^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}.\end{aligned}$$

Finally, by using the upper bound $B^2_\alpha(\bar{t}_{\epsilon,\alpha}) \leq \frac{R^2}{(\eta\bar{t}_{\epsilon,\alpha})^{1+\alpha}} \leq R^2(c')^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}$ for all $\alpha \in [0,1]$, one gets

$$\mathbb{P}_\varepsilon\left(\tau_\alpha > \bar{t}_{\epsilon,\alpha}\right) \leq 2\ \exp\left[-c_1\frac{R^2}{\sigma^2}n\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}\right], \tag{2.80}$$

for some positive numeric $c_1 > 0$ that depends only on $\mathcal{M}$.

■

What follows is the second deviation inequality for τ_α that will be further used in Lemma 2.10.5 to control the bias term.

Lemma 2.10.3. *Recall Definition 2.10.3 of $\widetilde{t}_{\epsilon,\alpha}$, then under Assumptions 1, 2, 3, 5,*

$$\mathbb{P}_\varepsilon\left(\tau_\alpha < \widetilde{t}_{\epsilon,\alpha}\right) \le 2\ \exp\left[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\right], \tag{2.81}$$

for positive constant c_2 that depends only on $\mathcal{M}$.

Proof of Lemma 2.10.3. Set $\kappa_\alpha \coloneqq \sigma^2 \mathrm{tr} K_n^\alpha/n$. Note that $\widetilde{t}_{\epsilon,\alpha} \le t_\alpha^*$ by construction. Further, for all $t \le t_\alpha^*$, due to the monotonicity of the smoothed empirical risk,

$$\begin{aligned}
\mathbb{P}_\varepsilon\Big(\tau_\alpha < t\Big) &= \mathbb{P}_\varepsilon\Big(R_{\alpha,t} - \mathbb{E}_\varepsilon R_{\alpha,t} \le -(\mathbb{E}_\varepsilon R_{\alpha,t} - \kappa_\alpha)\Big) \\
&\le \mathbb{P}_\varepsilon\Bigg(\underbrace{\frac{\sigma^2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha (1-\gamma_i^{(t)})^2\Big(\frac{\varepsilon_i^2}{\sigma^2}-1\Big)}_{\Sigma_1} \le -\frac{\mathbb{E}_\varepsilon R_{\alpha,t} - \kappa_\alpha}{2}\Bigg) \\
&+ \mathbb{P}_\varepsilon\Bigg(\underbrace{\frac{2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha (1-\gamma_i^{(t)})^2 G_i^* \varepsilon_i}_{\Sigma_2} \le -\frac{\mathbb{E}_\varepsilon R_{\alpha,t} - \kappa_\alpha}{2}\Bigg).
\end{aligned}$$

Consider $\Delta_{t,\alpha} \coloneqq \mathbb{E}_\varepsilon R_{\alpha,t} - \kappa_\alpha = B_\alpha^2(t) + V_\alpha(t) - 2\widetilde{V}_\alpha(t)$. At $t = \widetilde{t}_{\epsilon,\alpha}$, we have $B_\alpha^2(t) = 3\widetilde{V}_\alpha(t)$, thus

$$\Delta_{\widetilde{t}_{\epsilon,\alpha},\alpha} \ge \widetilde{V}_\alpha(\widetilde{t}_{\epsilon,\alpha}).$$

Then, by standard concentration results on linear and quadratic sums of Gaussian random variables (see, e.g., [28, Lemma 6.1]),

$$\begin{aligned}
\mathbb{P}_\varepsilon\left(\Sigma_1 \le -\frac{\Delta_{\widetilde{t}_{\epsilon,\alpha},\alpha}}{2}\right) &\le \exp\left[-\frac{\widetilde{V}_\alpha^2(\widetilde{t}_{\epsilon,\alpha})}{16\|a(\widetilde{t}_{\epsilon,\alpha})\|^2}\right], \\
\mathbb{P}_\varepsilon\left(\Sigma_2 \le -\frac{\Delta_{\widetilde{t}_{\epsilon,\alpha},\alpha}}{2}\right) &\le \exp\left[-\frac{-n\widetilde{V}_\alpha^2(\widetilde{t}_{\epsilon,\alpha})}{32\sigma^2 B_\alpha^2(\widetilde{t}_{\epsilon,\alpha})}\right],
\end{aligned} \tag{2.82}$$

where $a_i(\widetilde{t}_{\epsilon,\alpha}) = \frac{\sigma^2}{n}\widehat{\mu}_i^\alpha(1-\gamma_i^{(\widetilde{t}_{\epsilon,\alpha})}),\ i \in [r]$.

In what follows, we simplify the bounds above.

First, we deal with the Euclidean norm of $a_i(\widetilde{t}_{\epsilon,\alpha}),\ i \in [r]$. By $\widehat{\mu}_1 \le 1$ and Assumption 5 with Ineq. (2.43), we have

$$\begin{aligned}
\|a(\widetilde{t}_{\epsilon,\alpha})\|^2 = \frac{\sigma^4}{n^2}\sum_{i=1}^r \widehat{\mu}_i^{2\alpha}(1-\gamma_i^{(\widetilde{t}_{\epsilon,\alpha})})^4 &\le \frac{\sigma^4}{n^2}\left[\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha + \sum_{i=d_{n,\alpha}+1}^r \widehat{\mu}_i^{2\alpha}\right] \\
&\le \frac{4\sigma^2}{n}\left(1+\mathcal{M}\right)R^2\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}.
\end{aligned} \tag{2.83}$$

Recall that $\eta\widetilde{t}_{\epsilon,\alpha} = \frac{\eta\widehat{t}_{\epsilon,\alpha}}{c''}$ for $c'' \geq 1$. Therefore, it is sufficient to lower bound $\widetilde{V}_\alpha(\widetilde{t}_{\epsilon,\alpha})$ as follows.

$$\begin{aligned}\widetilde{V}_\alpha(\widetilde{t}_{\epsilon,\alpha}) \geq \frac{\sigma^2}{2n}\sum_{i=1}^{r}\widehat{\mu}_i^\alpha \min\{1, \frac{\eta\widehat{t}_{\epsilon,\alpha}}{c''}\widehat{\mu}_i\} &= \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{2nc''}\sum_{i=1}^{r}\widehat{\mu}_i^\alpha \min\left\{\frac{c''}{\eta\widehat{t}_{\epsilon,\alpha}}, \widehat{\mu}_i\right\} \\ &\geq \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{2R^2c''}\widehat{\mathcal{R}}_{n,\alpha}^2\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\right) \\ &= \frac{2R^2}{c''}\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}.\end{aligned}$$

By using the bound $B_\alpha^2(\widetilde{t}_{\epsilon,\alpha}) \leq \frac{R^2}{(\eta\widetilde{t}_{\epsilon,\alpha})^{1+\alpha}} \leq R^2(c'')^2\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}$ and inserting this expression with (2.83) into (2.82), it gives

$$\mathbb{P}_\varepsilon\left(\tau_\alpha < \widetilde{t}_{\epsilon,\alpha}\right) \leq 2\ \exp\Big[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big], \tag{2.84}$$

where c_2 depends only on $\mathcal{M}$. ■

2.10.2 Bounding the stochastic part of variance term at τ_α

Lemma 2.10.4. *Under Assumptions 1, 2, 3, 4, 5, the stochastic part of the variance at τ_α is bounded as follows.*

$$v(\tau_\alpha) \leq 8(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2$$

with probability at least $1 - 3\exp\Big[-c_1 n\frac{R^2}{\sigma^2}\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$, where constant c_1 depends only on $\mathcal{M}$.

Proof of Lemma 2.10.4. Due to Lemma 2.10.2,

$$\mathbb{P}_\varepsilon\left(\tau_\alpha > \bar{t}_{\epsilon,\alpha}\right) \leq 2\ \exp\left[-c_1\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\right]. \tag{2.85}$$

Therefore, thanks to the monotonicity of $\gamma_i^{(t)}$ in t, with probability at least $1-2\ \exp\Big[-c_1\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$,

$$v(\tau_\alpha) \leq v(\bar{t}_{\epsilon,\alpha}). \tag{2.86}$$

After that, due to the concentration inequality (2.67),

$$\mathbb{P}_\varepsilon\Big(|v(\bar{t}_{\epsilon,\alpha}) - V(\bar{t}_{\epsilon,\alpha})| \geq \delta\Big) \leq 2\ \exp\left[-\frac{cn\delta}{\sigma^2}\min\left\{1, \frac{R^2\delta}{\sigma^2\eta\bar{t}_{\epsilon,\alpha}\widehat{\mathcal{R}}_n^2(\frac{1}{\sqrt{\eta\bar{t}_{\epsilon,\alpha}}}, \mathcal{H})}\right\}\right].$$

Now, by setting $\delta = \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{R^2}\widehat{\mathcal{R}}_n^2\Big(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\Big) \geq \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{R^2}\widehat{\mathcal{R}}_{n,\alpha}^2\Big(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\Big)$ and recalling Lemma 2.14.3, it

yields

$$\begin{aligned} v(\bar{t}_{\epsilon,\alpha}) &\leq V(\bar{t}_{\epsilon,\alpha}) + \delta \\ &\leq \widetilde{V}(\widehat{t}_{\epsilon,\alpha}) + 4(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2 \\ &\leq \frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{R^2}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\right) + 4(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2 \\ &\leq 8(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2 \end{aligned} \tag{2.87}$$

with probability at least $1 - \exp\Big[-cn\frac{4R^2}{\sigma^2}\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$.

Combining all the pieces, we get

$$v(\tau_\alpha) \leq 8(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2 \tag{2.88}$$

with probability at least $1 - 3\exp\Big[-c_1 n\frac{R^2}{\sigma^2}\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$.

■

2.10.3 Bounding the bias term at τ_α

Lemma 2.10.5. *Under Assumptions 1, 2, 3, 5,*

$$B^2(\tau_\alpha) \leq c''R^2\widehat{\epsilon}_{n,\alpha}^2 \tag{2.89}$$

with probability at least $1 - 2\exp\Big[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$*, for positive numeric constant* $c'' \geq 1$ *and constant* c_2 *that depends only on* $\mathcal{M}$*.*

Proof of Lemma 2.10.5. Due to Lemma 2.10.3,

$$\mathbb{P}_\varepsilon\left(\tau_\alpha < \widetilde{t}_{\epsilon,\alpha}\right) \leq 2\ \exp\Big[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]. \tag{2.90}$$

Therefore, thanks to the monotonicity of the bias term, with probability at least $1 - 2\ \exp\Big[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$,

$$B^2(\tau_\alpha) \leq B^2(\widetilde{t}_{\epsilon,\alpha}) \leq \frac{R^2}{\eta\widetilde{t}_{\epsilon,\alpha}} = c''R^2\widehat{\epsilon}_{n,\alpha}^2. \tag{2.91}$$

■

2.11 Proof of Theorem 2.4.1

From Lemmas 2.10.1, 2.10.4, and 2.10.5, we get

$$\|f^{\tau_\alpha} - f^*\|_n^2 \leq 2c'' R^2 \widehat{\epsilon}_{n,\alpha}^2 + 16(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2 \tag{2.92}$$

with probability at least $1 - 5\exp\Big[-c_1 \frac{R^2}{\sigma^2} n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$. Therefore,

$$\|f^{\tau_\alpha} - f^*\|_n^2 \leq c_u R^2 \widehat{\epsilon}_{n,\alpha}^2 \tag{2.93}$$

with probability at least $1 - 5\exp\Big[-c_1 \frac{R^2}{\sigma^2} n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$, where c_1 depends only on $\mathcal{M}$, c_u is a positive constant that depends only on $\mathcal{A}$.

Moreover, taking the expectation in Ineq. (2.76) yields

$$\mathbb{E}_\varepsilon\|f^{\tau_\alpha} - f^*\|_n^2 \leq 2\mathbb{E}_\varepsilon[B^2(\tau_\alpha)] + 2\mathbb{E}_\varepsilon[v(\tau_\alpha)].$$

Let us upper bound $\mathbb{E}_\varepsilon\left[B^2(\tau_\alpha)\right]$ and $\mathbb{E}_\varepsilon\left[v(\tau_\alpha)\right]$. First, define $\widetilde{a} \coloneqq B^2(\widetilde{t}_{\epsilon,\alpha})$, thus

$$\begin{aligned}\mathbb{E}_\varepsilon\Big[B^2(\tau_\alpha)\Big] &= \mathbb{P}_\varepsilon\Big(B^2(\tau_\alpha) > \widetilde{a}\Big)\mathbb{E}_\varepsilon\Big[B^2(\tau_\alpha) \mid B^2(\tau_\alpha) > \widetilde{a}\Big] \\ &\quad + \mathbb{P}_\varepsilon\Big(B^2(\tau_\alpha) \leq \widetilde{a}\Big)\mathbb{E}_\varepsilon\Big[B^2(\tau_\alpha) \mid B^2(\tau_\alpha) \leq \widetilde{a}\Big].\end{aligned} \tag{2.94}$$

Defining $\delta_1 \coloneqq 2\ \exp\Big[-c_2\frac{R^2}{\sigma^2}n\widehat{\epsilon}_{n,\alpha}^{2(1+\alpha)}\Big]$ from Lemma 2.10.5 and using the upper bound

$$B^2(t) \leq \|f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n |f^*(x_i)|^2 = \frac{1}{n}\sum_{i=1}^n |\langle f^*, \mathbb{K}(\cdot, x_i)\rangle_{\mathcal{H}}|^2 \leq R^2$$

for any $t > 0$ gives the following.

$$\mathbb{E}_\varepsilon\Big[B^2(\tau_\alpha)\Big] \leq R^2\delta_1 + B^2(\widetilde{t}_{\epsilon,\alpha}) \leq R^2\left(\delta_1 + c''\widehat{\epsilon}_{n,\alpha}^2\right). \tag{2.95}$$

As for $\mathbb{E}_\varepsilon\left[v(\tau_\alpha)\right]$,

$$\begin{aligned}\mathbb{E}_\varepsilon\left[v(\tau_\alpha)\right] &= \mathbb{E}_\varepsilon\Big[v(\tau_\alpha)\mathbb{I}\Big\{v(\tau_\alpha) \leq 8(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2\Big\}\Big] \\ &\quad + \mathbb{E}_\varepsilon\Big[v(\tau_\alpha)\mathbb{I}\Big\{v(\tau_\alpha) > 8(1+\mathcal{A})R^2\widehat{\epsilon}_{n,\alpha}^2\Big\}\Big],\end{aligned} \tag{2.96}$$

and due to Lemma 2.10.4 and Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}_\varepsilon\left[v(\tau_\alpha)\right] &\le 8(1+\mathcal{A})R^2\widehat{\epsilon}^2_{n,\alpha} + \mathbb{E}_\varepsilon\Big[v(\tau_\alpha)\mathbb{I}\Big\{v(\tau_\alpha) > 8(1+\mathcal{A})R^2\widehat{\epsilon}^2_{n,\alpha}\Big\}\Big]\\ &\le 8(1+\mathcal{A})R^2\widehat{\epsilon}^2_{n,\alpha} + \sqrt{\mathbb{E}_\varepsilon v^2(\tau_\alpha)}\sqrt{\mathbb{E}_\varepsilon\Big[\mathbb{I}\Big\{v(\tau_\alpha) > 8(1+\mathcal{A})R^2\widehat{\epsilon}^2_{n,\alpha}\Big\}\Big]}. \end{aligned} \tag{2.97}$$

Notice that $v^2(\tau_\alpha) \le \frac{1}{n^2}\Big[\sum_{i=1}^r \varepsilon_i^2\Big]^2$, and

$$\mathbb{E}_\varepsilon\Big[v^2(\tau_\alpha)\Big] \le \frac{1}{n^2}\Big[\sum_{i=1}^r \mathbb{E}_\varepsilon\varepsilon_i^4 + 2\sum_{i<j}\mathbb{E}_\varepsilon\Big(\varepsilon_i^2\varepsilon_j^2\Big)\Big] \le \frac{3\sigma^4}{n^2}r^2 \le 3\sigma^4. \tag{2.98}$$

At the same time, thanks to Lemma 2.10.4,

$$\mathbb{E}_\varepsilon\Big[\mathbb{I}\Big\{v(\tau_\alpha) > 8(1+\mathcal{A})R^2\widehat{\epsilon}^2_{n,\alpha}\Big\}\Big] \le 3\exp\Big[-c_1 n\frac{R^2}{\sigma^2}\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}\Big].$$

Thus, inserting the last two inequalities into (2.97) gives

$$\mathbb{E}_\varepsilon\left[v(\tau_\alpha)\right] \le 8(1+\mathcal{A})R^2\widehat{\epsilon}^2_{n,\alpha} + 3\sigma^2\exp\Big[-c_1 n\frac{R^2}{\sigma^2}\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}\Big].$$

Finally, summing up all the terms together,

$$\begin{aligned}\mathbb{E}_\varepsilon\|f^{\tau_\alpha} - f^*\|_n^2 \le \left[16(1+\mathcal{A}) + 2c''\right]R^2\widehat{\epsilon}^2_{n,\alpha}\\ + 6\max\{\sigma^2, R^2\}\exp\left[-c_1 n\frac{R^2}{\sigma^2}\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}\right],\end{aligned}$$

where constant c_1 depends only on $\mathcal{M}$, constant c'' is numeric.

2.12 Proof of Theorem 2.3.4

Here, we prove Theorem 2.3.4 that shows a minimax optimality result for finite-rank kernels with rank r.

Let us proof that $\|f^\tau - f^*\|^2_{\mathcal{H}}$ is upper bounded with high probability by a constant depending only on R. If it is true, we are able to apply Lemma 2.3.3 to transfer the result of Corollary 2.3.2 to the $L_2(\mathbb{P}_X)$ norm. In order to do that, it is sufficient to upper bound $\|f^\tau\|^2_{\mathcal{H}}$ because $\|f^\tau - f^*\|^2_{\mathcal{H}} \le \|f^\tau\|^2_{\mathcal{H}} + R^2$.

We will use the definition of τ (2.24) with the threshold $\kappa \coloneqq \frac{r\sigma^2}{n}$ so that, due to the monotonicity

of the "reduced" empirical risk $\widetilde{R}_t$,

$$\mathbb{P}_\varepsilon(\tau > t) = \mathbb{P}_\varepsilon\Big(\widetilde{R}_t - \mathbb{E}_\varepsilon \widetilde{R}_t > \underbrace{\kappa - \mathbb{E}_\varepsilon \widetilde{R}_t}_{\Delta_t}\Big),$$

where

$$\Delta_t = -B^2(t) - V(t) + \underbrace{\frac{2\sigma^2}{n}\sum_{i=1}^{r}\gamma_i^{(t)}}_{2\widetilde{V}(t)}. \tag{2.99}$$

Assume that $\Delta_t \geq 0$. Remark that

$$\widetilde{R}_t - \mathbb{E}_\varepsilon \widetilde{R}_t = \underbrace{\frac{\sigma^2}{n}\sum_{i=1}^{r}(1-\gamma_i^{(t)})^2\left(\frac{\varepsilon_i^2}{\sigma^2}-1\right)}_{\Sigma_1} + \underbrace{\frac{2}{n}\sum_{i=1}^{r}(1-\gamma_i^{(t)})^2 G_i^*\varepsilon_i}_{\Sigma_2}. \tag{2.100}$$

Applying [28, Lemma 6.3] to Σ_1 yields

$$\mathbb{P}_\varepsilon\left(\Sigma_1 > \frac{\Delta_t}{2}\right) \leq \exp\left[\frac{-\Delta_t^2/4}{4(\|a(t)\|^2 + \frac{\Delta_t}{2}\|a(t)\|_\infty)}\right], \tag{2.101}$$

where $a_i(t) \coloneqq \frac{\sigma^2}{n}(1-\gamma_i^{(t)})^2,\ i \in [r]$.

Standard concentration bound [114, Proposition 2.5] for a sum of Gaussian variables Σ_2 gives us

$$\mathbb{P}_\varepsilon\left(\Sigma_2 > \frac{\Delta_t}{2}\right) \leq \exp\left[-\frac{n\Delta_t^2}{32\sigma^2 B^2(t)}\right]. \tag{2.102}$$

First, define the stopping rule $\bar{t}_\epsilon$ as follows.

$$\bar{t}_\epsilon \coloneqq \inf\{t > 0 : B^2(t) = \frac{1}{2}\widetilde{V}(t)\}. \tag{2.103}$$

Note that $\bar{t}_\epsilon$ serves as an upper bound on t^* and as a lower bound on $\widehat{t}_\epsilon$ (2.8.1). Moreover, $\bar{t}_\epsilon$ satisfies the critical inequality (2.68). Therefore due to Lemma 2.14.1, there is a positive numeric constant $c' \geq 1$ such that $\frac{1}{\eta\bar{t}_\epsilon} = c'\frac{1}{\eta\widehat{t}_\epsilon}$.

In what follows we simplify two high probability bounds (2.101) and (2.102) at $t = \bar{t}_\epsilon$.

Since from Lemma 2.7.4, $\widehat{\epsilon}_n^2 = c\frac{r\sigma^2}{nR^2}$, one can bound $\|a(\bar{t}_\epsilon)\|^2$ as follows.

$$\|a(\bar{t}_\epsilon)\|^2 = \frac{\sigma^4}{n^2}\sum_{i=1}^{r}(1-\gamma_i^{(\bar{t}_\epsilon)})^4 \leq \frac{r\sigma^4}{n^2} = \frac{R^2\sigma^2\widehat{\epsilon}_n^2}{cn}. \tag{2.104}$$

Remark that in Eq. (2.101):

$$\|a(\bar{t}_\epsilon)\|_\infty = \frac{\sigma^2}{n}\max_{i\in[r]}\Big[(1-\gamma_i^{(\bar{t}_\epsilon)})\Big] \leq \frac{\sigma^2}{n},$$

and

$$\frac{\Delta_{\bar{t}_\epsilon}}{2} \leq \frac{3}{4}\widetilde{V}(\bar{t}_\epsilon) \leq \frac{3}{4}\widetilde{V}(\widehat{t}_\epsilon) \leq \frac{3}{4}\frac{\sigma^2}{R^2}\eta\widehat{t}_\epsilon\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta\widehat{t}_\epsilon}},\mathcal{H}\right) = 3R^2\widehat{\epsilon}_n^2.$$

As for a lower bound on $\Delta_{\bar{t}_\epsilon}$,

$$\begin{aligned}\Delta_{\bar{t}_\epsilon} \geq \frac{1}{2}\widetilde{V}(\bar{t}_\epsilon) \geq \frac{\sigma^2}{4n}\sum_{i=1}^r \min\left\{1, \frac{\eta\widehat{t}_\epsilon}{c'}\widehat{\mu}_i\right\} &= \frac{\sigma^2\eta\widehat{t}_\epsilon}{4nc'}\sum_{i=1}^r \min\left\{\frac{c'}{\eta\widehat{t}_\epsilon}, \widehat{\mu}_i\right\} \\ &\geq \frac{\sigma^2\eta\widehat{t}_\epsilon}{4R^2c'}\widehat{\mathcal{R}}_n^2\left(\frac{1}{\sqrt{\eta\widehat{t}_\epsilon}},\mathcal{H}\right) \\ &= \frac{R^2}{c'}\widehat{\epsilon}_n^2.\end{aligned}$$

Knowing that $B^2(\bar{t}_\epsilon) \leq \frac{R^2}{\eta\bar{t}_\epsilon} = c'R^2\widehat{\epsilon}_n^2$ and summing up bounds (2.101), (2.102) with $t = \bar{t}_\epsilon$ yield the following.

$$\mathbb{P}_\varepsilon\left(\tau > \bar{t}_\epsilon\right) \leq 2\ \exp\Big[-C\frac{R^2}{\sigma^2}n\widehat{\epsilon}_n^2\Big], \tag{2.105}$$

where C is a numeric constant.

From Lemma 2.14.2, $\|f^{\bar{t}_\epsilon}\|_\mathcal{H} \leq \sqrt{7}R$ with probability at least $1 - 4\ \exp\Big[-c_3\frac{R^2}{\sigma^2}n\widehat{\epsilon}_n^2\Big]$, for some positive numeric constant c_3. Therefore, Ineq. (2.105) allows to say:

$$\|f^\tau\|_\mathcal{H} \leq \sqrt{7}R \text{ with probability at least } 1 - 6\ \exp\Big[-\tilde{c}_3\frac{R^2}{\sigma^2}n\widehat{\epsilon}_n^2\Big],\ \text{for a positive } \tilde{c}_3.$$

It implies that

$$\|f^\tau - f^*\|_\mathcal{H} \leq \|f^\tau\|_\mathcal{H} + \|f^*\|_\mathcal{H} \leq \left(1+\sqrt{7}\right)R$$

with the same probability.

Thus, according to Lemma 2.3.3, for some positive numeric constants $c_1, \tilde{c}_4, \tilde{c}_5$:

$$\|f^\tau - f^*\|_2^2 \leq 2\|f^\tau - f^*\|_n^2 + c_1R^2\epsilon_n^2$$

with probability (w.r.t. ε) at least $1 - 6\exp\Big[-\tilde{c}_3\frac{R^2}{\sigma^2}n\widehat{\epsilon}_n^2\Big]$ and with probability (w.r.t. $\{x_i\}_{i=1}^n$) at least $1 - \tilde{c}_4\ \exp\Big[-\tilde{c}_5\frac{R^2}{\sigma^2}n\epsilon_n^2\Big]$.

Moreover, by following the same arguments (with $\alpha = 0$ and without Assumptions 4 and 5) as in

the proof of Theorem 2.4.1, Lemma 2.7.6 yields

$$\|f^\tau - f^*\|_n^2 \leq c_u R^2 \widehat{\epsilon}_n^2 \leq \widetilde{c}_u R^2 \epsilon_n^2 \tag{2.106}$$

with probability at least $1 - c_1 \exp\Big[- c_2 \frac{R^2}{\sigma^2} n \epsilon_n^2 \Big]$.

The last step consists of recalling $R^2 \epsilon_n^2 \lesssim \frac{r\sigma^2}{n}$ due to Lemma 2.7.4.

2.13 Derivation of the smoothed empirical kernel complexity

In this section, we will show that

$$\mathbb{E}_{\mathbf{r}} \left[\sup_{\substack{\|f\|_{\mathcal{H}} \leq R \\ \|f\|_{n,\alpha} \leq R\epsilon}} | \frac{1}{n} \sum_{i=1}^n \widehat{\mu}_i^\alpha \mathbf{r}_i f(x_i) | \right] \leq R \sqrt{\frac{2}{n} \sum_{i=1}^r \widehat{\mu}_i^\alpha \min\{\epsilon^2, \widehat{\mu}_i\}}, \tag{2.107}$$

where $\{\mathbf{r}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \{-1, +1\}$ with probability $1/2$, $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq \ldots \geq \widehat{\mu}_r > 0$. The derivation of this result is inspired by [114, Lemma 13.22].

Define auxiliary random variables $\{\widetilde{\mathbf{r}}_i\}_{i=1}^n \overset{i.i.d.}{\sim} \{-\widehat{\mu}_i^\alpha, \widehat{\mu}_i^\alpha\}$ with probability $1/2$ for $i \in [n]$.

To start with, we recall that $B = 1$, thus $\widehat{\mu}_1 \leq 1$.

Since $f \in \mathcal{H}$, without loss of generality, we are able to restrict ourselves to the functions that take the form

$$f(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \theta_i \mathbb{K}(\cdot, x_i) \tag{2.108}$$

for some vector $\theta \in \mathbb{R}^n$. The condition $\|f\|_{n,\alpha} \leq \epsilon R$ is equivalent to $\|K_n^{1+\frac{\alpha}{2}} \theta\| \leq \epsilon R$. At the same time, the condition $\|f\|_{\mathcal{H}}^2 \leq R^2$ is equivalent to $\|f\|_{\mathcal{H}}^2 = \theta^\top K_n \theta \leq R^2$. Therefore, the smoothed localized Rademacher complexity could be expressed as

$$\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) = \frac{1}{\sqrt{n}} \mathbb{E}_{\widetilde{\mathbf{r}}} \left[\sup_{\substack{\theta^\top K_n \theta \leq R^2 \\ \theta^\top K_n^{2+\alpha} \theta \leq \epsilon^2 R^2}} | \widetilde{\mathbf{r}}^\top K_n \theta | \right]. \tag{2.109}$$

Recall the SVD decomposition $K_n = U \Lambda U^\top$, where $\Lambda = \operatorname{diag}\{\widehat{\mu}_1, \ldots, \widehat{\mu}_r, 0, \ldots, 0\}$. Therefore if $\beta = K_n \theta$, after some algebra Eq. (2.109) is equivalent to

$$\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) = \frac{1}{\sqrt{n}} \mathbb{E}_{\widetilde{\mathbf{r}}} \Big[\sup_{\beta \in \mathcal{D}} | \widetilde{\mathbf{r}}^\top \beta | \Big],$$

where

$$\mathcal{D} \coloneqq \left\{ \beta \in \mathbb{R}^n \mid \sum_{j=1}^{r} \widehat{\mu}_j^{\alpha} \beta_j^2 \leq \epsilon^2 R^2, \sum_{j=1}^{r} \frac{\beta_j^2}{\widehat{\mu}_j} \leq R^2 \right\}$$

Define the ellipsoid

$$\mathcal{E} \coloneqq \{\beta \in \mathbb{R}^n \mid \sum_{j=1}^{r} \eta^j \beta_j^2 \leq 2R^2\}, \quad \text{where } \eta^j = \max\left\{ \frac{\widehat{\mu}_j^{\alpha}}{\epsilon^2}, \widehat{\mu}_j^{-1} \right\}.$$

Notice that

$$\max\left\{ \frac{\widehat{\mu}_i^{\alpha}}{\epsilon^2}, \widehat{\mu}_i^{-1} \right\} \leq \frac{\widehat{\mu}_i^{\alpha}}{\epsilon^2} + \frac{1}{\widehat{\mu}_i}, \ i \in [r]. \tag{2.110}$$

Thus, $\mathcal{D} \subset \mathcal{E}$, and by Hölder's inequality,

$$\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) \leq \frac{1}{\sqrt{n}} \mathbb{E}_{\widetilde{\mathbf{r}}} \Big[\sup_{\beta \in \mathcal{E}} \mid \langle \widetilde{\mathbf{r}}, \beta \rangle \mid \Big] \leq R\sqrt{\frac{2}{n}} \mathbb{E}_{\widetilde{\mathbf{r}}} \sqrt{\sum_{i=1}^{r} \frac{\widetilde{\mathbf{r}}_i^2}{\eta^i}}. \tag{2.111}$$

By applying Jensen's inequality, it gives us

$$\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) \leq R\sqrt{\frac{2}{n}} \sqrt{\sum_{i=1}^{r} \frac{\mathbb{E}_{\widetilde{\mathbf{r}}}[\widetilde{\mathbf{r}}_i^2]}{\eta^i}} = R\sqrt{\frac{2}{n}} \sqrt{\sum_{i=1}^{r} \frac{\widehat{\mu}_i^{2\alpha}}{\eta^i}}, \tag{2.112}$$

where $\frac{1}{\eta^i} = \min\{\widehat{\mu}_i^{-\alpha}\epsilon^2, \widehat{\mu}_i\} \leq \widehat{\mu}_i^{-\alpha} \min\{\epsilon^2, \widehat{\mu}_i\}$, which leads to the claim.

2.14 Auxiliary results

Lemma 2.14.1. *Under Assumptions 1 and 2, for any $\alpha \in [0,1]$, the function $\epsilon \mapsto \frac{\widehat{\mathcal{R}}_{n,\alpha}(\epsilon,\mathcal{H})}{\epsilon}$ is non-increasing (as a function of ϵ) on the interval $(0, +\infty)$, and consequently, for any numeric constant $c > 0$, the inequality*

$$\frac{\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H})}{\epsilon} \leq c\frac{R^2}{\sigma}\epsilon^{1+\alpha} \tag{2.113}$$

has a smallest positive solution. In addition to that, $\widehat{\epsilon}_{n,\alpha}$ (2.41) exists and is unique.

Proof of Lemma 2.14.1. We will prove that $\widehat{\epsilon}_{n,\alpha}$ lies in the interval $(0, +\infty)$ and is unique. Recall the definition of $\widehat{\epsilon}_{n,\alpha}$:

$$\widehat{\epsilon}_{n,\alpha} = \min\left\{ \epsilon > 0 \mid \sum_{i=1}^{r} \widehat{\mu}_i^{\alpha} \min\{1, \widehat{\mu}_i \epsilon^{-2}\} \leq \frac{4R^4 n}{\sigma^2} \epsilon^{2+2\alpha} \right\}.$$

Note that $f(\epsilon) \coloneqq \sum_{i=1}^{r} \widehat{\mu}_i^{\alpha} \min\{1, \widehat{\mu}_i \epsilon^{-2}\}$ is non-increasing in ϵ, whereas $g(\epsilon) \coloneqq \frac{4R^4 n}{\sigma^2}\epsilon^{2+2\alpha}$ is increasing in ϵ. For $\epsilon \to 0$: $g(\epsilon) = 0 < f(\epsilon)$, and for $\epsilon \to \infty$: $g(\epsilon) > f(\epsilon)$. It proves that $\widehat{\epsilon}_{n,\alpha}$ exists, and due to

continuity of $\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H})$ w.r.t. ϵ, $\widehat{\epsilon}_{n,\alpha}$ is unique and satisfies

$$\frac{1}{\epsilon}\widehat{\mathcal{R}}_{n,\alpha}(\epsilon, \mathcal{H}) = \frac{2R^2}{\sigma}\epsilon^{1+\alpha}.$$

■

The following result establishes a condition, under which we can upper bound the $\mathcal{H}$-functional norm of f^t. It implies that the uniform norm of f^t is upper bounded, and we are at the point to change the $L_2(\mathbb{P}_n)$ and $L_2(\mathbb{P}_X)$ norms, with high probability, due to Lemma 2.3.3. The proof of the lemma below is inspired by Lemma 9 in [92].

Lemma 2.14.2. *Recall Definition 2.10.3 of $\bar{t}_{\epsilon,\alpha}$ and the discussion afterwards. Assume $\alpha \in [0, 1]$, then there exists a universal constant $c > 0$ such that, for all $t \leq \bar{t}_{\epsilon,\alpha}$: $\|f^t\|_{\mathcal{H}} \leq 7R^2$ with probability at least $1 - 4\exp(-cn\widehat{\epsilon}^2_{n,\alpha})$, where $\widehat{\epsilon}_{n,\alpha}$ is the smoothed empirical critical radius.*

Proof of Lemma 2.14.2. For any $t > 0$, let us write the following: f^t lies in $\mathcal{H}$, therefore it can be decomposed via the eigenvectors $\{\phi_k\}_{k=1}^{\infty}$ of the kernel integral operator T_k as

$$f^t = \sum_{k=1}^{\infty} \sqrt{\mu_k} a_k \phi_k \quad \text{such that} \quad \|f^t\|^2_{\mathcal{H}} = \sum_{k=1}^{\infty} a_k^2. \tag{2.114}$$

Consider the linear operator $\Phi_X : \ell^2(\mathbb{N}) \to \mathbb{R}^n$ defined via $[\Phi_X]_{jk} = \phi_j(x_k)$ and the diagonal operator $D : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$, with $[D]_{jj} = \mu_j$ and $[D]_{jk} = 0$ for $j \neq k$. Since $K_n = \frac{1}{n}\Phi_X D \Phi_X^\top$ and $a = \frac{1}{n} D^{\frac{1}{2}} \Phi_X^\top K_n^{-1} F^t$ (see [92, Lemma 9] for the derivation), one deduces an explicit expression for the $\mathcal{H}$-norm of f^t

$$\|f^t\|^2_{\mathcal{H}} = \|a\|^2_{\ell^2(\mathbb{N})} = \frac{1}{n}[F^t]^\top K_n^{-1} F^t.$$

Recall the SVD decomposition $K_n = U \Lambda U^\top$ with $\Lambda = \operatorname{diag}(\widehat{\mu}_1, \ldots, \widehat{\mu}_r)$, and $U^\top F^t = (I - S^t) U^\top Y$, where $I - S^t = \operatorname{diag}\{\gamma_i^{(t)},\ i \in [r]\}$. It gives the following:

$$\begin{aligned}
\|f^t\|^2_{\mathcal{H}} = \frac{1}{n} Y^\top U (I - S^t)^2 \Lambda^{-1} U^\top Y = &\underbrace{\frac{2}{n}\overline{\varepsilon}^\top U (I - S^t)^2 \Lambda^{-1} U^\top F^*}_{\mathcal{A}_t} \\
&+ \underbrace{\frac{1}{n}\overline{\varepsilon}^\top U (I - S^t)^2 \Lambda^{-1} U^\top \overline{\varepsilon}}_{\mathcal{B}_t} \\
&+ \underbrace{\frac{1}{n}[F^*]^\top U (I - S^t)^2 \Lambda^{-1} U^\top F^*}_{\mathcal{C}_t},
\end{aligned}$$

where $S^t = \operatorname{diag}\{1 - \gamma_i^{(t)},\ i \in [r]\}$. Firstly, by using Eq. (2.61), $\mathcal{C}_t \leq R^2$ for any $t > 0$.

Bounding $\mathcal{A}_t$. $\varepsilon = U^\top \bar{\varepsilon}$ is a zero-mean Gaussian vector with parameter σ, thus

$$\mathbb{P}_\varepsilon\Big(|\mathcal{A}_t| \geq R^2\Big) \leq 2\ \exp\Big(-\frac{n}{2\sigma^2\nu^2}\Big) \leq 2\ \exp\Big(-\frac{nR^2}{8\sigma^2\eta\bar{t}_{\epsilon,\alpha}}\Big) = 2\ \exp\Big(-\frac{c'R^2}{8\sigma^2}n\widehat{\epsilon}^2_{n,\alpha}\Big),$$

where the last inequality comes from

$$\begin{aligned}
\nu^2 &= \frac{4}{nR^4}[F^*]^\top U(I-S^t)^4\Lambda^{-2}U^\top F^* \\
&\overset{\text{(i)}}{\leq} \frac{4}{nR^4}\sum_{i=1}^r \frac{(G_i^*)^2}{\widehat{\mu}_i^2}\min\{1,\eta t\widehat{\mu}_i\} \\
&\leq \frac{4\eta t}{nR^4}\sum_{i=1}^r \frac{(G_i^*)^2}{\widehat{\mu}_i} \\
&\overset{\text{(ii)}}{\leq} \frac{4\eta t}{R^2}.
\end{aligned}$$

(i) is true since $(\gamma_i^{(t)})^4 \leq \gamma_i^{(t)} \leq \min\{1,\eta t\widehat{\mu}_i\},\ i \in [r]$. The upper bound (ii) was due to the bound (2.61).

Bounding $\mathcal{B}_t$:

$$\mathcal{B}_t = \frac{1}{n}\sum_{i=1}^r \frac{(\gamma_i^{(t)})^2}{\widehat{\mu}_i}\varepsilon_i^2.$$

Let us define the matrix $Q_t \coloneqq \operatorname{diag}\Big\{\frac{(\gamma_i^{(t)})^2}{\widehat{\mu}_i},\ i \in [r]\Big\}$. Now, we will bound the quadratic form $\mathcal{B}_t$ by utilizing the following concentration result [95]: there is a universal constant $c > 0$ such that

$$\mathbb{P}_\varepsilon\Big(|\mathcal{B}_t - \mathbb{E}_\varepsilon\mathcal{B}_t| \geq R^2\Big) \leq 2\ \exp\left[-c\ \min\left(\frac{nR^2}{\sigma^2}\|UQ_tU^\top\|_{\text{op}}^{-1}, \frac{n^2R^4}{\sigma^4}\|UQ_tU^\top\|_F^{-2}\right)\right].$$

In the following, we will bound $\mathbb{E}_\varepsilon\mathcal{B}_t$, $\|UQ_tU^\top\|_{\text{op}}$, and $\|UQ_tU^\top\|_F$.

Bounding the mean $\mathbb{E}_\varepsilon\mathcal{B}_t = \frac{\sigma^2}{n}\sum_{i=1}^r \frac{(\gamma_i^{(t)})^2}{\widehat{\mu}_i}$. So, using $\gamma_i^{(t)} \leq \min\{1,\eta t\widehat{\mu}_i\}$, we can write the following

$$\begin{aligned}
\mathbb{E}_\varepsilon\mathcal{B}_t &\leq \frac{\sigma^2}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \frac{\min^2\{1,\eta\bar{t}_{\epsilon,\alpha}\widehat{\mu}_i\}}{\widehat{\mu}_i^{1+\alpha}} \\
&\leq \frac{\sigma^2(\eta\bar{t}_{\epsilon,\alpha})^{2+\alpha}}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \min\left\{\frac{1}{(\eta\bar{t}_{\epsilon,\alpha})(\eta\bar{t}_{\epsilon,\alpha}\widehat{\mu}_i)^{1+\alpha}}, (\eta\bar{t}_{\epsilon,\alpha}\widehat{\mu}_i)^{1+\alpha}\widehat{\mu}_i\right\} \\
&\leq \frac{\sigma^2}{R^2}(\eta\bar{t}_{\epsilon,\alpha})^{2+\alpha}\widehat{\mathcal{R}}^2_{n,\alpha}\Big(\frac{1}{\sqrt{\eta\bar{t}_{\epsilon,\alpha}}}, \mathcal{H}\Big) \leq 4R^2.
\end{aligned}$$

Bounding the operator and Frobenius norms. For the normalized operator norm:

$$\frac{1}{n}\|UQ_tU^\top\|_{\mathrm{op}} = \frac{1}{n}\max_{j\in[r]}\Big[\frac{(\gamma_j^{(t)})^2}{\widehat{\mu}_j}\Big] \le \frac{\eta\bar{t}_{\epsilon,\alpha}}{n} = \frac{1}{c'n\widehat{\epsilon}^2_{n,\alpha}}.$$

As for the normalized Frobenius norm,

$$\begin{aligned}
\frac{1}{n}\|UQ_tU^\top\|_F^2 = \frac{1}{n}\sum_{i=1}^r \frac{(\gamma_i^{(t)})^4}{\widehat{\mu}_i^2} &\le \frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \frac{\min^4\{1,\eta\bar{t}_{\epsilon,\alpha}\widehat{\mu}_i\}}{\widehat{\mu}_i^{2+\alpha}} \\
&\le \frac{1}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \min\Big\{\frac{1}{\widehat{\mu}_i^{2+\alpha}}, (\eta\bar{t}_{\epsilon,\alpha})^4\widehat{\mu}_i^{2-\alpha}\Big\} \\
&= \frac{(\eta\bar{t}_{\epsilon,\alpha})^{3+\alpha}}{n}\sum_{i=1}^r \widehat{\mu}_i^\alpha \min\Big\{\frac{1}{(\eta\bar{t}_{\epsilon,\alpha})(\eta\bar{t}_{\epsilon,\alpha}\widehat{\mu}_i)^{2+\alpha}}, \\
&\quad (\eta\bar{t}_{\epsilon,\alpha}\widehat{\mu}_i)^{1-\alpha}\widehat{\mu}_i\Big\} \\
&\le \frac{1}{R^2}(\eta\bar{t}_{\epsilon,\alpha})^{3+\alpha}\widehat{\mathcal{R}}^2_{n,\alpha}\Big(\frac{1}{\sqrt{\eta\bar{t}_{\epsilon,\alpha}}}, \mathcal{H}\Big) \le \frac{4R^2}{\sigma^2c'\widehat{\epsilon}^2_{n,\alpha}}.
\end{aligned}$$

Finally, we are able to conclude that there exists a numeric constant $c > 0$ such that it holds

$$\mathbb{P}_\varepsilon\Big(|\mathcal{B}_t - \mathbb{E}_\varepsilon\mathcal{B}_t| \ge R^2\Big) \le 2\ \exp\Big[-c\frac{R^2}{\sigma^2}n\widehat{\epsilon}^2_{n,\alpha}\Big].$$

By combining all the pieces, there exists a numeric constant $\widetilde{\widetilde{c}}_1 > 0$ such that

$$\mathbb{P}_\varepsilon\Big(|\mathcal{B}_t| \ge 5R^2 \text{ or } |\mathcal{A}_t| \ge R^2\Big) \le 2\ \exp\left(-\widetilde{\widetilde{c}}_1\frac{R^2}{\sigma^2}n\widehat{\epsilon}^2_{n,\alpha}\right).$$

■

The following lemma establishes a connection between the smoothed critical inequality and its non-smooth version.

Lemma 2.14.3. *Under Assumptions 1, 2, 3, 4, $\widehat{t}_{\epsilon,\alpha}$ from Definition 2.10.2 satisfies*

$$\frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{4R^2}\widehat{\mathcal{R}}^2_n\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}}, \mathcal{H}\right) \le \frac{(1+\mathcal{A})R^2}{\eta\widehat{t}_{\epsilon,\alpha}}. \tag{2.115}$$

Thus, $\widehat{t}_{\epsilon,\alpha}$ provides a smallest positive solution to the non-smooth version of the critical inequality.

Proof of Lemma 2.14.3. On one side, let us start by recalling that

$$\frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{4R^2}\widehat{\mathcal{R}}^2_{n,\alpha}\Big(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}},\mathcal{H}\Big) = R^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}.$$

Then for $d_{n,\alpha} = \min\{j \in [r] \mid \widehat{\mu}_j \leq \widehat{\epsilon}^2_{n,\alpha}\}$,

$$\begin{aligned}\frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{4R^2}\widehat{\mathcal{R}}^2_{n,\alpha}\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}},\mathcal{H}\right) &= \frac{\sigma^2}{4n\widehat{\epsilon}^2_{n,\alpha}}\sum_{i=1}^r \widehat{\mu}_i^\alpha \min\{\widehat{\mu}_i, \widehat{\epsilon}^2_{n,\alpha}\} \\ &= \frac{\sigma^2}{4n\widehat{\epsilon}^2_{n,\alpha}}\left[\widehat{\epsilon}^2_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha + \sum_{i=d_{n,\alpha}+1}^r \widehat{\mu}_i^{1+\alpha}\right] \\ &= R^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}.\end{aligned} \tag{2.116}$$

The last two lines of (2.116) yield

$$\frac{\sigma^2}{4n\widehat{\epsilon}^2_{n,\alpha}} = \frac{R^2\widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}}{\widehat{\epsilon}^2_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha + \sum_{i=d_{n,\alpha}+1}^r \widehat{\mu}_i^{1+\alpha}}. \tag{2.117}$$

On the other side, consider the left-hand part of the non-smooth version of the critical inequality (2.71) at $t = \widehat{t}_{\epsilon,\alpha}$:

$$\begin{aligned}\frac{\sigma^2\eta\widehat{t}_{\epsilon,\alpha}}{4R^2}\widehat{\mathcal{R}}^2_{n}\left(\frac{1}{\sqrt{\eta\widehat{t}_{\epsilon,\alpha}}},\mathcal{H}\right) &= \frac{\sigma^2}{4n\widehat{\epsilon}^2_{n,\alpha}}\sum_{i=1}^r \min\{\widehat{\mu}_i, \widehat{\epsilon}^2_{n,\alpha}\} \\ &= R^2\frac{\sum_{i=1}^{d_{n,\alpha}}\widehat{\epsilon}^{4+2\alpha}_{n,\alpha} + \widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}\sum_{i=d_{n,\alpha}+1}^r\widehat{\mu}_i}{\widehat{\epsilon}^2_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha + \sum_{i=d_{n,\alpha}+1}^r\widehat{\mu}_i^{1+\alpha}} \\ &\leq R^2\frac{\sum_{i=1}^{d_{n,\alpha}}\widehat{\epsilon}^{4+2\alpha}_{n,\alpha} + \widehat{\epsilon}^{2(1+\alpha)}_{n,\alpha}\sum_{i=d_{n,\alpha}+1}^r\widehat{\mu}_i}{\widehat{\epsilon}^2_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha}.\end{aligned} \tag{2.118}$$

Notice that $\widehat{\mu}_i \geq \widehat{\epsilon}^2_{n,\alpha}$ and $\widehat{\mu}_i^\alpha \geq \widehat{\epsilon}^{2\alpha}_{n,\alpha}$ for $i \leq d_{n,\alpha}$. This implies $\sum_{i=1}^{d_{n,\alpha}}\widehat{\epsilon}^{4+2\alpha}_{n,\alpha} \leq \widehat{\epsilon}^4_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha$, and also that $\sum_{i=d_{n,\alpha}+1}^r\widehat{\mu}_i \leq \mathcal{A}\widehat{\epsilon}^{2(1-\alpha)}_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha$, using Assumption 4. Hence,

$$\widehat{\epsilon}^{2\alpha}_{n,\alpha}\sum_{i=d_{n,\alpha}+1}^r\widehat{\mu}_i \leq \mathcal{A}\widehat{\epsilon}^2_{n,\alpha}\sum_{i=1}^{d_{n,\alpha}}\widehat{\mu}_i^\alpha,$$

which leads to the desired upper bound with $\widehat{\epsilon}^2_{n,\alpha} = (\eta \widehat{t}_{\epsilon,\alpha})^{-1}$:

$$\frac{\sigma^2 \eta \widehat{t}_{\epsilon,\alpha}}{4R^2} \widehat{\mathcal{R}}_n^2 \left(\frac{1}{\sqrt{\eta \widehat{t}_{\epsilon,\alpha}}}, \mathcal{H} \right) \leq (1 + \mathcal{A}) R^2 \widehat{\epsilon}^2_{n,\alpha}.$$

■

2.15 Proof of Lemma 2.5.1

Let us prove the lemma only for kernel ridge regression. Notice that

$$\mathbb{E}_\varepsilon \left[\frac{R_{1,t}}{1/n \sum_{i=1}^r \widehat{\mu}_i (1 - \gamma_i^{(t)})^2} \right] = \sigma^2 + \frac{B_1^2(t)}{\frac{1}{n} \sum_{i=1}^r \widehat{\mu}_i (1 - \gamma_i^{(t)})^2}. \tag{2.119}$$

From Lemma 2.8.1, $B_1^2(t) \leq \frac{R^2}{(\eta t)^2}$. As for the denominator,

$$\frac{1}{n} \sum_{i=1}^r \widehat{\mu}_i (1 - \gamma_i^{(t)})^2 \geq \frac{c}{n} \sum_{i=1}^r i^{-\beta} \frac{1}{(1 + \eta c i^{-\beta} t)^2}.$$

Define an index $i_0 \in [r]$ such that $\eta c i^{-\beta} t \geq 1$ if $i \leq i_0$, and $\eta c i^{-\beta} t < 1$ if $i > i_0$, hence

$$\frac{c}{n} \sum_{i=1}^r \widehat{\mu}_i (1 - \gamma_i^{(t)})^2 \geq \frac{1}{4nc\eta^2 t^2} \sum_{i=1}^{i_0} i^\beta + \frac{c}{4n} \sum_{i=i_0+1}^r i^{-\beta} \geq \frac{1}{4nc\eta^2 t^2} \sum_{i=1}^{i_0} i^\beta. \tag{2.120}$$

By lower bounding the last sum in (2.120), we achieve the following

$$\sum_{i=1}^{i_0} i^\beta \geq \int_0^{i_0} x^\beta dx = \frac{i_0^{\beta+1}}{\beta + 1}.$$

Assume that t is large enough, meaning that $i_0 > (\eta c t)^{\frac{1}{\beta}} - 1 \geq \frac{1}{2} (\eta c t)^{\frac{1}{\beta}}$, thus we obtain

$$\frac{B_1^2(t)}{\frac{1}{n} \sum_{i=1}^r \widehat{\mu}_i (1 - \gamma_i^{(t)})^2} \leq \frac{2^{\beta+3} (\beta + 1) c n R^2}{(\eta c t)^{1 + \frac{1}{\beta}}}.$$

In addition to that, one knows (see Lemma 2.7.4) that $(\widehat{\epsilon}_n^2)^{-1} = \eta \widehat{t}_\epsilon \asymp \frac{1}{\left[1 + \sqrt{\frac{C}{\beta - 1}}\right]^{\frac{2\beta}{\beta - 1}}} \left[\frac{2nR^2}{\sigma^2} \right]^{\frac{\beta}{\beta+1}}$, therefore for all $t \gtrsim \widehat{t}_\epsilon$,

$$\frac{B_1^2(t)}{\frac{1}{n} \sum_{i=1}^r \widehat{\mu}_i (1 - \gamma_i^{(t)})} \lesssim \frac{2^{\beta+2} (\beta + 1) \sigma^2}{c^{1/\beta}} \left[1 + \sqrt{\frac{C}{\beta - 1}} \right]^{\frac{2(\beta+1)}{\beta - 1}}. \tag{2.121}$$

Chapter 3

Minimum discrepancy principle for choosing k in k-NN regression

Abstract

This chapter presents a novel data-driven strategy to choose the tuning parameter k in the k-NN regression estimator. We treat the problem of choosing the tuning parameter as an iterative procedure (over k) and propose using an easily implemented in practice strategy based on the idea of *early stopping* and the *minimum discrepancy principle*. This estimation strategy is proven to be minimax optimal, under the fixed-design assumption on covariates, over a range of smoothness function classes, for instance, the Lipschitz functions class on a bounded domain. The proof relies on careful analysis of the bias-variance trade-off and standard concentration inequalities for linear/quadratic forms of Gaussian variables. The novelty of the strategy comes from reducing the computational time of model selection while preserving the statistical (minimax) optimality. In particular, if one should choose k among $\{1, \dots, n\}$, the strategy reduces the computational time of the generalized cross-validation, AIC or Mallow's C_p criteria from $\mathcal{O}\left(n^3\right)$ to $\mathcal{O}\left(n^2(n-k)\right)$, where k is the proposed (minimum discrepancy principle) value of the nearest neighbors.

3.1 Introduction

Nonparametric regression estimation is a fundamental problem in statistics and machine learning. The k-NN regression estimator [26, 67] is a very simple and popular choice in practice. For this estimator, the central issue is choosing properly the number of neighbors k.

The theoretical performance of the k-NN regression estimator has been widely studied since the 1970s [24, 25, 26, 52, 55, 56, 73, 128]. For example, in [26, Chapter 12] the uniform consistency of the k-NN estimator is proved under the condition that $k(n)/n \to 0$ as $n \to \infty$, where n is the sample size. However, as it was shown in [67], the nearest neighbor estimator ($k = 1$) is proved to be consistent only in the noiseless case. Therefore, it is necessary to let k grow with n.

Recently, researchers started to be interested in choosing k optimally from the data [7, 13, 67, 73]. Apparently, the most common (and the simplest) strategy to choose k is to assume some smoothness assumption on the regression function (e.g., the Lipschitz condition [67]), and to find k that makes an

upper bound on the bias and the variance of the k-NN regression estimator equal. This method has a clear lack: one needs to know the smoothness of the regression function (e.g., the Lipschitz constant). The seminal paper [7] gives a *data-driven strategy* for choosing a tuning parameter with different linear estimators (e.g., the k-NN estimator) developed from the idea of minimal penalty, introduced previously in [27]. The main inconvenience of this strategy is that one needs to compute all estimators $\mathbf{F}_n = \{f^k,\ k = 1, \dots, n\}$ of the regression function in order to choose the optimal one among them (by comparing them via a criterion). To list other (similar) strategies, one can think about Akaike's AIC [2] or BIC [100] criterion or generalized cross-validation [70, 76] where one has to compute the empirical risk error plus a penalty term for any $k = 1, \dots, n$. Often, it is computationally expensive and restricts the use in practice. This gives rise to the problem of choosing the tuning parameter "in real-time", meaning that the practitioner should compute iteratively $f^k \in \mathbf{F}_n$. Eventually, this iterative process has to be stopped. This problem can be solved by applying the *early stopping rule*.

Review on early stopping rule

The early stopping rule (ESR) is a regularization method that consists in stopping an iterative learning algorithm prior to its convergence. The main idea of ESR is preserving statistical optimality while lowering the computational complexity of a learning algorithm. Early stopping dates back to the 1970s and was originally proposed for solving ill-posed operator (matrix) problems (see the book [60] for a thorough review on the subject). After that, there was a great interest in applying early stopping to train artificial neural networks [91]. The main concern of this heuristics was to show that during the training phase of learning one can benefit from leaving apart a part of the data called the validation data. This way, the validation error on this part should give an approximation of the true risk error. This approach was purely practical, and until the 2000s there were no theoretical justifications for the ESR at all. Furthermore, until the work [92], all the developed stopping rules [17, 39, 122, 126] were not data-dependent. Raskutti et al. [92] proposed using the so-called localized Rademacher complexities [16, 114] to recover the bias-variance trade-off for two learning algorithms: gradient descent and ridge (Tikhonov) regression in the unit ball of Reproducing kernel Hilbert space $\mathcal{H}$. The subsequent work [118] extended the previous result to boosting algorithms with the same idea of controlling the localized Gaussian complexities in RKHS. The main inconvenience was that the results in [92] and [118] were derived under the assumption that the regression function lies in the unit ball of $\mathcal{H}$, which restricts the use of these stopping rules in practice.

The first early stopping rule that could be potentially data-driven was proposed by [28, 30, 50] for spectral filter iterative algorithms (see, e.g. [19, 64] for examples of such algorithms). The idea behind the construction of this early stopping rule is the so-called *minimum discrepancy principle* that is based on finding a first iteration for which a learning algorithm starts to fit the noise. The key quantity for the analysis of the minimum discrepancy principle is the *empirical risk error* (the train error in the terminology of the machine learning community), which is monitored throughout

the whole learning process. The process thus is stopped if the empirical risk starts to fit the noise.

Contribution. In the present paper, we propose applying the minimum discrepancy principle stopping rule for the k-NN regression estimator in order to select k. We prove a non-asymptotic bound on the performance of the minimum dicrepancy principle stopping rule measured in the empirical $L_2(\mathbb{P}_n)$ norm. This bound implies that, under a quite mild assumption on the regression function, the minimum discrepancy principle stopping rule provides a minimax-optimal functional estimator, in particular, over the class of Lipschitz functions on a bounded domain.

Outline of the chapter. The organization of the chapter is as follows. Section 3.2 describes the statistical model, its main assumption, and introduces the notation that will be used in the chapter. In Section 3.3, we introduce the k-NN estimator and explain how to compute the minimum discrepancy early stopping rule. Section 3.4 provides the main theoretical result that shows that the proposed rule achieves statistical optimality over a range of functional classes (e.g., the well-known class of Lipschitz functions on a bounded domain). Section 3.5 is devoted to the discussion of the obtained results. All the technical proofs are in Appendix.

3.2 Statistical model, main assumption and notation

In the nonparametric regression setting, we work with a sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X}^n \times \mathbb{R}^n$ that satisfies the statistical model

$$y_i = f^*(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $f^* : \mathcal{X} \mapsto \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$, is a measurable function on some set $\mathcal{X}$, and $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Gaussian noise variables $\mathcal{N}(0, \sigma^2)$. Assume that the parameter $\sigma^2 > 0$ is fixed and known (except for Section 4.4 where we consider real-data simulated experiments). One should point out here that the assumption of known σ^2 is quite typical in model selection literature with nonparametric regression setting (see, e.g., [42, 76, 77, 119]). In addition to that, we assume that $\{x_i \in \mathcal{X}\}_{i=1}^n$ are *fixed* covariates (corresponds to the so-called fixed design setting), thus we observe noise only in the responses $\{y_i\}_{i=1}^n$. The goal of the present chapter is to estimate optimally the regression function f^*. The term "optimally" will be explained in Section 3.3.

In the context of the *fixed design* setting, the performance of an estimator $\widehat{f}$ of f^* is measured in terms of the so-called *empirical norm* defined as

$$\|\widehat{f} - f^*\|_n^2 := \frac{1}{n} \sum_{i=1}^n \left[\widehat{f}(x_i) - f^*(x_i)\right]^2, \tag{3.2}$$

where $\|h\|_n := \sqrt{1/n \sum_{i=1}^n h(x_i)^2}$ for any bounded on $\mathcal{X}$ function h. We denote the empirical norm as $L_2(\mathbb{P}_n)$. For each bounded over $\mathcal{X}$ functions h_1, h_2, $\langle h_1, h_2 \rangle_n$ denotes the related inner product defined

as $\langle h_1, h_2 \rangle_n \coloneqq 1/n \sum_{i=1}^n h_1(x_i) h_2(x_i)$. Further, $\mathbb{P}_\varepsilon$ and $\mathbb{E}_\varepsilon$ denote the probability and expectation with respect to $\{\varepsilon_i\}_{i=1}^n$.

Notation. Throughout the chapter, $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ are the usual Euclidean norm and inner product in $\mathbb{R}^n$. $\|M\|_2$ and $\|M\|_F$ signify the operator and Frobenius norms of the matrix $M \in \mathbb{R}^{n \times n}$, respectively. We denote the trace of the matrix M by $\mathrm{tr}(M)$. In addition to that, $\mathbb{I}\{\mathcal{E}\}$ is equal to 1 if the probabilistic event $\mathcal{E}$ holds true, otherwise it is equal to 0. For $a \geq 0$, we denote by $\lfloor a \rfloor$ the largest natural number that is smaller than or equal to a. We denote by $\lceil a \rceil$ the smallest natural number that is greater than or equal to a.

We make the following assumption on the regression function f^* introduced earlier in Eq. (3.1).

Assumption 6 (Boundness of the r.f.)**.** *f^* is bounded on $\mathcal{X}$, meaning that there exists a constant $\mathcal{M} > 0$ such that*

$$|f^*(x)| \leq \mathcal{M} \qquad \text{for all } x \in \mathcal{X}. \tag{3.3}$$

Assumption 6 is quite standard in the nonparametric regression literature [67, 128]. In particular, Assumption 6 holds when the set $\mathcal{X}$ is bounded, and the regression function f^* is L-Lipschitz with some positive constant L [67].

Along the chapter, we use the notation $c, c_1, C, \widetilde{c}, \widetilde{C}, \ldots$ to show that the numeric constants $c, c_1, C, \widetilde{c}, \widetilde{C}, \ldots$ can depend only on d, σ, and $\mathcal{M}$. The values of all the constants may change from line to line or even in the same line.

3.3 k-NN estimator and minimum discrepancy stopping rule

3.3.1 k-NN regression estimator

Let us transform the initial model (3.1) into its vector form

$$Y = [y_1, \ldots, y_n]^\top = F^* + \varepsilon \in \mathbb{R}^n, \tag{3.4}$$

where the vectors $F^* \coloneqq [f^*(x_1), \ldots, f^*(x_n)]^\top$ and $\varepsilon \coloneqq [\varepsilon_1, \ldots, \varepsilon_n]^\top$.

Define a k-nearest neighbor estimator f^k of f^* from (3.1) at the point x_i, $i = 1, \ldots, n$, as

$$f^k(x_i) \coloneqq F_i^k = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} y_j, \qquad k = 1, \ldots, n, \tag{3.5}$$

where $\mathcal{N}_k(i)$ are the indices of the k nearest neighbors of x_i among $\{1, \ldots, n\}$ in the usual Euclidean norm in $\mathbb{R}^d$, where ties are broken at random. In words, in Eq. (3.5), one weights by $1/k$ the response y_j if x_j is a k nearest neighbor of x_i, measured in the Euclidean norm. Note that other adaptive metrics (instead of the Euclidean one) have been also considered in the literature [70, Chap. 14].

One can notice that the k-NN regression estimator (3.5) belongs to the class of linear estimators [6, 70], meaning that the vector $F^k \in \mathbb{R}^n$ estimates the vector F^* as it follows.

$$F^k \coloneqq \Big(f^k(x_1), \dots, f^k(x_n)\Big)^\top = A_k Y, \tag{3.6}$$

where $A_k \in \mathbb{R}^{n\times n}$ is the matrix described below.

$$\begin{cases} \forall 1 \le i,j \le n,\ (A_k)_{ij} \in \{0, 1/k\} \text{ with } k \in \{1,\dots,n\}, \\ \forall 1 \le i \le n,\ (A_k)_{ii} = 1/k \text{ and } \sum_{j=1}^n (A_k)_{ij} = 1. \end{cases} \tag{3.7}$$

Saying differently, $(A_k)_{ij} = 1/k$ if x_j is a k nearest neighbor of x_i, otherwise $(A_k)_{ij} = 0$, $i,j \in \{1,\dots,n\}$.

Define the mean-squared error (the risk error) of the estimator f^k as

$$\mathrm{MSE}(k) \coloneqq \mathbb{E}_\varepsilon \|f^k - f^*\|_n^2 = \frac{1}{n}\mathbb{E}_\varepsilon \sum_{i=1}^n \Big(\frac{1}{k}\sum_{j\in\mathcal{N}_k(i)} y_j - f^*(x_i)\Big)^2. \tag{3.8}$$

Further, we will introduce the (squared) bias and variance of the functional estimator f^k (see, e.g. [6, Eq. (7)]),

$$\mathrm{MSE}(k) = B^2(k) + V(k), \tag{3.9}$$

where

$$B^2(k) = \|(I_n - A_k)F^*\|_n^2, \qquad V(k) = \frac{\sigma^2}{n}\mathrm{tr}\left(A_k^\top A_k\right).$$

Moreover, we are able to simplify a bit the expression for the variance, which shows the lemma below.

Lemma 3.3.1 (Proposition 1 in [6])**.** *For any $k \in \{1,\dots,n\}$,*

$$V(k) = \frac{\sigma^2}{n}\mathrm{tr}(A_k) = \frac{\sigma^2}{k}.$$

Proof of Lemma 3.3.1. Notice that

$$\mathrm{tr}\left(A_k^\top A_k\right) = \mathrm{tr}\left(A_k A_k^\top\right) = \sum_{i=1}^n\sum_{j=1}^n (A_k)_{ij}^2 = \frac{n}{k}. \tag{3.10}$$

■

Thus, due to Lemma 3.3.1, the variance term σ^2/k is a decreasing function of k. Note that $B^2(1) = 0$, $V(1) = \sigma^2$, and $B^2(n) = (1-1/n)^2\|f^*\|_n^2$, $V(n) = \sigma^2/n$. Importantly, the bias term $B^2(k)$ can have *arbitrary* behavior on the interval $[1,n]$.

Ideally, we would like to minimize the mean-squared error (3.8) as a function of k. However, since

the bias term is not known (it contains the unknown regression function), one should introduce other quantities that will be related to the bias. In our case, this quantity will be the *empirical risk* at k:

$$R_k := \|(I_n - A_k)Y\|_n^2. \tag{3.11}$$

R_k measures how well the estimator f^k fits Y. Remark that $R_1 = 0$ (corresponds to the "overfitting" regime) and $R_n = (1 - 1/n)^2 \frac{1}{n} \sum_{i=1}^n y_i^2$ (corresponds to the "underfitting" regime), but there is no information about the monotonicity of R_k on the interval $[1, n]$.

Furthermore, some information about the bias is contained in the expectation of the empirical risk. To be precise, since $Y = F^* + \varepsilon$ and $\operatorname{tr}\left(A_k^\top A_k\right) = \operatorname{tr}(A_k)$, for any $k \in \{1, \dots, n\}$,

$$\begin{aligned} \mathbb{E}_\varepsilon R_k &= \sigma^2 + B^2(k) - \frac{\sigma^2(2\operatorname{tr}(A_k) - \operatorname{tr}(A_k^\top A_k))}{n} \\ &= \sigma^2 + B^2(k) - \frac{\sigma^2}{n}\operatorname{tr}(A_k) \\ &= \sigma^2 + B^2(k) - V(k). \end{aligned} \tag{3.12}$$

Let us illustrate all the mentioned quantities in one example in Fig 3.1. We take the regression function equal to $f^*(x) = \|x - 0.5\|/\sqrt{3} - 0.5$, the noise variance $\sigma = 0.1$. We take $n = 50$, $x_i \overset{i.i.d.}{\sim} \mathbb{U}[0,1]^3$, $i = 1, \dots, n$, and plot the bias term $B^2(k)$, the variance term $V(k)$, the risk error $\mathrm{MSE}(k)$, the empirical risk R_k, and its expectation $\mathbb{E}_\varepsilon R_k$ versus the number of neighbours k. We start with the maximum number of neighbours $k_{\max} = n/2$ and decrease it until $k = 1$. By doing that, one is able to increase successively the complexity of the model measured by its "degree of freedom" [7, 70] $\operatorname{tr}(A_k) = n/k$.

Note that among all defined quantities, only the variance term can be proved monotonic (without an additional assumption on the smoothness of f^*). Importantly, Fig 3.1 indicates that choosing $k = 5$ will provide the user with the global optimum of the risk (the mean-squared error) curve. Thus, for instance, it would be meaningless (according to the risk curve) to compute all the estimators f^k (3.5) for $k = 1, \dots, 5$.

Our main concern is to design a data-driven strategy to choose $\widehat{k} \in \{1, \dots, n\}$, which can be seen as a mapping from the data $\{(x_i, y_i)\}_{i=1}^n$ to a positive integer number so that the prediction error (the mean squared error) $\mathbb{E}_\varepsilon \|f^{\widehat{k}} - f^*\|_n^2$ is as small as possible. To be precise, the goal is to define a data-driven $\widehat{k}$ such that it satisfies the following upper bound [108, 114]

$$\|f^{\widehat{k}} - f^*\|_n^2 \le C_n \mathbb{E}_\varepsilon \|f^{k_{\text{opt}}} - f^*\|_n^2 + r_n, \tag{3.13}$$

with high (exponential) probability over $\{\varepsilon_i\}_{i=1}^n$, where $f^{k_{\text{opt}}}$ is a minimax optimal estimator of $f^* \in \mathcal{F}$, $\mathcal{F}$ is some a priori chosen function space. The leading constant C_n should be bounded and not depend on the regression function f^*, the remainder term r_n is negligible (smaller) with respect to $\mathbb{E}_\varepsilon \|f^{k_{\text{opt}}} - f^*\|_n^2$.

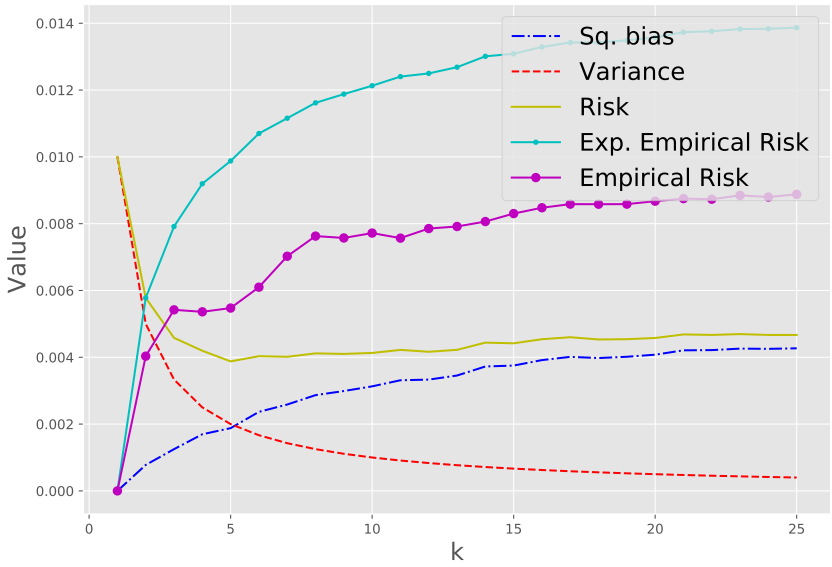


Figure 3.1 – Sq. bias, variance, risk, and (expected) empirical risk behavior.

3.3.2 Related work

The idea of choosing the tuning parameter $k \in \{1, \dots, n\}$ from the data has been already considered in the literature. For example, the classical procedures such as generalized cross-validation [42, 53, 76], penalized estimators [7, 10, 76, 79] and different cross-validation methods [8] are popular choices for linear estimators. Let us consider them in more detail.

Generalized CV [42, 70, 76]. This model selection method has been widely studied in the case of (kernel) ridge regression [53] and smoothing splines [42]. In particular, [42] proved a non-asymptotic oracle inequality for the generalized CV estimator when the variance σ^2 is known. However, in a more general case, GCV estimates σ^2 implicitly, which is an advantage of the method. In addition to that, GCV for k-NN regression is proved [76] to be asymptotically optimal under the assumption $\|A_k\|_2 \leq c,\ \forall k = 1, \dots, n$, for some positive constant c. It is worth to mention that generalized cross-validation provides an approximation to the so-called leave-one-out cross-validation [8, 48], which is an exhaustive model selection procedure. In this case, if the nearest neighbors' matrices are precomputed, the computational time is reduced to $\mathcal{O}(n^3)$. The GCV strategy will be later considered in our simulations in Chapter 4.

Penalized estimators date back to the works on the AIC [1] or Mallow's C_p [79] criteria, where a penalty proportional to the dimension of the model is added to the least-squares loss (i.e., *the empirical risk* in our notation (3.11)) when the noise level σ^2 is known. As for GCV strategy, the computational time of AIC and Mallow's C_p are $\mathcal{O}(n^3)$. After that, a new approach was developed by [27], where

the authors introduced the so-called "slope heuristics" for projection matrices. This notion relied on the introduction of the penalty $\text{pen}(k) = K\text{tr}(A_k)$, where $\text{tr}(A_k)$ is the "dimension of the model", and K is a constant that depends on σ^2, in particular. It appeared that there exists a constant $K_{\min}$ such that $2K_{\min}\text{tr}(A_k)$ yields an asymptotically optimal model selection procedure. This gives rise to some strategies for the estimation of the constant $K_{\min}$ from the data, as it was done, for instance, in [6] for a general linear estimator when σ^2 is unknown.

Cross-validation methods [8]. These model selection methods are the most used in practice. Compared to generalized cross-validation, for instance, V-fold cross-validation method [8, 63] incurs a large computational cost (with V, which is not too small). To be precise, V-fold cross-validation requires the model selection procedure to be performed V times for each value of $k \in \{1, \dots, n\}$. Another alternative could be the Hold-out method [8, 117], which consists in randomly splitting the data into two parts for each value $k \in \{1, \dots, n\}$: one is dedicated to training the estimator (3.5), and the other one is dedicated to testing.

3.3.3 Minimum discrepancy principle rule

In this section, we present a minimum discrepancy principle stopping rule.

We are at the point to define our first reference rule. Based on the nonparametric statistics literature [108, 116], the bias-variance trade-off usually provides an optimal functional estimator:

$$k^* = \inf\left\{k \in \{1, \dots, n\} \mid B^2(k) \geq V(k)\right\}. \tag{3.14}$$

In general, the bias-variance trade-off stopping rule k^* does not exist due to an arbitrary behaviour of the bias term $B^2(k)$. Thus, if no such k^* exists, set $k^* = n$. If it exists, then $k^* \geq 2$ since $V(1) = \sigma^2 > B^2(1) = 0$.

Notice that the stopping rule k^* is not computable in practice since it depends on the unknown bias. Nevertheless, we can create a data-driven version of k^* using the empirical risk R_k.

Eq. (3.12) gives us that the event $\{B^2(k) \geq V(k)\}$ is equivalent to the event $\{\mathbb{E}_\varepsilon R_k \geq \sigma^2\}$, so we conclude that $k^* = \inf\{k \in \{1, \dots, n\} \mid \mathbb{E}_\varepsilon R_k \geq \sigma^2\}$. It gives rise to an estimator of k^* that we denote as k^τ. This stopping rule is called *the minimum discrepancy principle* stopping rule and is defined as

$$\boxed{k^\tau \coloneqq \sup\left\{k \in \{1, \dots, n\} \mid R_k \leq \sigma^2\right\}}. \tag{3.15}$$

Remark. If no such k^τ exists, then set $k^\tau = 1$. Note that in Eq. (3.15), we introduced a supremum instead of the infimum from Eq. (3.14). That was done on purpose because there could be several points of the bias-variance trade-off, and the bias (and the empirical risk) could behave badly in the areas "in-between". In order to calculate k^τ, the user should, first, compute the empirical risk R_k at $k = n$ (thus, the matrix A_n of n nearest neighbors). After that, one needs to decrease k until the event

$\{R_k \leq \sigma^2\}$ holds. It is worth to mention that it is not necessary to compute explicitly all the matrices $A_k,\ k = n,\ n-1, \ldots$, since, for instance, the matrix A_{n-1} could be easily derived from the matrix A_n (assuming that one has already arranged the neighbors and removed the n^{th} neighbors from the matrix A_n), i.e.,

$$[A_{n-1}]_{ij} = \frac{n}{n-1}[A_n]_{ij},\ \forall i, j \in \{1, \ldots, n\}. \tag{3.16}$$

It is one of the main computational advantages of the proposed rule (3.15). For more details on the efficient computation of the nearest neighbors, see, e.g., [22, 87]. In addition to all of that, we emphasize that the definition (3.15) of k^τ does not require the knowledge of the constant $\mathcal{M}$ from Assumption 6, and k^τ does not require computing the empirical risk R_k for all values $k = 1, \ldots, n$, as it is the case, for instance, for generalized cross-validation or Mallow's C_p (see Section 3.3.2). Moreover, we need to point out that the stopping rule (3.15) depends on the noise level σ^2, which should be estimated, as for AIC or Mallow's C_p criteria [2, 70, 79]. We provide a consistent estimator of σ^2 in Chapter 4. Regarding the computational time of k^τ, if the nearest neighbors' matrices are already computed, it is of the order $\mathcal{O}\left(n^2\left(n - k^\tau\right)\right)$, which is less than $\mathcal{O}\left(n^3\right)$ for AIC/Mallow's C_p or GCV.

There is a large amount of literature [19, 28, 30, 50, 60] on the minimum discrepancy principle for spectral filter algorithms such as gradient descent, ridge (Tikhonov) regularization, and spectral cut-off regression (e.g., [28, 50] provide a thorough review). We should emphasize that intuitively the minimum discrepancy principle determines the first iteration (time) at which a learning algorithm starts to fit noise, which is measured by σ^2 in the present context.

Moreover, one is able to notice that if the empirical risk is close to its expectation, k^τ should produce an optimal estimator in some sense. The main question that should be asked is "In which setting is it possible to quantify this gap between R_k and $\mathbb{E}_\varepsilon R_k$ that will not be statistically large?". This question is the main technical obstacle of the present chapter. In what follows, we show that for a quite large class of functions, k^τ is optimal in the sense of Ineq. (3.13).

3.4 Theoretical optimality result

Let us start to describe the main theoretical result of the chapter. The following theorem applies to the estimator defined in Eq. (3.6).

Theorem 3.4.1. *Under Assumption 6, for arbitrary $u \geq 0$,*

$$\|f^{k^\tau} - f^*\|_n^2 \leq 8V(k^*) + C_1\left(\frac{u}{n} + \frac{\sqrt{u}}{\sqrt{n}}\right) + C_2\sqrt{\frac{\log n}{n}} \tag{3.17}$$

with probability at least $1 - 16\exp(-u)$, where positive constants C_1, C_2 can depend on d, σ, and $\mathcal{M}$.

Moreover, if k^* from Eq. (3.14) exists, then for arbitrary $u \geq 0$,

$$\|f^{k^\tau} - f^*\|_n^2 \leq \underbrace{4\ \mathrm{MSE}(k^*)}_{\text{Main term}} + \underbrace{C_1\left(\frac{u}{n} + \frac{\sqrt{u}}{\sqrt{n}}\right) + C_2\sqrt{\frac{\log n}{n}}}_{\text{Rem. term}} \tag{3.18}$$

with probability at least $1 - 16\exp(-u)$, where the constants C_1, C_2 are from Ineq. (3.17).

Sketch of the proof of Theorem 3.4.1. The full proof is deferred to Appendix 3.11. Let us provide a sketch of the proof here.

The main ingredients of the proof are two deviation inequalities: for any $x \geq 0$,

$$\mathbb{P}_\varepsilon\left(V(k^\tau) > 2V(k^*) + x\right) \leq 2\exp\left(-\tilde{c}n\min\left(x^2, x\right)\right), \tag{3.19}$$

and

$$B^2(k^\tau) \leq 2V(k^*) + c_1\sqrt{\frac{\log n}{n}} + 2x, \tag{3.20}$$

where Ineq. (3.20) holds with probability at least $1 - 12\exp\left(-cn\min\left(x^2, x\right)\right)$.

After that, one can split the $L_2(\mathbb{P}_n)$ error at k^τ into two parts:

$$\|f^{k^\tau} - f^*\|_n^2 \leq 2B^2(k^\tau) + 2\|A_{k^\tau}\varepsilon\|_n^2. \tag{3.21}$$

By considering Eq. (3.21), it is sufficient to derive high probability control of $\sup_k \left|\|A_k\varepsilon\|_n^2 - V(k)\right|$ for $k = 1, \ldots, n$ (see Appendix 3.8). That was the reason why the term $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$ appeared in Eq. (3.17).

Finally, one can combine Eq. (3.21), Eq. (3.20), and Eq. (3.19), and apply $V(k^*) \leq \frac{1}{2}\mathrm{MSE}(k^*)$, if k^* exists, and $u = cn\min\left(x^2, x\right)$. The claim follows.

■

In order to gain some intuition of the claim of Theorem 3.4.1, let us make some comments.

First of all, Ineq. (3.18) is non-asymptotic, meaning that it holds true for any $n \geq 1$. Second, Ineq. (3.18) holds with high (exponential) probability, which is a stronger result than in expectation since [76] there are model selection procedures that are asymptotically optimal with high probability but not in expectation.

Third, the main term in Ineq. (3.18) is the risk error at the bias-variance trade-off times 4 (this constant could be improved). Ideally, one should rather introduce the oracle risk $\inf_{k=1,\ldots,n} \mathbb{E}_\varepsilon\|f^k - f^*\|_n^2$ and compare $\|f^{k^\tau} - f^*\|_n^2$ with it. However, to the best of our knowledge, a smoothness assumption is needed to connect the bias-variance trade-off risk and the oracle risk. It was the reason to keep the main term as it was stated. Fourth, the right hand side term of Ineq. (3.18) is of the order $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$.

Notice that the same rate for this term was achieved in [13], but in terms of the expectation over the noise.

A natural question would be to understand if the rate $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$ is sufficiently fast. In order to do that, one should precise the function space $\mathcal{F}$, where f^* lies in. In what follows, we will mention one famous example (among others) of a such function space $\mathcal{F}$.

Example 7. *Consider the class of functions*

$$\mathcal{F}_{\text{Lip}}(L) \coloneqq \left\{ f : [0,1]^d \mapsto \mathbb{R} \mid f(0) = 0,\ f \text{ is } L - \text{Lipschitz} \right\}, \tag{3.22}$$

where f is L-Lipschitz means that $|f(x) - f(x')| \leq L\|x - x'\|$ for all $x, x' \in [0,1]^d$. In this case (see, e.g., [67, Theorem 3.2] with $p = 1$),

$$\mathbb{E}_\varepsilon \|\widehat{f} - f^*\|_n^2 \geq c_l n^{-\frac{2}{2+d}}, \tag{3.23}$$

for some positive constant c_l, for any measurable of the input data $\widehat{f}$.

Therefore, for the class of L-Lipshitz functions, the rate $\mathcal{O}\left(\sqrt{\log n / n}\right)$ is faster than the minimax rate $\mathcal{O}(n^{-\frac{2}{2+d}})$ for any $d > 2$.

As for the main term $8V(k^*)$ in Ineq. (3.17), it should be of a minimax optimal order since the common strategy for obtaining optimal rates for the k-NN regression estimator is twofold. First, one should derive a uniform (over k) upper bound on the bias term (knowing the smoothness of the regression function), which is a non-decreasing function of k. After that, this upper bound is made equal to the variance term, which results in the optimal $k^{\text{b/v}}$. Following this argument, one can conclude that $k^{\text{b/v}} \leq k^*$, which implies $V(k^*) \leq V(k^{\text{b/v}})$. We summarize our findings in the theorem and the corollary below.

Theorem 3.4.2. *(e.g., [67, Theorem 3.2]) Under the Lipschitz condition (3.22) on the regression function f^*, for any $k \in \{1, \ldots, n\}$,*

$$\text{MSE}(k) \leq C \left(\frac{k}{n}\right)^{2/d} + \frac{\sigma^2}{k}, \tag{3.24}$$

where constant C may depend on d and L. Thus, Ineq. (3.24) yields $k^{\text{b/v}} = \left\lceil \left(\frac{\sigma^2}{C}\right)^{d/(2+d)} n^{\frac{2}{2+d}} \right\rceil$.

Corollary 3.4.3. *Set $u = \log n$ in Ineq. (3.17), then under the L-Lipschitz condition (3.22) on the regression function f^*, the early stopping rule k^τ from Eq. (3.15) satisfies*

$$\mathbb{E}_\varepsilon \|f^{k^\tau} - f^*\|_n^2 \leq c_u n^{-\frac{2}{2+d}}, \tag{3.25}$$

where positive constant c_u depends on d, σ, and L; $d > 2$.

Proof of Corollary 3.4.3. First, by taking the expectation of Ineq. (3.17), it gives

$$\begin{aligned}\mathbb{E}_\varepsilon\|f^{k^\tau} - f^*\|_n^2 = \mathbb{E}_\varepsilon\left[\|f^{k^\tau} - f^*\|_n^2 \mathbb{I}\left\{\|f^{k^\tau} - f^*\|_n^2 \leq 8V(k^*) + C_1\frac{\sqrt{\log n}}{\sqrt{n}} + C_2\frac{\log n}{n}\right\}\right] \\ + \mathbb{E}_\varepsilon\left[\|f^{k^\tau} - f^*\|_n^2 \mathbb{I}\left\{\|f^{k^\tau} - f^*\|_n^2 > 8V(k^*) + C_1\frac{\sqrt{\log n}}{\sqrt{n}} + C_2\frac{\log n}{n}\right\}\right].\end{aligned} \tag{3.26}$$

After that, due to Lemma 3.6.4 from Appendix, $\|I_n - A_k\|_2 \leq c$ for any $k \in \{1, \ldots, n\}$, and $|f^*(x_i)| \leq \mathcal{M}$ for $i \in \{1, \ldots, n\}$ due to the Lipschitz condition (3.22), which implies that

$$\begin{aligned}\|f^{k^\tau} - f^*\|_n^2 &= \|(I_n - A_k)F^*\|_n^2 + \|A_k\varepsilon\|_n^2 + 2\langle A_k\varepsilon, (I_n - A_k)F^*\rangle_n \\ &\leq 2\|(I_n - A_{k^\tau})F^*\|_n^2 + 2\|A_{k^\tau}\varepsilon\|_n^2 \\ &\leq 2\|I_n - A_{k^\tau}\|_2^2\|f^*\|_n^2 + 2\|A_{k^\tau}\|_2^2\|\varepsilon\|_n^2 \\ &\leq c_1 + c_2\|\varepsilon\|_n^2,\end{aligned}$$

where constants c_1 and c_2 depend only on $\mathcal{M}$ and d. Thus,

$$\|f^{k^\tau} - f^*\|_n^4 \leq c_1 + c_2\|\varepsilon\|_n^4 + c_3\|\varepsilon\|_n^2. \tag{3.27}$$

From Ineq. (3.26) and Cauchy-Schwarz inequality, it comes

$$\begin{aligned}\mathbb{E}_\varepsilon\|f^{k^\tau} - f^*\|_n^2 \leq 8V(k^*) + C_1\frac{\sqrt{\log n}}{\sqrt{n}} + C_2\frac{\log n}{n} \\ + \sqrt{\mathbb{E}_\varepsilon\|f^{k^\tau} - f^*\|_n^4}\sqrt{\mathbb{P}_\varepsilon\left(\|f^{k^\tau} - f^*\|_n^2 > 8V(k^*) + C_1\frac{\sqrt{\log n}}{\sqrt{n}} + C_2\frac{\log n}{n}\right)}.\end{aligned}$$

Further, by applying Ineq. (3.17) and Ineq. (3.27), we obtain

$$\mathbb{E}_\varepsilon\|f^{k^\tau} - f^*\|_n^2 \leq 8V(k^*) + C_1\frac{\sqrt{\log n}}{\sqrt{n}} + C_2\frac{\log n}{n} + \sqrt{c_1 + c_3\sigma^4 + c_2\sigma^2}\frac{4}{\sqrt{n}}.$$

The claim follows from $V(k^*) \leq V(k^{\mathrm{b/v}})$, for $k^{\mathrm{b/v}}$ defined in Theorem 3.4.2. ■

Therefore, the function estimator f^{k^τ} achieves (up to a constant) the minimax bound presented in Eq. (3.23), thus non-improvable in general.

3.5 Conclusion

In the present chapter, we tackled the problem of choosing the tuning parameter k in the k-NN regression estimator. A strategy based on early stopping and the minimum discrepancy principle was

proposed. In Section 3.4 it was shown that the minimum discrepancy stopping rule k^τ (3.15) provides a minimax optimal estimator, in particular, over the class of Lipschitz functions on a bounded domain. We remark that in the next chapter, this theoretical result will be confirmed empirically on artificial and real data sets: the stopping rule has comparable performance to other stopping rules that use, for instance, hold-out while reducing the computational time. The main inconvenience of the proposed strategy is that one has to estimate the variance σ^2 of the regression model (as it is the case for AIC or Mallow's C_p criteria), thus a plug-in estimator is needed. We will construct such an estimator for simulated experiments with real-world data in the next chapter.

Appendix

Below, one can find a plan of Appendix.

In Appendix 3.6, we state some already known results that will be used along the other sections of Appendix.

Appendix 3.7 is devoted to the introduction of the main quantities for the derivation of the proofs.

The main goal of Appendix 3.8 is to provide a concentration inequality for the difference of the variance $V(k^\tau)$ and its stochastic part $\|A_{k^\tau}\varepsilon\|_n^2$ as well as a concentration inequality for $\sup_{k\in\{1,\dots,n\}} |\, R_k - \mathbb{E}_\varepsilon R_k \,|$.

In Appendix 3.9, we derived a concentration inequality for controlling the variance term.

Appendix 3.10 is devoted to the derivation of a concentration inequality that deals with the deviation of the bias term.

After that, by combining all the results from Appendices 3.8, 3.9, and 3.10, we are able to provide a proof of Theorem 3.4.1.

3.6 Auxiliary lemmas

The first result is concerned with the derivation of the concentration of a Gaussian linear form around 0.

Lemma 3.6.1 (Concentration of a linear term)**.** *Let ε be a standard Gaussian vector in $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, and $Z \coloneqq \langle \varepsilon, \alpha\rangle = \sum_{j=1}^n \alpha_j\varepsilon_j$. Then for every $x > 0$, one has*

$$\mathbb{P}_\varepsilon\left(|Z| \geq x\right) \leq 2\exp\left[-\frac{x^2}{2\sigma^2\|\alpha\|^2}\right].$$

Further, we need to recall a concentration result for a quadratic form of Gaussian random variables.

Lemma 3.6.2 (Hanson-Wright's inequality for Gaussian random variables [95])**.** *If $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^\top \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2 I_n)$ and A is a $n\times n$ matrix, then for any $t > 0$,*

$$\mathbb{P}_\varepsilon\left(|\varepsilon^\top A\varepsilon - \mathbb{E}_\varepsilon[\varepsilon^\top A\varepsilon]| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{\sigma^4\|A\|_F^2}, \frac{t}{\sigma^2\|A\|_2}\right)\right]. \tag{3.28}$$

The next lemma provides us with a result that shows that the number of points among $\{x_1,\dots,x_n\}$ such that x_i is one of their k nearest neighbors, is not more than a constant times k.

Lemma 3.6.3 (Corollary 6.1 in [67]). *Assume that* $(X_1, \dots, X_n) \sim \mathbb{P}_X$ *for some probability measure* $\mathbb{P}_X$ *on* $\mathcal{X}$*, and* X *is an independent copy of* X_i*,* $i = 1, \dots, n$*, then if there are no ties, a.s.*

$$\sum_{i=1}^{n} \mathbb{I}\{X \text{ is among the kNNs of } X_i \text{ in the set } \{X_1, \dots, X_{i-1}, X, X_{i+1}, \dots, X_n\}\} \leq k c_d,$$

where constant c_d *depends only on* d*.*

After that, the operator norm of the matrix $I_n - A_k$ is proved to be bounded.

Lemma 3.6.4. *Recall that* $\mathcal{N}_k(i)$ *denotes the set of the* k *nearest neighbors of* x_i*. For any* $k \in \{1, \dots, n\}$*, define the matrix* $M_k \in \mathbb{R}^{n \times n}$ *as*

$$(M_k)_{ij} = \begin{cases} 1 - 1/k, \text{ if } i = j, \\ 0, \text{ if } j \notin \mathcal{N}_k(i), \\ -1/k, \text{ if } j \in \mathcal{N}_k(i). \end{cases}$$

Then, $\|M_k\|_2 \leq c_d$*, where positive constant* c_d *depends only on* d*. Moreover, it implies that for the matrix* $A_k = I_n - M_k$ *:* $\|A_k\|_2 \leq 1 + c_d$*.*

Proof of Lemma 3.6.4. We will adapt the proof of [13, Lemma 3.3].

Take $x \in \mathcal{X}$ such that $\|x\| = 1$ and denote $(M_k)_{i\cdot}$ as the i^{th} row of the matrix M_k. Then, the following holds.

$$\begin{aligned} \|M_k x\|^2 &= \sum_{i=1}^{n} \langle (M_k)_{i\cdot}, x \rangle^2 \\ &\leq 2 \sum_{i=1}^{n} (1 - 1/k)^2 x_i^2 + 2 \sum_{i=1}^{n} \left(\frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} x_j \right)^2 \\ &\overset{(i)}{\leq} 2\|x\|^2 + \frac{2}{k} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_k(i)} x_j^2 \\ &= 2\|x\|^2 + \frac{2}{k} \sum_{j=1}^{n} \sum_{i : j \in \mathcal{N}_k(i)} x_j^2 \\ &\overset{(ii)}{\leq} c_d \|x\|^2. \end{aligned}$$

(i) holds due to Jensen's inequality, and (ii) is due to Lemma 3.6.3. Hence, $\|M_k\|_2 \leq c_d$. ∎

Lemma 3.6.5. *For any* $k \in \{2, \dots, n\}$*,*

$$\frac{1}{2} V(k-1) \leq V(k) \leq V(k-1).$$

Proof of Lemma 3.6.5. It is sufficient to notice that

$$V(k-1) - V(k) = \frac{\sigma^2}{k(k-1)} \leq \frac{\sigma^2}{k} = V(k).$$

■

3.7 Main quantities and notations

For more theoretical convenience (the variance will be an increasing function, the empirical risk will be approximately a decreasing function), define the following notation and stopping rules:

$$\lambda[k] \coloneqq \mathrm{tr}(A_k) = n/k \ \in \ \{1, n/(n-1), n/(n-2), \ldots, n\}, \tag{3.29}$$

and

$$\begin{aligned}
\lambda_1^* &\coloneqq \inf\left\{\lambda \in \left\{1, \frac{n}{n-1}, \ldots, n\right\} \mid B^2(\lambda) \leq V(\lambda)\right\}, \quad \lambda_1^\tau \coloneqq \inf\left\{\lambda \in \left\{1, \frac{n}{n-1}, \ldots, n\right\} \mid R_\lambda \leq \sigma^2\right\} \\
\lambda_2^* &\coloneqq \sup\left\{\lambda \in \left\{1, \frac{n}{n-1}, \ldots, n\right\} \mid B^2(\lambda) \geq V(\lambda)\right\}, \quad \lambda_2^\tau \coloneqq \sup\left\{\lambda \in \left\{1, \frac{n}{n-1}, \ldots, n\right\} \mid R_\lambda \geq \sigma^2\right\}.
\end{aligned} \tag{3.30}$$

Notice that there is a one-to-one map between k and $\lambda[k]$, as it is suggested in Eq. (3.29).

If λ_1^* does not exist, set $\lambda_1^* = n$ whereas, if λ_2^* does not exists, set $\lambda_2^* = 1$. If λ_1^τ does not exist, set $\lambda_1^\tau = n$; if λ_2^τ does not exist, set $\lambda_2^\tau = 1$.

In Eq. (3.30), we omit for simplicity the notation $\lambda[k]$. Moreover, in Eq. (3.30), we used the notation $A_{\lambda[k]}$ (inside the definitions of $B^2(\lambda), V(\lambda)$, and R_λ) to denote the matrix A_k for $k = n/\lambda$ corresponding to λ, i.e., $A_{\lambda[k]} \equiv A_k$.

Note that $\lambda_1^* \leq \lambda_2^*$, and $\lambda_1^\tau \leq \lambda_2^\tau$. Besides that, the bias, variance, and (expected) empirical risk at λ_1^τ are equal to the bias, variance, (expected) empirical risk at k^τ, defined in Eq. (3.15), respectively. The bias, variance, (expected) empirical risk at λ_2^* are equal to the bias, variance, (expected) empirical risk at k^*, defined in Eq. (3.14), respectively.

The behavior of the bias term, variance, risk error, and (expected) empirical risk w.r.t. the new notation λ is presented in Fig. 3.2. One can conclude that only the variance term is monotonic w.r.t. λ.

Denote $\widetilde{R}_\lambda$ as the tightest non-increasing lower bound on R_λ and $\overline{R}_\lambda$ as the tightest non-increasing upper bound on R_λ. We precise the definitions of the latter quantities below.

Definition 3.7.1. Assume that one has the grid of values $\Lambda = \{1, n/(n-1), n/(n-2), \ldots, n\}$, and the empirical risk curve is observed successively, meaning that one starts from $\lambda = 1$ (corresponds to

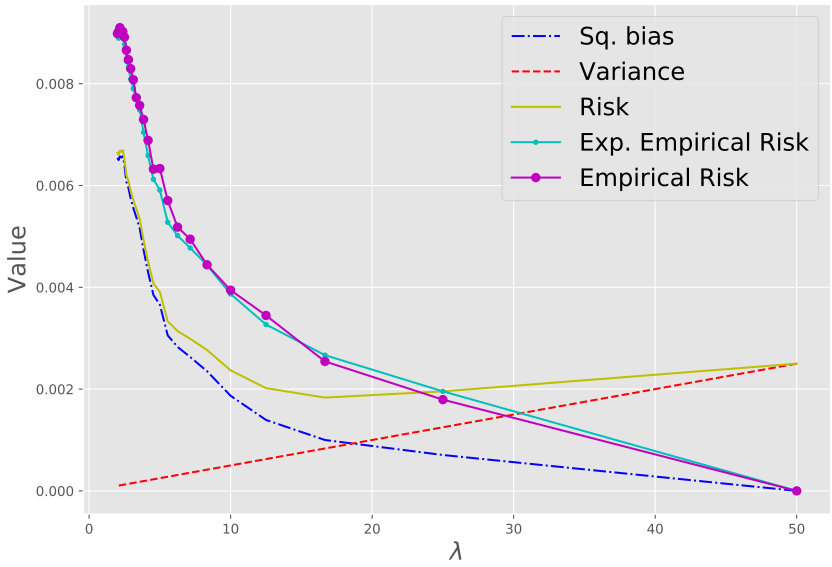


Figure 3.2 – Sq. bias, variance, risk, and (expected) empirical risk behavior in the λ notation.

$k = n$) and increases λ until the value n (corresponds to $k = 1$). Then, consider the value of R_λ and its next increment $R_{\lambda+\Delta}$ such that $\lambda + \Delta \in \Lambda$. Define $\widetilde{R}_1 \coloneqq R_1$ and

$$\widetilde{R}_{\lambda+\Delta} \coloneqq \begin{cases} R_{\lambda+\Delta}, \text{ if } R_{\lambda+\Delta} - R_\lambda \leq 0, \\ R_\lambda, \text{ otherwise; in this case, one should wait until } R_{\widetilde{\lambda}} \leq \widetilde{R}_{\widetilde{\lambda}} \text{ for some } \widetilde{\lambda} > \lambda, \ \widetilde{\lambda} \in \Lambda. \end{cases} \tag{3.31}$$

Definition 3.7.2. Assume that one has the grid of values $\Lambda = \{1, n/(n-1), n/(n-2), \ldots, n\}$, and the empirical risk curve is observed successively, meaning that one starts from $\lambda = n$ (corresponds to $k = 1$) and decreases λ until the value 1 (corresponds to $k = n$). Then, consider the value of R_λ and its next increment $R_{\lambda-\Delta}$ such that $\lambda - \Delta \in \Lambda$. Define $\overline{R}_n \coloneqq R_n$ and

$$\overline{R}_{\lambda-\Delta} \coloneqq \begin{cases} R_{\lambda-\Delta}, \text{ if } R_{\lambda-\Delta} - R_\lambda \geq 0, \\ R_\lambda, \text{ otherwise; in this case, one should wait until } R_{\widetilde{\lambda}} \geq \overline{R}_{\widetilde{\lambda}} \text{ for some } \widetilde{\lambda} < \lambda, \ \widetilde{\lambda} \in \Lambda. \end{cases} \tag{3.32}$$

The typical behavior of the defined lower and upper bound $\widetilde{R}_\lambda, \overline{R}_\lambda$ is illustrated in Fig. 3.3. Note

that with these definitions:

$$\lambda_1^\tau = \inf\{\lambda \in \{1, \dots, n\} \mid \widetilde{R}_\lambda \leq \sigma^2\},$$
$$\lambda_2^\tau = \sup\{\lambda \in \{1, \dots, n\} \mid \overline{R}_\lambda \geq \sigma^2\}.$$

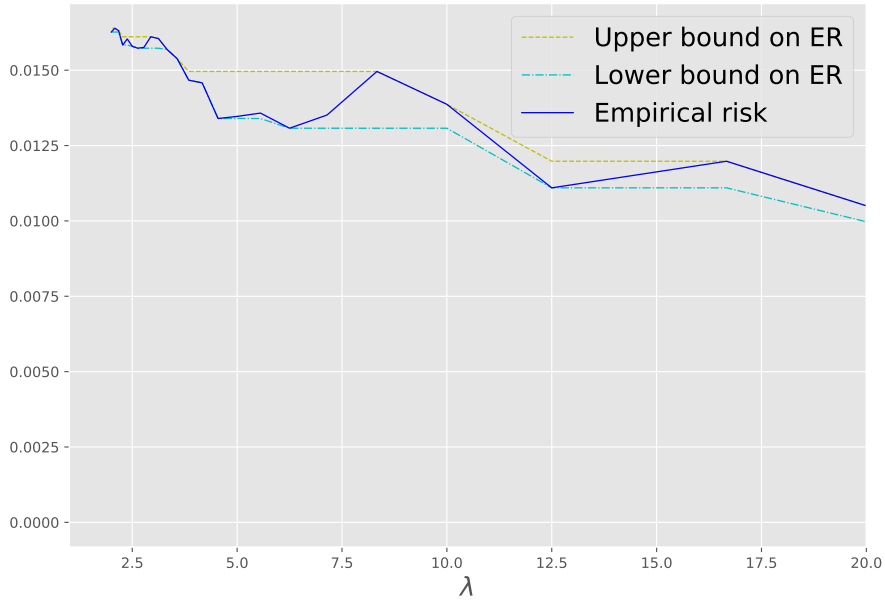


Figure 3.3 – Lower and upper bounds on the empirical risk.

Define an additional stopping rule $\lambda^{\star\star}$ that will be helpful in the analysis.

$$\lambda^{\star\star} := \sup\left\{\lambda \in \{1, \dots, n\} \mid B^2(\lambda) \geq V(\lambda) + c_1\sqrt{\frac{\log n}{n}} + \widetilde{y}\right\}, \tag{3.33}$$

for some $\widetilde{y} \geq 0$ and positive constant c_1 that will be precised later (Lemma 3.10.2).

3.8 Control of the stochastic part of the variance / the empirical risk

3.8.1 Control of the stochastic part of the variance

Consider $v(\lambda_1^\tau) = \|A_{\lambda_1^\tau[k]}\varepsilon\|_n^2$ and $V(\lambda_1^\tau) = \frac{\sigma^2}{n}\mathrm{tr}\left(A_{\lambda_1^\tau[k]}\right)$ from Section 3.7. Then for any $x > 0$,

$$\begin{aligned}\mathbb{P}_\varepsilon\left(v(\lambda_1^\tau) > V(\lambda_1^\tau) + x\right) &= \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda_1^\tau[k] < 1\} \bigcap \{v(\lambda_1^\tau) - V(\lambda_1^\tau) > x\}\right)}_{=0} \\ &+ \mathbb{P}_\varepsilon\left(\{\lambda_1^\tau[k] \geq 1\} \bigcap \{v(\lambda_1^\tau) - V(\lambda_1^\tau) > x\}\right) \\ &\leq \mathbb{P}_\varepsilon\left(\sup_{k\in\{1,\dots,n\}} \left|\|A_k\varepsilon\|_n^2 - V(k)\right| > x\right).\end{aligned} \tag{3.34}$$

In what follows, we will bound $\mathbb{P}_\varepsilon\left(\sup\limits_{k\in\{1,\dots,n\}} |\|A_k\varepsilon\|_n^2 - V(k)| > x\right)$.

Let us define the set of matrices $\overline{\mathcal{A}} \coloneqq \{A_k,\ k = 1,\dots,n\}$, then [74, Theorem 3.1]

$$\mathbb{P}_\varepsilon\left(\sup_{\mathbf{A}\in\overline{\mathcal{A}}} \left|\|\mathbf{A}\varepsilon\|^2 - \mathbb{E}_\varepsilon\|\mathbf{A}\varepsilon\|^2\right| \geq c_1 E + t\right) \leq 2\exp\left(-c_2 \min\left(\frac{t^2}{V^2}, \frac{t}{U}\right)\right), \tag{3.35}$$

where

$$\begin{aligned}E &= \gamma_2(\overline{\mathcal{A}}, \|\cdot\|_2)\left(\gamma_2(\overline{\mathcal{A}}, \|\cdot\|_2) + \sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_F\right) + \sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_F \sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_2, \\ U &= \left[\sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_2\right]^2, \\ V &= \sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_2\left(\gamma_2(\overline{\mathcal{A}}, \|\cdot\|_2) + \sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_F\right),\end{aligned}$$

and $\gamma_2(\overline{\mathcal{A}}, \|\cdot\|_2)$ can be bounded via the metric entropy of $(\overline{\mathcal{A}}, \|\cdot\|_2)$ as

$$\gamma_2(\overline{\mathcal{A}}, \|\cdot\|_2) \leq c\int_0^{\sup_{\mathbf{A}\in\overline{\mathcal{A}}}\|\mathbf{A}\|_2} \sqrt{\log N(\overline{\mathcal{A}};\ \|\cdot\|_2;\ u)}du.$$

First, notice that due to Lemma 3.6.4, for any $\mathbf{A} \in \overline{\mathcal{A}}$, one has $\|\mathbf{A}\|_2 \leq c_d$. Moreover, $\log N(\overline{\mathcal{A}};\ \|\cdot\|_2;\ u) \leq \log n$ due to the definition of the metric entropy (see, e.g., [114, Chapter 5]). These arguments imply

$$\begin{aligned}U &\leq c_d, \qquad \text{and} \\ \gamma_2(\overline{\mathcal{A}}, \|\cdot\|_2) &\leq c_{\gamma,d}\sqrt{\log n},\end{aligned}$$

where constants c_d and $c_{\gamma,d}$ depend only on d.

Second, as for the Frobenius norm,

$$\sup_{\mathbf{A}\in\overline{\mathcal{A}}} \|\mathbf{A}\|_F \leq \sqrt{n}$$

due to the definition (3.7). Combining all the pieces together, for any $t > 0$,

$$\mathbb{P}_\varepsilon \left(\sup_{\mathbf{A}\in\overline{\mathcal{A}}} \left| \|\mathbf{A}\varepsilon\|_n^2 - \mathbb{E}_\varepsilon \|\mathbf{A}\varepsilon\|_n^2 \right| \geq c_1 \sqrt{\frac{\log n}{n}} + t \right) \leq 2\exp\left(-c_2 \min\left(nt^2, nt \right) \right),$$

where c_1 and c_2 may depend on d and σ^2.

Take $x = c_1\sqrt{\frac{\log n}{n}} + t$ in (3.34), then for any $t > 0$,

$$\mathbb{P}_\varepsilon \left(v(\lambda_1^\tau) > V(\lambda_1^\tau) + c_1 \sqrt{\frac{\log n}{n}} + t \right) \leq 2\exp\left(-cn \min\left(t^2, t \right) \right).$$

3.8.2 Control of the empirical risk around its expectation

Define now the set of matrices $\overline{\mathcal{M}} \coloneqq \{M_k = I_n - A_k,\ k = 1, \ldots, n\}$, then by the same arguments presented above, for any $t > 0$,

$$\mathbb{P}_\varepsilon \left(\sup_{\mathbf{M}\in\overline{\mathcal{M}}} \mid \|\mathbf{M}\varepsilon\|_n^2 - \mathbb{E}_\varepsilon \|\mathbf{M}\varepsilon\|_n^2 \mid \geq c_1 \sqrt{\frac{\log n}{n}} + t \right) \leq 2\exp\left(-c_2 \min(nt^2, nt) \right) \tag{3.36}$$

with c_1 and c_2 depending only on d and σ^2. Further, notice that for any $k \in \{1, \ldots, n\}$,

$$R_k - \mathbb{E}_\varepsilon R_k = \|M_k Y\|_n^2 - \mathbb{E}_\varepsilon \|M_k Y\|_n^2 = \|M_k\varepsilon\|_n^2 - \sigma^2 \left(1 - \frac{1}{k} \right) + 2\langle F^*, M_k^\top M_k \varepsilon \rangle_n.$$

Ineq. (3.36) implies that for any $t > 0$,

$$\mathbb{P}_\varepsilon \left(\sup_{k\in\{1,\ldots,n\}} \mid \|M_k\varepsilon\|_n^2 - \sigma^2 \left(1 - \frac{1}{k} \right) \mid \geq c_1 \sqrt{\frac{\log n}{n}} + t \right) \leq 2\exp(-c_2 \min(nt^2, nt)). \tag{3.37}$$

Moreover, Lemma 3.6.1 gives us that for any $y > 0$ and $k \in \{1, \dots, n\}$,

$$\begin{aligned}
\mathbb{P}_\varepsilon \left(2 \mid \langle F^*, M_k^\top M_k \varepsilon \rangle_n \mid \geq y\right) &\leq 2 \exp\left[-\frac{n^2 y^2}{8\sigma^2 \|M_k^\top M_k F^*\|^2}\right] \\
&\leq 2 \exp\left[-\frac{n^2 y^2}{8\sigma^2 \|M_k^\top M_k\|_2^2 \|F^*\|^2}\right] \\
&\leq 2 \exp\left[-\frac{n y^2}{8 c_d \sigma^2 \|f^*\|_n^2}\right] \\
&\leq 2 \exp\left[-\frac{n y^2}{8 c_d \sigma^2 \mathcal{M}^2}\right].
\end{aligned} \tag{3.38}$$

Then, using the union bound for the linear term above with the deviation $y = c_1 \sqrt{\frac{\log n}{n}} + t$ and combining all the pieces together,

$$\mathbb{P}_\varepsilon \left(\sup_{k \in \{1,\dots,n\}} \mid R_k - \mathbb{E}_\varepsilon R_k \mid \geq c_1 \sqrt{\frac{\log n}{n}} + t \right) \leq 4 \exp\left[-c_2 \min(nt^2, nt)\right] \tag{3.39}$$

for any $t > 0$.

3.9 Deviation inequality for the variance term

This is the first deviation inequality for λ_1^τ that will be used to control the variance term.

Lemma 3.9.1. *Under Assumption 6, define $\mathcal{K}_V \subseteq \{1, \dots, n\}$ such that, for any $\lambda \in \mathcal{K}_V$, one has $V(\lambda) \geq V(\lambda[k^* - 1]) + y$ for some $y \geq 0$. Recall the definition of λ_1^τ from Eq. (3.30), then for any $\lambda \in \mathcal{K}_V$,*

$$\mathbb{P}_\varepsilon \left(\lambda_1^\tau > \lambda\right) \leq 2 \exp\left[-c_d n \min\left(\frac{y^2}{\sigma^4}, \frac{y}{\sigma^2}\right)\right], \tag{3.40}$$

where constant c_d depends only on d.

Proof of Lemma 3.9.1. We start with the following series of inequalities that can be derived from the definition of λ_1^τ and lower bound on the empirical risk $\widetilde{R}_\lambda$ from Eq. (3.31).

$$\begin{aligned}
\mathbb{P}_\varepsilon \left(\lambda_1^\tau > \lambda\right) &= \mathbb{P}_\varepsilon \left(\widetilde{R}_\lambda > \sigma^2\right) \\
&= \mathbb{P}_\varepsilon \left(\widetilde{R}_\lambda - \mathbb{E}_\varepsilon R_\lambda > \sigma^2 - \mathbb{E}_\varepsilon R_\lambda\right) \\
&\leq \mathbb{P}_\varepsilon \left(R_\lambda - \mathbb{E}_\varepsilon R_\lambda > \sigma^2 - \mathbb{E}_\varepsilon R_\lambda\right).
\end{aligned}$$

Due to Eq. (3.12), one has

$$\sigma^2 - \mathbb{E}_\varepsilon R_\lambda = V(\lambda) - B^2(\lambda) \geq V(\lambda) - V(\lambda[k^* - 1]) \geq y.$$

Moreover,

$$R_\lambda - \mathbb{E}_\varepsilon R_\lambda = \|(I_n - A_{\lambda[k]})\varepsilon\|_n^2 - \frac{\sigma^2}{n}\left(n - \mathrm{tr}(A_{\lambda[k]})\right) + 2\langle (I_n - A_{\lambda[k]})F^*, (I_n - A_{\lambda[k]})\varepsilon\rangle_n.$$

Define for simplicity $M_{\lambda[k]} \coloneqq I_n - A_{\lambda[k]}$, then

$$\mathbb{P}_\varepsilon\left(\lambda_1^\tau > \lambda\right) \le \mathbb{P}_\varepsilon\left(\|M_{\lambda[k]}\varepsilon\|_n^2 - \frac{\sigma^2}{n}\left(n - \mathrm{tr}(A_{\lambda[k]})\right) \ge \frac{y}{2}\right) + \mathbb{P}_\varepsilon\left(2\langle M_{\lambda[k]}F^*, M_{\lambda[k]}\varepsilon\rangle_n \ge \frac{y}{2}\right).$$

Further, we will concentrate the quadratic and linear terms as follows.

First term. The linear term $2\langle M_{\lambda[k]}F^*, M_{\lambda[k]}\varepsilon\rangle_n$: using Lemma 3.6.1 and Lemma 3.6.4 gives us

$$\begin{aligned}
\mathbb{P}_\varepsilon\left(2\langle M_{\lambda[k]}F^*, M_{\lambda[k]}\varepsilon\rangle_n \ge \frac{y}{2}\right) &= \mathbb{P}_\varepsilon\left(\langle M_{\lambda[k]}^\top M_{\lambda[k]}F^*, \varepsilon\rangle \ge \frac{ny}{4}\right) \\
&\le \exp\left[-\frac{n^2y^2}{32\sigma^2\|M_{\lambda[k]}^\top M_{\lambda[k]}F^*\|^2}\right] \\
&\le \exp\left[-\frac{ny^2}{32\sigma^2\|M_{\lambda[k]}^\top\|_2^2 B^2(\lambda)}\right] \\
&\le \exp\left[-\frac{ny^2}{32c_d\sigma^2 V(\lambda)}\right] \\
&\le \exp\left[-\frac{ny^2}{32c_d\sigma^4}\right].
\end{aligned}$$

Second term. Consider the quadratic term $\|M_{\lambda[k]}\varepsilon\|_n^2 - \frac{\sigma^2}{n}\left(n - \mathrm{tr}A_{\lambda[k]}\right)$: combining Lemma 3.6.2 and Lemma 3.6.4 gives

$$\begin{aligned}
\mathbb{P}_\varepsilon\left(\|M_{\lambda[k]}\varepsilon\|_n^2 - \frac{\sigma^2}{n}\left(n - \mathrm{tr}A_{\lambda[k]}\right) \ge \frac{y}{2}\right) &\le \exp\left[-c\min\left(\frac{n^2y^2}{4\sigma^4\|M_{\lambda[k]}^\top M_{\lambda[k]}\|_F^2}, \frac{ny}{2\sigma^2\|M_{\lambda[k]}^\top M_{\lambda[k]}\|_2}\right)\right] \\
&\le \exp\left[-c_d\min\left(\frac{ny^2}{4\sigma^4}, \frac{ny}{2\sigma^2}\right)\right],
\end{aligned}$$

where constant c_d depends only on d.

■

Based on Lemma 3.9.1, due to the fact that the variance $V(\lambda)$ is increasing w.r.t. $\lambda \in \{1, n/(n-1), \ldots, n\}$, the following corollary holds.

Corollary 3.9.2. *For any $y > 0$, define $0 \le \Delta y \le y$ as the distance between $V(\lambda[k^* - 1]) + y$ and $V(\lambda_0)$, where $V(\lambda_0)$ is the closest to $V(\lambda[k^* - 1]) + y$ value of $V(\lambda)$, which is lower than or equal to $V(\lambda[k^* - 1]) + y$, over the grid of $\lambda \in \{\lambda[k^* - 1], \lambda[k^* - 2], \ldots, n\}$. Then due to the monotonicity of*

the variance term,

$$\mathbb{P}_\varepsilon\left(V(\lambda_1^\tau) > V(\lambda[k^*-1]) + y - \Delta y\right) \leq 2\exp\left[-c_d n \min\left(\frac{y-\Delta y}{\sigma^2}, \frac{(y-\Delta y)^2}{\sigma^4}\right)\right], \tag{3.41}$$

for constant c_d that depends only on d. Moreover, due to the definition of k^* (3.14) and Lemma 3.6.5, $\frac{1}{2}V(\lambda[k^*-1]) \leq V(\lambda_2^*) \leq V(\lambda[k^*-1])$, which implies that

$$\mathbb{P}_\varepsilon\left(V(\lambda_1^\tau) > 2V(\lambda_2^*) + y - \Delta y\right) \leq 2\exp\left[-c_d n \min\left(\frac{y-\Delta y}{\sigma^2}, \frac{(y-\Delta y)^2}{\sigma^4}\right)\right], \ \forall y > 0.$$

Thus, one is able to control $V(\lambda_1^\tau)$ via $V(\lambda_2^*)$, which is equal to $V(k^*)$.

3.10 Deviation inequality for the bias term

What follows is the second deviation inequality for λ_1^τ that will be further used to control the bias term.

Lemma 3.10.1. *Under Assumption 6, define $\mathcal{K}_B \subseteq \{1,\dots,n\}$ such that for any $\lambda \in \mathcal{K}_B$, one has $B^2(\lambda) \geq V(\lambda) + c_1\sqrt{\frac{\log n}{n}}$ for some positive constant c_1. Then if $\mathcal{K}_B$ is not empty, λ_1^τ from Eq. (3.30) satisfies*

$$\mathbb{P}_\varepsilon\left(\lambda_1^\tau < \lambda\right) \leq 10\exp\left(-cn\min\left(y^2, y\right)\right), \tag{3.42}$$

where $y = B^2(\lambda) - V(\lambda) - c_1\sqrt{\frac{\log n}{n}}$ for any $\lambda \in \mathcal{K}_B$, constant c depends only on d, σ, and $\mathcal{M}$.

Proof of Lemma 3.10.1. Consider Ineq. (3.39) and the event

$$\mathcal{E}_{\text{er}}(t) \coloneqq \left\{\sup_{\lambda\in\{1,\dots,n\}} \mid R_\lambda - \mathbb{E}_\varepsilon R_\lambda \mid \geq c_1\sqrt{\frac{\log n}{n}} + t\right\}$$

for any $t > 0$. Take $t \coloneqq B^2(\lambda) - V(\lambda),\ \lambda \in \mathcal{K}_B$. One notes from Ineq. (3.39) that

$$\mathbb{P}_\varepsilon\left(\mathcal{E}_{\text{er}}\left(B^2(\lambda) - V(\lambda)\right)\right) \leq 4\exp\left(-cn\min\left(\left[B^2(\lambda) - V(\lambda)\right]^2, B^2(\lambda) - V(\lambda)\right)\right). \tag{3.43}$$

Further, recall that $\lambda_1^\tau \leq \lambda_2^\tau$, and $\overline{R}_\lambda$ is the upper bound on R_λ from Section 3.7, which implies that

$$\begin{aligned}
\mathbb{P}_\varepsilon\left(\lambda_1^\tau < \lambda\right) &= \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda_1^\tau < \lambda\} \bigcap \{\lambda > \lambda_2^\tau\}\right)}_{\mathcal{A}} + \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda_1^\tau < \lambda\} \bigcap \{\lambda \leq \lambda_2^\tau\}\right)}_{\mathcal{B}}, \\
\mathcal{A} &= \mathbb{P}_\varepsilon\left(\overline{R}_\lambda < \sigma^2\right) \leq \mathbb{P}_\varepsilon\left(R_\lambda < \sigma^2\right) \leq \mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right), \\
\mathcal{B} &= \mathbb{P}_\varepsilon\left(\lambda \in (\lambda_1^\tau, \lambda_2^\tau]\right).
\end{aligned} \tag{3.44}$$

Consider the probability $\mathcal{B}$ from (3.44).

$$\mathcal{B} = \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda \in (\lambda_1^\tau, \lambda_2^\tau]\} \bigcap \left\{R_\lambda > \sigma^2\right\}\right)}_{\mathcal{C}} + \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda \in (\lambda_1^\tau, \lambda_2^\tau]\} \bigcap \left\{R_\lambda \leq \sigma^2\right\}\right)}_{\mathcal{D}}.$$

On the one hand,

$$\mathcal{D} \leq \mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2\right) \leq \mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right).$$

On the other hand, Ineq. (3.43) and the equality $\mathbb{E}_\varepsilon R_\lambda = \sigma^2 + B^2(\lambda) - V(\lambda)$ imply that the event

$$R_\lambda \in \left(\sigma^2 - c_1\sqrt{\frac{\log n}{n}}, \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right) \qquad \text{for any } \lambda \in \{1, \ldots, n\}$$

holds with probability at least $1 - 4\exp\left(-cn\min\left(\left[B^2(\lambda) - V(\lambda)\right]^2, B^2(\lambda) - V(\lambda)\right)\right)$. Let us denote this event as $\overline{\mathcal{E}}$. Then,

$$\mathcal{C} = \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda \in (\lambda_1^\tau, \lambda_2^\tau]\} \bigcap \left\{R_\lambda > \sigma^2\right\} \bigcap \left\{\overline{\mathcal{E}}\right\}\right)}_{\widetilde{\mathcal{F}}} + \underbrace{\mathbb{P}_\varepsilon\left(\{\lambda \in (\lambda_1^\tau, \lambda_2^\tau]\} \bigcap \left\{R_\lambda > \sigma^2\right\} \bigcap \left\{\overline{\mathcal{E}}^c\right\}\right)}_{\mathcal{G}}.$$

First,

$$\mathcal{G} \leq \mathbb{P}_\varepsilon\left(\overline{\mathcal{E}}^c\right) \leq 4\exp\left(-cn\min\left([B^2(\lambda) - V(\lambda)]^2, B^2(\lambda) - V(\lambda)\right)\right).$$

Second,

$$\widetilde{\mathcal{F}} \leq \mathbb{P}_\varepsilon\left(R_\lambda \in \left(\sigma^2,\ \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right)\right) \leq \mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right).$$

Combining the terms $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \widetilde{\mathcal{F}}$, and $\mathcal{G}$, one gets

$$\begin{aligned}\mathbb{P}_\varepsilon\left(\lambda_1^\tau < \lambda\right) &\leq 3\ \mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right) + 4\exp\left(-cn\min\left(\left[B^2(\lambda) - V(\lambda)\right]^2, B^2(\lambda) - V(\lambda)\right)\right)\\ &\leq 3\ \mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right)\\ &\quad + 4\exp\left(-cn\min\left(\left[B^2(\lambda) - V(\lambda) - c_1\sqrt{\frac{\log n}{n}}\right]^2, B^2(\lambda) - V(\lambda) - c_1\sqrt{\frac{\log n}{n}}\right)\right).\end{aligned}$$

Then, one has

$$\begin{aligned}\mathbb{P}_\varepsilon\left(R_\lambda \leq \sigma^2 + c_1\sqrt{\frac{\log n}{n}}\right) &= \mathbb{P}_\varepsilon\left(R_\lambda - \mathbb{E}_\varepsilon R_\lambda \leq \sigma^2 - \mathbb{E}_\varepsilon R_\lambda + c_1\sqrt{\frac{\log n}{n}}\right)\\ &= \mathbb{P}_\varepsilon\left(R_\lambda - \mathbb{E}_\varepsilon R_\lambda \leq -\left(\mathbb{E}_\varepsilon R_\lambda - \sigma^2 - c_1\sqrt{\frac{\log n}{n}}\right)\right).\end{aligned} \tag{3.45}$$

Since $\mathbb{E}_\varepsilon R_\lambda - \sigma^2 - c_1\sqrt{\frac{\log n}{n}} = B^2(\lambda) - V(\lambda) - c_1\sqrt{\frac{\log n}{n}} =: y$ for any $\lambda \in \mathcal{K}_B$, and

$$R_\lambda - \mathbb{E}_\varepsilon R_\lambda = \|(I_n - A_{\lambda[k]})\varepsilon\|_n^2 - \frac{\sigma^2}{n}(n - \operatorname{tr}(A_{\lambda[k]})) + 2\langle (I_n - A_{\lambda[k]})F^*, (I_n - A_{\lambda[k]})\varepsilon\rangle_n,$$

we have

$$\mathbb{P}_\varepsilon\left(\lambda_2^\tau < \lambda\right) \leq \mathbb{P}_\varepsilon\left(\|M_{\lambda[k]}\varepsilon\|_n^2 - \frac{\sigma^2}{n}(n - \operatorname{tr}(A_{\lambda[k]})) \leq -\frac{y}{2}\right) + \mathbb{P}_\varepsilon\left(2\langle M_{\lambda[k]}F^*, M_{\lambda[k]}\varepsilon\rangle_n \leq -\frac{y}{2}\right),$$

where the matrix $M_{\lambda[k]} = I_n - A_{\lambda[k]}$, where $A_{\lambda[k]} \equiv A_k$.

Further, we will concentrate the quadratic and linear terms above as follows.

First term. The linear term $2\langle M_{\lambda[k]}F^*, M_{\lambda[k]}\varepsilon\rangle_n$: using Lemma 3.6.1 and Lemma 3.6.4 gives us

$$\begin{aligned}\mathbb{P}_\varepsilon\left(2\langle M_{\lambda[k]}F^*, M_{\lambda[k]}\varepsilon\rangle_n \leq -\frac{y}{2}\right) &= \mathbb{P}_\varepsilon\left(\langle M_{\lambda[k]}^\top M_{\lambda[k]}F^*, \varepsilon\rangle \leq -\frac{ny}{4}\right)\\ &\leq \exp\left[-\frac{n^2y^2}{32\sigma^2\|M_{\lambda[k]}^\top M_{\lambda[k]}F^*\|^2}\right]\\ &\leq \exp\left[-\frac{n^2y^2}{32\sigma^2\|M_{\lambda[k]}^\top M_{\lambda[k]}\|_2^2\|F^*\|^2}\right]\\ &\leq \exp\left[-\frac{ny^2}{32c_d\sigma^2\|f^*\|_n^2}\right].\end{aligned}$$

Second term. Consider the quadratic term $\|M_{\lambda[k]}\varepsilon\|_n^2 - \frac{\sigma^2}{n}(n - \mathrm{tr}A_{\lambda[k]})$: combining Lemma 3.6.2 and Lemma 3.6.4 gives

$$\mathbb{P}_\varepsilon\left(\|M_{\lambda[k]}\varepsilon\|_n^2 - \frac{\sigma^2}{n}(n-\mathrm{tr}A_{\lambda[k]}) \le -\frac{y}{2}\right) \le \exp\left[-c\min\left(\frac{n^2y^2}{4\sigma^4\|M_{\lambda[k]}^\top M_{\lambda[k]}\|_F^2}, \frac{ny}{2\sigma^2\|M_{\lambda[k]}^\top M_{\lambda[k]}\|_2}\right)\right]$$
$$\le \exp\left[-c_d\min\left(\frac{ny^2}{4\sigma^4}, \frac{ny}{2\sigma^2}\right)\right],$$

where constant c_d depends only on d.

Finally, it is sufficient to recall Assumption 6 in order to apply $\|f^*\|_n^2 \le \mathcal{M}^2$.

■

Lemma 3.10.2. *Under Assumption 6, recall the definitions of λ_1^τ and λ_2^* from Eq. (3.30). Then for any $y > 0$ and Δy from Corollary 3.9.2,*

$$B^2(\lambda_1^\tau) \le 2V(\lambda_2^*) + c_1\sqrt{\frac{\log n}{n}} + 2(y - \Delta y) \tag{3.46}$$

with probability at least $1 - 12\exp\left(-cn\min\left((y-\Delta y)^2, y-\Delta y\right)\right)$, where constants c, c_1 depend only on d, σ, and $\mathcal{M}$.

Proof of Lemma 3.10.2. Consider the event $\mathcal{E}(\lambda)$ from Lemma 3.10.1 for each $\lambda \in \mathcal{K}_B$. Then,

$$\mathbb{P}_\varepsilon\left(\mathcal{E}(\lambda)\right) \le 10\exp\left(-cn\min\left(x^2, x\right)\right),$$

for $x = B^2(\lambda) - V(\lambda) - c_1\sqrt{\frac{\log n}{n}}$.

In what follows, two cases are distinguished.

Case 1: If $\lambda_1^\tau > \lambda_2^*$, then by definition of λ_2^*, Corollary 3.9.2, and the monotonicity of the variance term,

$$B^2(\lambda_1^\tau) < V(\lambda_1^\tau) \le 2V(\lambda_2^*) + y - \Delta y \tag{3.47}$$

with probability at least $1 - 2\exp\left(-c_d n\min\left(\frac{y-\Delta y}{\sigma^2}, \frac{(y-\Delta y)^2}{\sigma^4}\right)\right),\ \forall y > 0$.

Case 2: If $\lambda_1^\tau \le \lambda_2^*$, then take $y - \Delta y$ from Ineq. (3.47) and define $\lambda^{\star\star} \le \lambda_2^*$ as in Eq. (3.33) with $\widetilde{y} = y - \Delta y$.

If no such point $\lambda^{\star\star}$ exists, then for any $\lambda \le \lambda_2^*$, one has $B^2(\lambda) < V(\lambda) + c_1\sqrt{\frac{\log n}{n}} + y - \Delta y$. In particular, it holds true for λ_1^τ, which implies that

$$B^2(\lambda_1^\tau) < V(\lambda_1^\tau) + c_1\sqrt{\frac{\log n}{n}} + y - \Delta y \le 2V(\lambda_2^*) + c_1\sqrt{\frac{\log n}{n}} + 2(y - \Delta y)$$

with probability at least $1 - 2\exp\left(-c_d n\min\left(\frac{y-\Delta y}{\sigma^2}, \frac{(y-\Delta y)^2}{\sigma^4}\right)\right)$, due to Corollary 3.9.2.

If $\lambda^{\star\star}$ exists, notice that $\lambda^{\star\star} \in \mathcal{K}_B$ by its definition. Therefore, due to Lemma 3.10.1, under the event $\mathcal{E}^c(\lambda^{\star\star})$, $\lambda_1^\tau \geq \lambda^{\star\star}$, and

$$B^2(\lambda_1^\tau) < V(\lambda_1^\tau) + c_1\sqrt{\frac{\log n}{n}} + y - \Delta y \leq 2V(\lambda_2^*) + c_1\sqrt{\frac{\log n}{n}} + 2(y - \Delta y)$$

with probability at least $1 - 10\exp\left(-cn\min\left((y-\Delta y)^2, y - \Delta y\right)\right)$.

Combining **Case 1** and **Case 2** together,

$$B^2(\lambda_1^\tau) \leq 2V(\lambda_2^*) + c_1\sqrt{\frac{\log n}{n}} + 2(y - \Delta y) \tag{3.48}$$

with probability at least $1 - 12\exp\left(-cn\min\left((y-\Delta y)^2, y - \Delta y\right)\right)$.

The claim is proved.

■

3.11 Proof of Theorem 3.4.1

Define $v(\lambda[k]) \coloneqq \|A_{\lambda[k]}\varepsilon\|_n^2$, where $\lambda[k] = \mathrm{tr}(A_k) = n/k$ (see Section 3.7 for the definitions related to the notation λ). Then, due to the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \geq 0$, Lemma 3.10.2, Corollary 3.9.2, and the control of the stochastic term in Appendix 3.8 (with $t = y - \Delta y$), for $\lambda_1^\tau[k]$ and $\lambda_2^*[k]$ from Section 3.7, one obtains

$$\begin{aligned}
\|f^{\lambda_1^\tau[k]} - f^*\|_n^2 &= \|(I_n - A_{\lambda_1^\tau[k]})F^*\|_n^2 + \|A_{\lambda_1^\tau[k]}\varepsilon\|_n^2 + 2\langle A_{\lambda_1^\tau[k]}\varepsilon, (I_n - A_{\lambda_1^\tau[k]})F^*\rangle_n \\
&\leq 2B^2(\lambda_1^\tau[k]) + 2v(\lambda_1^\tau[k]) \\
&\leq 4V(\lambda_2^*[k]) + 6(y - \Delta y) + 2V(\lambda_1^\tau[k]) + c_1\sqrt{\frac{\log n}{n}} \\
&\leq 8V(\lambda_2^*[k]) + 8(y - \Delta y) + c_1\sqrt{\frac{\log n}{n}}
\end{aligned}$$

with probability at least $1 - 16\exp\left(-c_2 n\min\left((y-\Delta y)^2, y - \Delta y\right)\right)$, where $y > 0$ is arbitrary, $y - \Delta y \geq 0$.

In addition to that, if λ_2^* from Eq. (3.30) exists, then $V(\lambda_2^*[k]) \leq 1/2\mathrm{MSE}(\lambda_2^*[k])$, and

$$\|f^{\lambda_1^\tau[k]} - f^*\|_n^2 \leq 4\mathrm{MSE}(\lambda_2^*[k]) + 8(y - \Delta y) + c_1\sqrt{\frac{\log n}{n}} \tag{3.49}$$

with the same probability.

Define $u \coloneqq c_2 n \min\left((y-\Delta y)^2, y-\Delta y\right)$, then one concludes that

$$\|f^{\lambda_1^\tau[k]} - f^*\|_n^2 \leq 4\mathrm{MSE}(\lambda_2^*[k]) + C\left(\frac{\sqrt{u}}{\sqrt{n}} + \frac{u}{n}\right) + c_1\sqrt{\frac{\log n}{n}} \tag{3.50}$$

with probability at least $1 - 16\exp(-u)$, where $u \geq 0$ is arbitrary since $y - \Delta y$ is arbitrary, constants C and c_1 can depend on d, σ, and $\mathcal{M}$.

Chapter 4

Empirical evaluation of MDP rule for linear estimators

Abstract

The present chapter aims at comparing the practical behavior of the minimum discrepancy stopping rule for choosing the tuning parameter in linear estimators, with other existing and the most used in practice model selection methods. We split the chapter into four parts. Section 4.1 introduces the linear estimator and provides some examples of this estimator. Section 4.2 defines the competitive stopping rules and "oracle" stopping rule. Section 4.3 presents experiments on some artificial data sets, while Section 4.4 presents experiments on some real data sets.

4.1 Introduction

Let us first recall examples of the linear estimator that we take into account. Assume that we have data $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathbb{R}$, and a model selection set $\Lambda = \{\lambda_1, \ldots, \lambda_S\}$ for some $S \in \mathbb{N}$ that can potentially depend on the sample size n. A general linear (functional) estimator f^λ of the regression function from the statistical model $y_i = f^*(x_i) + \varepsilon_i,\ \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n$, could be defined as

$$F^\lambda := [f^\lambda(x_1), \ldots, f^\lambda(x_n)]^\top = A_\lambda Y, \qquad \lambda \in \Lambda, \tag{4.1}$$

where A_λ is an $n \times n$ matrix, and $Y = [y_1, \ldots, y_n]^\top$, λ is the parameter to choose (tune/learn). In what follows, three linear estimators are considered.

— **k-nearest neighbor regression [26, 67].** Theoretical analysis of this estimator has been carried out in Chapter 3. We recall that for the case of the k-NN regression estimator, λ is the number of neighbors k to choose; $(A_k)_{ij} = 1/k$ if x_i is a k-nearest neighbor of x_j (measured in the Euclidean norm), otherwise $(A_k)_{ij} = 0,\ \forall i, j, k \in \{1, \ldots, n\}$. Moreover, $(A_k)_{ii} = 1/k$, and $\sum_{j=1}^n (A_k)_{ij} = 1,\ \forall i, k \in \{1, \ldots, n\}$.

— **Nadaraya-Watson regression [85, 116].** For this estimator, the corresponding matrix $A_h = WD^{-1}$, where $D = \mathrm{diag}(W\mathbf{1})$ ($\mathbf{1}$ is the $n \times n$ unit matrix) is the diagonal matrix of row sums and $W_{ij} = \exp(-\|x_i - x_j\|^2/(dh)),\ h > 0$ is the smoothing parameter (bandwidth) to learn.

— **Variable selection in the regression model [102].** Assume the standard nonparametric regression model $Y = F^* + \varepsilon$, where $F^* = [f^*(x_1), \dots, f^*(x_n)]^\top$, $\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ is a *full-rank* fixed design matrix with $d \leq n$, $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_n)$. Given this model, the goal is to choose a subset $J \subseteq \{1, \dots, d\}$ such that redundant variables (features) are omitted. Denote X_J as the matrix of size $n \times |J|$ composed of columns of $\mathbf{X}$ and indexed by J. Then, the final linear (functional) estimator is $F^J = [f^J(x_1), \dots, f^J(x_n)]^\top = A_J Y$, with the matrix $A_J = X_J(X_J^\top X_J)^{-1} X_J^\top$ (also called a projection matrix on the space induced by the columns J). Moreover, one can show [70, p. 233] that $\mathrm{tr}(A_J) = |J|$, which implies that $\frac{\sigma^2}{n}\mathrm{tr}(A_J^\top A_J) = \frac{\sigma^2}{n}\mathrm{tr}(A_J) = \frac{\sigma^2}{n}|J|$. Thus, as we will see later, two subsets of $\{1, \dots, d\}$ with the same cardinality will have the same variance term, and the tuning parameter λ can be chosen equal to the *cardinality* of a subset. We will discuss this point in more detail in Section 4.3.

4.2 Description of the stopping rules to compare

In what follows, we will briefly describe five competitive stopping rules as well as the "undefeated" oracle rule.

Before starting, we should recall [6, Eq. (7)] the expression of the risk error (mean squared error) of the linear estimator (4.1) that can be split into the bias and variance parts (we recall from Chapter 1 that $\mathbb{E}_\varepsilon$ denotes the expectation w.r.t. the noise $\{\varepsilon_i\}_{i=1}^n$, and $\|\cdot\|_n$ denotes the usual $L_2(\mathbb{P}_n)$ empirical norm):

$$\begin{aligned} \mathrm{MSE}(\lambda) &:= \mathbb{E}_\varepsilon \|f^\lambda - f^*\|_n^2 = B^2(\lambda) + V(\lambda), \quad \text{where} \\ B^2(\lambda) &:= \|(I_n - A_\lambda)F^*\|_n^2, \quad V(\lambda) := \frac{\sigma^2}{n}\mathrm{tr}\left(A_\lambda^\top A_\lambda\right). \end{aligned} \tag{4.2}$$

Minimum discrepancy principle.

First, mimicking results from Chapter 3, assume that when we iterate over the grid $\lambda \in \{\lambda_1, \dots, \lambda_S\}$, the variance term $V(\lambda)$ of the linear estimator $A_\lambda Y$ decreases. In particular, for the k-NN estimator, it means that $V(k) = \sigma^2/k$ and $\{\lambda_1, \dots, \lambda_S\} = \{1, \dots, n\}$.

Second, notice that for the k-NN and variable selection estimators, we have

$$\mathrm{tr}(A_\lambda^\top A_\lambda) = \mathrm{tr}(A_\lambda), \quad \lambda \in \Lambda. \tag{4.3}$$

Thus, from Eq. (4.3), one concludes that given the expressions for the empirical risk

$$R_\lambda := \|Y - A_\lambda Y\|_n^2, \tag{4.4}$$

bias term $B^2(\lambda)$, and variance term $V(\lambda)$ from Eq. (4.2) of the linear estimator $A_\lambda Y$, where $Y = F^* + \varepsilon$, it comes

$$\begin{aligned}\mathbb{E}_\varepsilon R_\lambda &= \sigma^2 + B^2(\lambda) - \frac{\sigma^2}{n}\left(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right) \\ &= \sigma^2 + B^2(\lambda) - V(\lambda).\end{aligned}$$

Therefore, the bias-variance trade-off [70, Chapter 7] leads to the reference stopping rule

$$\lambda^* = \inf\left\{\lambda \in \{\lambda_1, \ldots, \lambda_S\} \mid B^2(\lambda) \geq V(\lambda)\right\}, \tag{4.5}$$

that can be equivalently defined as

$$\lambda^* = \inf\left\{\lambda \in \{\lambda_1, \ldots, \lambda_S\} \mid \mathbb{E}_\varepsilon R_\lambda \geq \sigma^2\right\}. \tag{4.6}$$

If λ^* from Eq. (4.5) or Eq. (4.6) does not exists, set $\lambda^* = \lambda_S$. Eq. (4.6) is the population justification of the so-called minimum discrepancy principle rule λ^τ, which will be an estimator of λ^*:

$$\lambda^\tau = \sup\left\{\lambda \in \{\lambda_1, \ldots, \lambda_S\} \mid R_\lambda \leq \sigma^2\right\}. \tag{4.7}$$

If λ^τ from Eq. (4.7) does not exists, set $\lambda^\tau = \lambda_1$. As we have already discussed in Chapter 3, we change inf to sup in Eq. (4.7) due to an arbitrary and uncontrolled behavior of the empirical risk and bias term (lack of monotonicity) between these two points. In simple words, we are not able to provide a tight control of these two quantities "in-between", that is, between the infimum and supremum.

Third, consider now the Nadaraya-Watson regressor with the tuning parameter h to learn, then $\mathrm{tr}\left(A_h^\top A_h\right) \neq \mathrm{tr}\left(A_h\right)$, which results in a slightly different formula for the minimum discrepancy principle:

$$h^\tau = \sup\left\{h \in \{h_1, h_2, \ldots, h_S\} \mid R_h \leq \sigma^2 + \frac{2\sigma^2}{n}\left(\mathrm{tr}\left(A_h^\top A_h\right) - \mathrm{tr}\left(A_h\right)\right)\right\}; \tag{4.8}$$

if h^τ does not exists, set $h^\tau = h_1$.

We recall (see Chapter 3) that the minimum discrepancy principle stopping rule k^τ for the k-NN regression estimator has been proved to output a minimax optimal functional estimator as soon as the minimax rate of the function class, where the regression function f^* lies in, is not faster than $\mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$. It is true, for instance, with the class of Lipschitz functions on a bounded domain that can be defined as

$$\mathcal{F}_{\mathrm{Lip}}(L) = \left\{f : [0,1]^d \mapsto \mathbb{R} \mid f(0) = 0,\ f \text{ is L-Lipschitz}\right\}, \tag{4.9}$$

where f is L-Lipschitz means that $|f(x) - f(x')| \leq L\|x - x'\|$ for all $x, x' \in [0,1]^d$.

Generalized cross-validation [42, 53, 70].

The generalized (GCV) cross-validation strategy [8, 53] was introduced in least-squares regression as a rotation-invariant version of the leave-one-out cross-validation procedure. The GCV estimator of the risk error of the linear estimator $A_\lambda Y,\ \lambda \in \Lambda = \{\lambda_1, \dots, \lambda_S\}$, is defined as

$$R_{GCV}(f^\lambda) = \frac{n^{-1}\|Y - A_\lambda Y\|^2}{(1 - n^{-1}\mathrm{tr}(A_\lambda))^2}.$$

The final generalized cross-validation stopping rule is

$$\lambda_{GCV} \coloneqq \underset{\lambda \in \{\lambda_1, \dots, \lambda_S\}}{\mathrm{argmin}} \left\{ R_{GCV}(f^\lambda) \right\}. \tag{4.10}$$

Note that for the k-NN regression estimator, we exclude the case $k = 1$ in Eq. (4.10) because $A_1 Y = Y$. GCV should be close to C_L model selection procedure (e.g., Mallows' C_p generalized to linear estimators [79]). The efficiency (a.k.a. the asymptotic optimality of a model selection procedure) of GCV has been proved, for instance, for the k-NN regression estimator in [76]. As its main feature, in smoothing problems, GCV is able to alleviate the tendency of other cross-validation methods to undersmooth. Notice that, if the matrices $\{A_\lambda\}_{\lambda \in \Lambda}$ are already computed, the computational time of the generalized cross-validation is $\mathcal{O}\left(n^2|\Lambda|\right)$, which is higher than $\mathcal{O}\left(n^2\left(|\Lambda| - k^\tau\right)\right)$ for the minimum discrepancy principle stopping rule λ^τ (4.7).

Hold-out cross-validation stopping rule [8, 63, 117].

The Hold-out cross-validation strategy [8, 63] is described as follows. The data $\{x_i, y_i\}_{i=1}^n$ are randomly split into two parts of equal size: the training sample $S_{\text{train}} = \{x_{\text{train}}, y_{\text{train}}\}$ and the test sample $S_{\text{test}} = \{x_{\text{test}}, y_{\text{test}}\}$ so that the training and test samples represent a half of the whole data set. For each $\lambda \in \Lambda = \{\lambda_1, \dots, \lambda_S\}$, one trains a linear estimator (4.1) on S_{train} and evaluates its performance by $R_{HO}(f^\lambda) = \frac{1}{n}\sum_{i \in S_{\text{test}}}(f^\lambda(x_i) - y_i)^2$, where $f^\lambda(x_i)$ denotes the output of a learning algorithm trained for λ and evaluated at the point $x_i \in x_{\text{test}}$. Then, the Hold-out CV stopping rule is defined as

$$\lambda_{HO} \coloneqq \underset{\lambda \in \{\lambda_1, \dots, \lambda_S\}}{\mathrm{argmin}} \left\{ R_{HO}(f^\lambda) \right\}. \tag{4.11}$$

The main inconvenience of this stopping rule is the fact that a part of the data is lost, which increases the risk error. As a result, the Hold-out strategy is not stable [8], which often requires some aggregation of it. Nevertheless, [117] derived a non-asymptotic oracle inequality when combining a penalized least-squares estimator with the hold-out. As it was for GCV, the (asymptotic) computational time of the Hold-out strategy is $\mathcal{O}\left(n^2|\Lambda|\right)$.

Mallows' C_p stopping rule [70, 76, 79].

We will apply Mallows' C_p stopping rule [79] specifically for the case of variable selection in the model $Y = F^* + \varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(F^*, \sigma^2 I_n)$, where $F^* = [f^*(x_1), \dots, f^*(x_n)]^\top$ and the full-rank design matrix $\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$, $d \leq n$. Assume that one selects a subset of the variables $J \subseteq \{1, \dots, d\}$ and construct a matrix X_J based on this subset (the matrix of size $n \times |J|$), then $A_J = X_J (X_J^\top X_J)^{-1} X_J^\top$, and the estimator of the risk of the linear estimator $A_J Y$ is defined as

$$R_{Cp}(f^J) = n^{-1} \|Y - A_J Y\|^2 + 2 \frac{|J| \widehat{\sigma}^2}{n}, \tag{4.12}$$

where $\widehat{\sigma}^2 = \frac{\|Y - AY\|^2}{n-d}$, $A = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ with $\mathbf{X} = X_{\{1,\dots,d\}}$, is the noise variance estimation. In Eq. (4.12), $|J|$ denotes the cardinality of the subset J. Then, the Mallows' C_p stopping rule is defined as

$$J_{Cp} \coloneqq \underset{|J| \in \{1,\dots,d\}}{\operatorname{argmin}} \left\{ R_{Cp}(f^J) \right\}. \tag{4.13}$$

C_p and related model selection procedures have been proved to be efficient (a.k.a. asymptotically optimal) or to satisfy oracle inequalities in some frameworks (see, e.g., [27] and references therein for more details). If the matrices $\{A_J\}_{J \subseteq \{1,\dots,d\}}$ are already computed, the computational time of the Mallow's C_p strategy is $\mathcal{O}\left(n^2 |J|\right)$, which is higher than $\mathcal{O}\left(n^2 \left(|J| - J^\tau\right)\right)$ for the minimum discrepancy principle stopping rule J^τ (4.7).

Notice that the competitive stopping rules (generalized cross-validation, Hold-out, and Mallows' C_p) involve the computation of the empirical risk $\|Y - A_\lambda Y\|_n^2$, as we do when using the minimum discrepancy principle (4.7). However for MDP, we will not compute the empirical risk for all $\lambda \in \Lambda$, whereas it is the case for the mentioned rules.

Bias-variance trade-off stopping rule.

The third stopping rule is the one introduced in Eq. (4.5). This stopping rule is the classical bias-variance trade-off stopping rule that provides minimax-optimal rates (see the monographs [108, 116]):

$$\lambda^* = \inf \left\{ \lambda \in \{\lambda_1, \dots, \lambda_S\} \mid B^2(\lambda) \geq V(\lambda) \right\}, \quad S \in \mathbb{N}. \tag{4.14}$$

This stopping rule is introduced for comparison purposes only because it cannot be computed in practice (the bias term is unknown). However, it could serve as a reference in the present simulated experiments.

Oracle stopping rule.

The "oracle" stopping rule is defined as

$$\lambda_{\text{or}} \coloneqq \underset{\lambda \in \{\lambda_1,\ldots,\lambda_S\}}{\operatorname{argmin}} \left\{ \mathbb{E}_\varepsilon \|f^\lambda - f^*\|_n^2 \right\} \tag{4.15}$$

and minimizes the risk error. Note that this stopping rule is not computable from the data, since one has to know the regression function f^* to compute it. Moreover, we do not have access to the whole curve of the risk error. Nevertheless, it serves as a convenient lower bound on the risk error for the simulations with artificial data.

4.3 Artificial data

First, the goal is to perform simulation experiments for making a comparison of all mentioned stopping rules on artificial data.

4.3.1 Description of the simulation design for k-NN and Nadaraya-Watson regression

We start with the description of the simulation design for the k-NN and Nadaraya-Watson regression estimators. In this case, the data is generated according to the regression model $y_j = f^*(x_j) + \varepsilon_j$, where $\varepsilon_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $j = 1, \ldots, n$, is Gaussian noise. We choose the covariates $x_j \overset{i.i.d.}{\sim} \mathbb{U}[0,1]^3$ (uniform) or $x_j \overset{i.i.d.}{\sim} \mathcal{N}(0, I_3)$ (standard normal), $j = 1, \ldots, n$, and $\sigma \in \{0.1, 0.15, 0.4\}$ is assumed to be known. Consider two regression functions with different smoothness: a "smooth" $f_1^*(x) = 1.5 \cdot \left[\|x - 0.5\|/\sqrt{3} - 0.5 \right]$ and a "sinus" $f_2^*(x) = 1.5 \cdot \sin(\|x\|/\sqrt{3})$, for $x \in [0,1]^3$ or $x \in \mathbb{R}^3$. Notice that both functions belong to the class of Lipschitz functions (4.9) on $[0,1]^3$. The sample size n varies from 50 to 250.

Assume that we have the grids of values $k \in \{1, 2, \ldots, n\}$ for the k-NN regression and $h \in \{h_1, h_2, \ldots, h_n\}$ (thus, $S = n$) for the Nadaraya-Watson regression, where $h_1 = \min\limits_{i,j \in \{1,\ldots,n\}} \|x_i - x_j\|^2$, $h_i = h_1 + \frac{(h_n - h_1)(i-1)}{n-1}$, $i = 1, \ldots, n$, where $h_n = \max\limits_{i,j \in \{1,\ldots,n\}} \|x_i - x_j\|^2/10$ for the "smooth" function; $h_n = \max\limits_{i,j \in \{1,\ldots,n\}} \|x_i - x_j\|^2/30$ for the "sinus" function (constants 10 and 30 were calibrated so that one can observe the oracle rule (4.15) around $h_{\lfloor n/2 \rfloor}$).

For $\lambda \in \{k, h\}$, the k-NN and Nadaraya-Watson learning algorithms (4.1) are trained, first, for $\lambda = \lambda_n$, after that we decrease the value of λ until $\lambda = \lambda_1$ such that at each step of the iteration procedure we increase the variance term $V(\lambda)$. In other words, the model becomes more complex successively due to the increase of its "degree of freedom" measured by $\operatorname{tr}(A_\lambda)$. We should remark here that for the Nadaraya-Watson estimator, the variance term $V(h)$ is proportional to $\operatorname{tr}\left(A_h^\top A_h\right)$, and

not to $\operatorname{tr}(A_h)$, so the previous statement holds only approximately, meaning $\operatorname{tr}\left(A_h^\top A_h\right) \approx \operatorname{tr}(A_h)$. If the condition in Eq. (4.7) or Eq. (4.8) is satisfied, the process is stopped and it outputs the stopping rule λ^τ. An illustration of the discussed strategy is presented in Figure 4.1 (panel (a)).

The performance of the stopping rules is measured in terms of the empirical $L_2(\mathbb{P}_n)$ norm $\|f^\lambda - f^*\|_n^2$ averaged over $N = 80$ repetitions (over the noise $\{\varepsilon_j\}_{j=1}^n$).

4.3.2 Description of the simulation design for variable selection regression.

The simulation design for the variable selection problem is a bit more involved and needs some theoretical justifications that we will mention in what follows.

Recall that we consider the regression model

$$Y = F^* + \varepsilon \in \mathbb{R}^n, \ \varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_n), \tag{4.16}$$

where $F^* = [f^*(x_1), \ldots, f^*(x_n)]^\top$ and $\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ is a full-rank matrix with the rank $r = d \leq n$. Then, [67, Theorem 3.2] a typical minimax optimal lower bound of the risk error in terms of the $L_2(\mathbb{P}_n)$ norm is defined as follows.

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{\mathrm{Lip}}(L)} \mathbb{E}_\varepsilon \left[\|\widehat{f} - f^*\|_n^2 \right] \geq c_l n^{-\frac{2}{2+d}}, \tag{4.17}$$

where c_l is some positive constant, functional space $\mathcal{F}_{\mathrm{Lip}}(L)$ was defined in Eq. (4.9), and $\widehat{f}$ is any measurable of the data functional estimator.

At the same time, assume that one has at hand an abstract set of functional estimators $\Theta_f = \{f_1, \ldots, f_M\}$ for some $M \in \mathbb{N}$, and the goal is to select "the best estimator" $\widehat{f}$ from Θ_f (the so-called model selection aggregation setting introduced in [86, 109]). Define $f_\omega = \sum_{j=1}^M \omega_j f_j$ for any $\omega = (\omega_1, \ldots, \omega_M) \in \mathbb{R}^M$. Assume that the performance of $\widehat{f}$ is assessed via the following oracle inequality

$$\mathbb{E}_\varepsilon \|\widehat{f} - f^*\|_n^2 \leq \inf_{\omega \in \Omega^M} \|f_\omega - f^*\|_n^2 + \Delta_{n,M}, \tag{4.18}$$

where $\Delta_{n,M} \geq 0$ is a remainder term independent of f^* characterizing the price to pay to select an estimator from Θ_f, and the set Ω^M is the set of all vertices of $\{\omega = (\omega_1, \ldots, \omega_M) \in \mathbb{R}^M \mid \omega_j \geq 0, \sum_{j=1}^M \omega_j \leq 1\}$, except the vertex $(0, \ldots, 0) \in \mathbb{R}^M$. Then, [40, Theorem 5.1] proved that under the uniform boundness assumption of f^* and $\{f_1, \ldots, f_M\}$, the smallest possible (minimax) remainder term $\Delta_{n,M}$ is of the order $\mathcal{O}\left(\frac{\log M}{n}\right)$.

Suppose that we have at hand *all* possible subsets of the set $\{1, 2, \ldots, d\}$ and the estimators associated with these subsets, then, in total, there are $M = \sum_{i=1}^d \binom{d}{i} = 2^d$ estimators. The next step

would be, of course, to compare the rate from Ineq. (4.17) and the rate

$$\mathcal{O}\left(\frac{\log M}{n}\right) = \mathcal{O}\left(\frac{d}{n}\right). \tag{4.19}$$

One can conclude that, in the worst case $d \asymp n$ (which is of the main interest for variable selection), and the best achievable rate from (4.19) is *always* slower than the minimax rate presented in Eq. (4.17). Furthermore, it will be computationally infeasible to deal with all subsets of $\mathbf{X}$ starting from $d > 20$ (approximately). These two obstacles force us to reduce the number of estimators at hand. Otherwise, there would be only a sub-optimal solution. Notice that this problem was a reason why Arlot and Bach [6] restricted the cardinality of the model selection set Λ (see Assumption $(H\Lambda)$). Therefore, to overcome these obstacles, we propose the following procedure to choose a subset from the set $\{1, 2, \ldots, d\}$.

1: $|J| = 0$
2: **repeat**
3: $\quad |J| = |J| + 1$
4: $\quad$ Choose randomly $|J|$ variables from $\{1, \ldots, d\}$
5: $\quad$ Given the full-rank matrix $\mathbf{X}$, construct the matrix X_J from the chosen variables and calculate $A_J = X_J(X_J^\top X_J)^{-1}X_J^\top$
6: $\quad$ Calculate the empirical risk $R_J = \|Y - A_J Y\|_n^2$
7: **until** $R_J \leq \sigma^2$ or $|J| = d$

The procedure above will output the minimum discrepancy principle rule $J^\tau \in \{1, \ldots, d\}$ and linear estimator $F^{J^\tau} := [f^{J^\tau}(x_1), \ldots, f^{J^\tau}(x_n)]^\top = A_{J^\tau} Y$ associated with this rule. This model selection procedure is meaningful, i.e., its statistical performance is comparable to the performance of the bias-variance trade-off from Eq. (4.5) up to the remainder term $\mathcal{O}\left(\sqrt{\frac{\log d}{n}}\right) = \mathcal{O}\left(\sqrt{\frac{\log r}{n}}\right)$. The latter is justified by what follows.

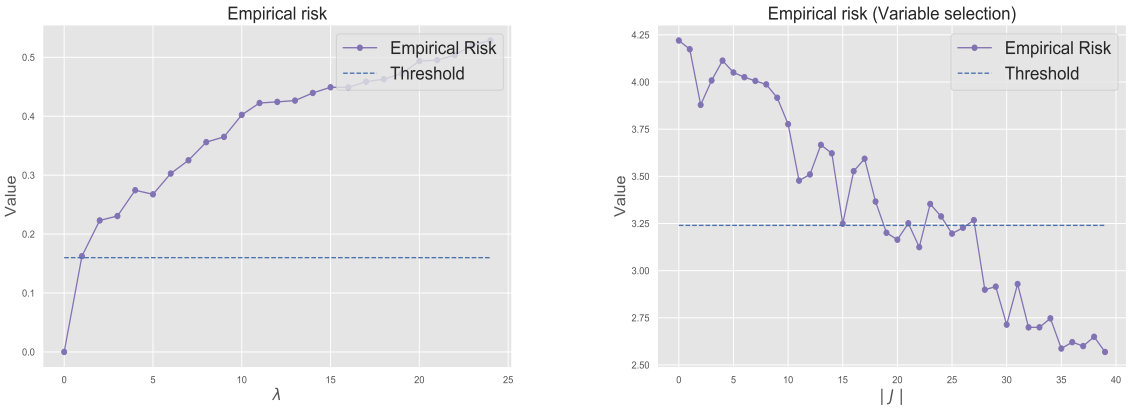
Theorem 4.3.1. *Under the assumption that, for all $x \in \mathcal{X}$, $|f^*(x)| \leq \mathcal{Q}$ for some constant $\mathcal{Q} > 0$, for arbitrary $u \geq 0$,*

$$\|f^{J^\tau} - f^*\|_n^2 \leq \underbrace{8\ \mathrm{MSE}(J^*)}_{\text{Main term}} + \underbrace{C_1\left(\frac{u}{n} + \frac{\sqrt{u}}{\sqrt{n}}\right) + C_2\sqrt{\frac{\log r}{n}}}_{\text{Rem. term}} \tag{4.20}$$

with probability at least $1 - 16\exp(-u)$, where positive constants C_1 and C_2 can depend on σ and $\mathcal{Q}$; J^* is the bias-variance trade-off defined in Eq. (4.5).

Proof of Theorem 4.3.1. The proof is a direct adaptation of the proof of Theorem 3.4.1 in Chapter 3. Let us list the main steps of the proof.

First, we notice that the operator norm of the matrix A_J : $\|A_J\|_2 \leq 1$ for any $J \in \{1, \ldots, d\}$. It



(a) (b)

Figure 4.1 – The minimum discrepancy principle: k-NN regression for the panel (a); variable selection for the panel (b). "Threshold" corresponds to the value σ^2.

implies that $\|M_J\|_2 \coloneqq \|I_n - A_J\|_2 \leq 1$. After that,

$$\begin{aligned}\|f^{J^\tau} - f^*\|_n^2 &= B^2(J^\tau) + \|A_{J^\tau}\varepsilon\|_n^2 + 2\langle A_{J^\tau}\varepsilon, (I_n - A_{J^\tau})F^*\rangle_n \\ &\leq 2B^2(J^\tau) + 2\|A_{J^\tau}\varepsilon\|_n^2.\end{aligned}$$

Further, it is sufficient to upper bound the bias $B^2(J^\tau)$ in the same way as in Lemma 3.10.2 from Chapter 3 and control $\|A_J\varepsilon\|_n^2$ via the variance term $V(J) = \sigma^2 \mathrm{tr}\left(A_J\right)/n$ with high probability, for all $J \in \{1, \ldots, d\}$. The only difference between the proof of Theorem 3.4.1 in Chapter 3 and the present one is the remainder term in the oracle-type inequality – it becomes $\mathcal{O}\left(\sqrt{\frac{\log d}{n}}\right) = \mathcal{O}\left(\sqrt{\frac{\log r}{n}}\right)$ – since we have r estimators instead of n, as it was for the k-NN estimator. ■

Let us comment the statement of Theorem 4.3.1. The main conclusion we can make is that, with high probability, the performance (prediction error) of J^τ is close to the performance of the bias-variance trade-off rule J^* (constant 8 could be improved) up to the term $\mathcal{O}\left(\sqrt{\log r/n}\right)$. This remainder term should be sufficiently fast (compared to a minimax lower bound), for example, in the case of L-Lipschitz functions (4.17) when $d > 2$. Besides that, we should mention that the bias-variance trade-off J^*, in this case, is random itself, meaning that it depends on the particular choice of chosen subsets made in Algorithm 0. Nevertheless, for sufficiently smooth regression functions f^*, the bias term will not fluctuate much, and J^* should provide a good approximation to the "true bias-variance trade-off" (when one considers all subsets of $\mathbf{X}$).

We take into account everything what has been said previously, the regression model from Eq. (4.16), and the covariates $x_i \overset{i.i.d.}{\sim} \mathbb{U}[0,1]^d$, $i \in \{1, \ldots, n\}$, with $d = 40$, $\sigma \in \{0.15, 0.4\}$, where n changes as follows: $n \in \{80, 100, 150, 200, 250, 400\}$. As usual, we are interested in the estimation of

two functions: a "smooth" $f_1^*(x) = 1.5 \cdot \left[\|x - 0.5\| / \sqrt{40} - 0.5 \right]$ and a "sinus" $f_2^*(x) = 1.5 \cdot \sin\left(\|x\| / \sqrt{40} \right)$, where $x \in [0,1]^{40}$.

Assume that we have the grid of values $|J| \in \{1, 2, \ldots, d\}$ for the variable selection (projection) estimator. The linear estimator (4.1) is trained first for $|J| = 1$ (by choosing a random column of $\mathbf{X}$), further we increase the value of $|J|$ by one until $|J| = n$ (corresponds to taking the whole design matrix $\mathbf{X}$). However, if the condition in Eq. (4.7) is satisfied, the learning process is stopped, producing J^τ. An illustration of the discussed strategy is presented in Figure 4.1 (panel (b)). Besides already introduced rules (4.10), (4.14), (4.15), we consider Mallows C_p criterion from Eq. (4.12), and what we call "the full-rank model selection criterion" J_{FR}, meaning simply that $J_{FR} = \{1, \ldots, d\}$, and consequently,

$$F^{J_{FR}} = A_{\{1,\ldots,d\}} Y = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y. \tag{4.21}$$

The performance of the stopping rules will be measured in terms of the empirical $L_2(\mathbb{P}_n)$ norm $\|f^J - f^*\|_n^2$ averaged over $N = 50$ repetitions (over the noise $\{\varepsilon_j\}_{j=1}^n$).

4.3.3 Results of the simulation experiments for k-NN and Nadaraya-Watson regression.

In this subsection, we explain the results achieved by using the k-NN and Nadaraya-Watson regression estimators. Figure 4.2 and Figure 4.3 display the resulting (averaged over 80 repetitions) $L_2(\mathbb{P}_n)$ error of k^τ and h^τ from Eq. (4.7) and Eq. (4.8), respectively, $k_{\text{or}}/h_{\text{or}}$ from Eq. (4.15), k^*/h^* from Eq. (4.14), k_{HO}/h_{HO} from Eq. (4.11), and k_{GCV}/h_{GCV} from Eq. (4.10), versus the sample size n. In particular, Figure 4.2 shows the results for the k-NN regression estimator, whereas Figure 4.3 provides the results for the Nadaraya-Watson regression estimator.

Let us start to discuss the results from Figure 4.2. At first, from all the graphs, (almost) all the curves do not increase as the sample size n grows. Without accounting the oracle performance, one achieves the best performance by either the k^* or k_{GCV} stopping rules. This good behavior was expected since k^* represents the well-known bias-variance trade-off, and k_{GCV} has been proved to be an asymptotically optimal model selection criterion (see, e.g., [76]).

In more detail, Figure 4.2a (the "smooth" regression function and uniform covariates) indicates that k^* achieves the best performance (if we do not take into account the oracle performance). Besides that, the minimum discrepancy principle rule k^τ is almost uniformly better than k_{HO}. Moreover, the gap between k^τ and k^*/k_{GCV} is getting smaller as the sample size increases. This behavior supports the theoretical part of the present work (see Chapter 3) because k^τ should serve as an estimator of k^*. Since k^* is the well-known bias-variance trade-off, the minimum discrepancy principle stopping rule seems a meaningful model selection strategy.

Now, let us move to Figure 4.2b (the "sinus" regression function and uniform covariates), where the situation is slightly different. In this case, the best performance is achieved again by k^* (except

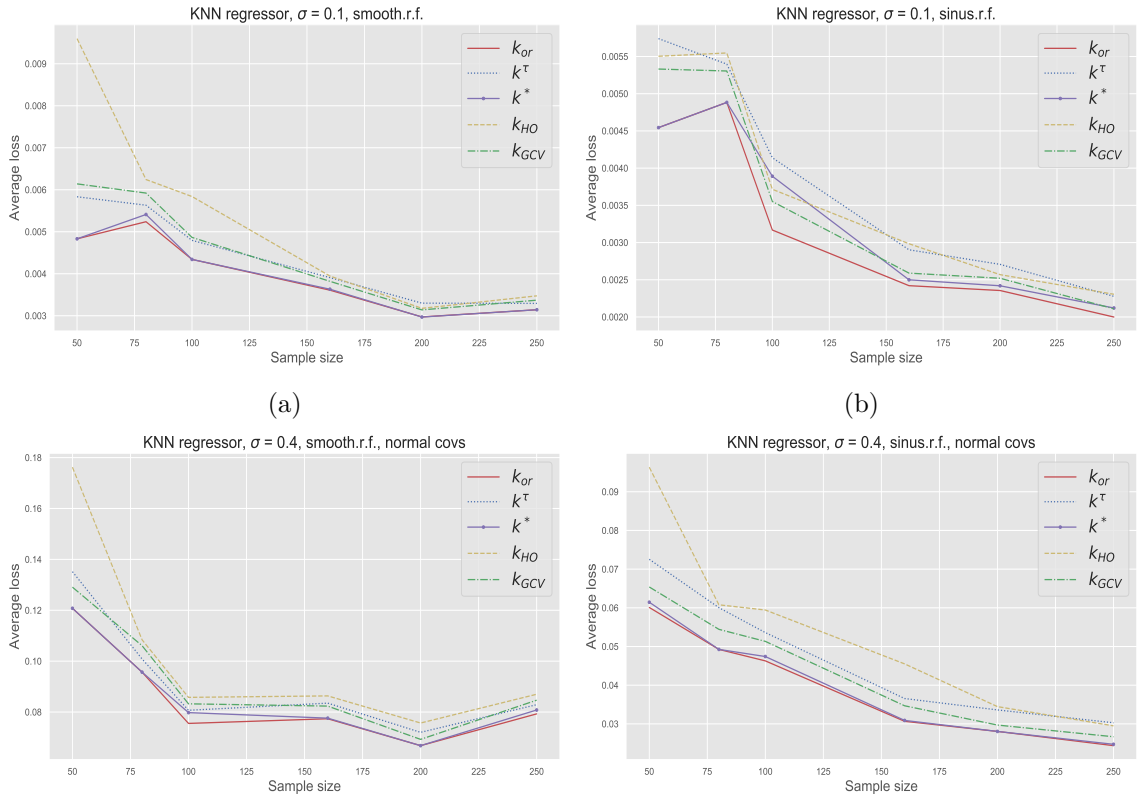


Figure 4.2 – The k-NN estimator (4.1) performance with two noised regression functions: smooth $f_1^*(x) = 1.5 \cdot \left[\|x - 0.5\|/\sqrt{3} - 0.5\right]$ for the panels (a) and (c), and "sinus" $f_2^*(x) = 1.5 \cdot \sin(\|x\|/\sqrt{3})$ for the panels (b) and (d), with uniform covariates $x_j \overset{i.i.d.}{\sim} \mathbb{U}[0,1]^3$ (panels (a) and (b)) or standard normal covariates $x_j \overset{i.i.d.}{\sim} \mathcal{N}(0, I_3)$ (panels (c) and (d)), $j = 1, \ldots, n$. Each curve corresponds to the $L_2(\mathbb{P}_n)$ squared norm error for the stopping rules (4.7), (4.14), (4.15), (4.11), (4.10), averaged over 80 independent trials, versus the sample size $n = \{50, 80, 100, 160, 200, 250\}$.

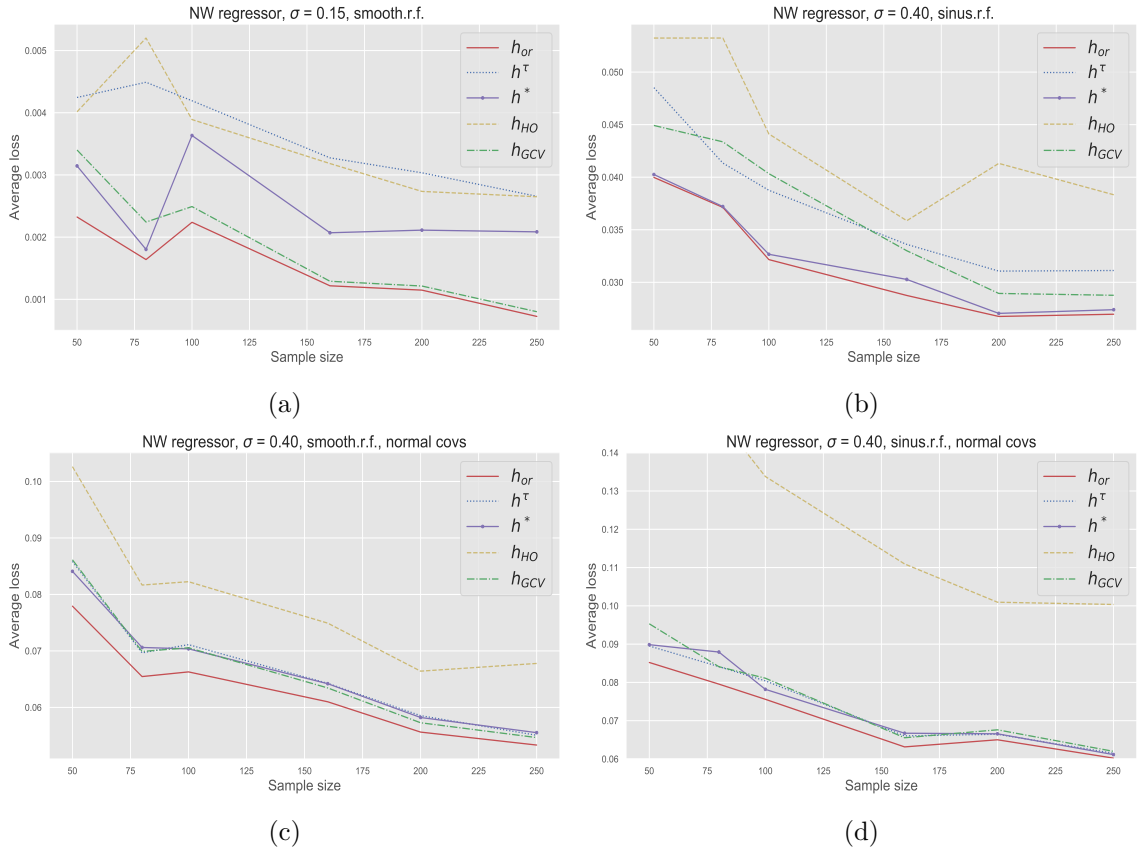
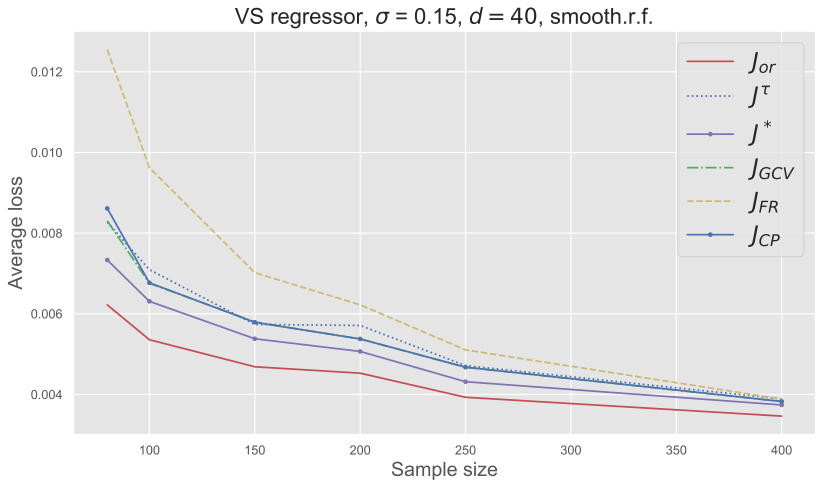
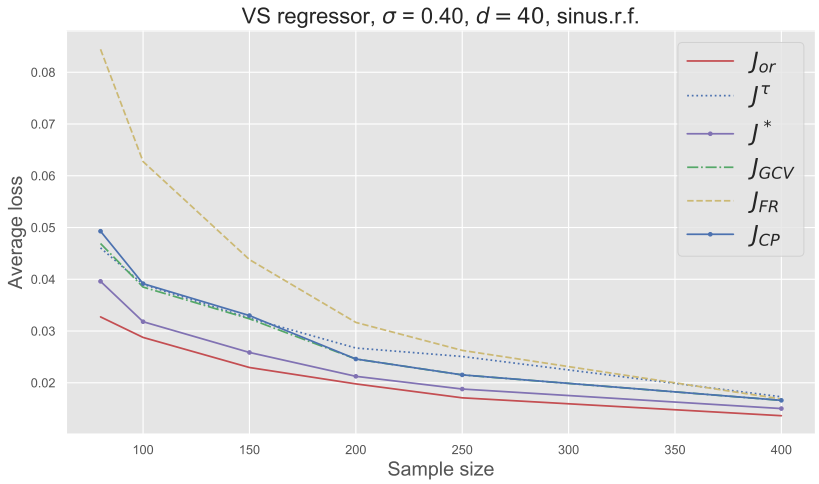


Figure 4.3 – The Nadaraya-Watson estimator (4.1) performance with two noised regression functions: smooth $f_1^*(x) = 1.5 \cdot \left[\|x - 0.5\|/\sqrt{3} - 0.5\right]$ for the panels (a) and (c), and "sinus" $f_2^*(x) = 1.5 \cdot \sin(\|x\|/\sqrt{3})$ for the panels (b) and (d), with uniform covariates $x_j \overset{i.i.d.}{\sim} \mathbb{U}[0,1]^3$ (panels (a) and (b)) or standard normal covariates $x_j \overset{i.i.d.}{\sim} \mathcal{N}(0, I_3)$ (panels (c) and (d)), $j = 1, \dots, n$. Each curve corresponds to the $L_2(\mathbb{P}_n)$ squared norm error for the stopping rules (4.8), (4.14), (4.15), (4.11), (4.10), averaged over 80 independent trials, versus the sample size $n = \{50, 80, 100, 160, 200, 250\}$. Moreover, $\mathrm{SNR} = \|f_j^*\|_n/\sigma \in [1, 5]$, $j \in \{1, 2\}$.



(a)



(b)

Figure 4.4 – The variable selection estimator (4.1) performance in nonparametric regression $Y_i = f^*(x_i) + \varepsilon_i$, $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, where $x_j \overset{i.i.d.}{\sim} \mathbb{U}[0,1]^{40}$, $j = 1, \ldots, n$: "smooth" regression function $f_1^*(x) = 1.5 \cdot \left[\|x - 0.5\| / \sqrt{40} - 0.5 \right]$ for the panel (a), and "sinus" regression function $f_2^*(x) = 1.5 \cdot \sin\left(\|x\| / \sqrt{40} \right)$ for the panel (b). Each curve corresponds to the $L_2(\mathbb{P}_n)$ squared norm error for the stopping rules (4.7), (4.14), (4.15), (4.11), (4.10), averaged over 50 independent trials, versus the sample size $n = \{80, 100, 150, 200, 250, 400\}$. Moreover, $\mathrm{SNR} = \|f_j^*\|_n / \sigma \approx 2$, $j \in \{1, 2\}$.

for $k = 100$): its results are close to the results for the oracle rule. As for the data-driven model selection methods, the stopping rules k^τ and k_{HO} perform almost equivalently. Increasing the number of repetitions of simulations experiments should reduce the performance gap between k_{GCV} and k^τ.

If we consider Figures 4.2c and 4.2d (the "smooth"/"sinus" regression functions and standard normal covariates), then one can conclude that the MDP-based stopping rule k^τ performs favorably in comparison to the generalized cross-validation stopping rule, which is asymptotically optimal [76].

Further, we move to Figure 4.3, where the Nadaraya-Watson estimator is analyzed. Overall, without accounting the performance of h_{or}, the winners are (most of the time) the same as for Figure 4.2 – h^* and h_{GCV}.

More precisely, starting from the panel (a) (the "smooth" regression function and uniform covariates), the performances of the hold-out rule h_{HO} and the minimum discrepancy principle h^τ are comparable. However, there is a bizarre behavior of h^τ and h^* for the sample sizes $n \geq 100$, which could be explained by the randomness of the covariates $\{x_i\}_{i=1}^n$. Apart from that, h_{GCV} shows the best results.

Moving to the panel (b) of Figure 4.3 (the "sinus" regression function and uniform covariates), the best performance is achieved by the bias-variance trade-off h^* while the MDP rule h^τ is largely better than the hold-out rule h_{HO} and comparable to h_{GCV}.

If one considers Figures 4.3c and 4.3d (the "smooth"/"sinus" regression functions and standard normal covariates), then we can conclude that the performance of the MDP-based rule h^τ is similar to that of h^*(the bias-variance trade-off) and h_{GCV} (generalized cross-validation).

It is worth to mention that even though h^τ shows comparable performance w.r.t., e.g., h_{GCV} or h_{HO}, it is still in our best interest to extend the theoretical results achieved in Chapter 3 to the Nadaraya-Watson regression estimator.

4.3.4 Results of the simulation experiments for variable selection regression

Here, let us explain the results that we obtained in Figure 4.4, where the variable selection estimator is analyzed. We start with the panel (b), where the results for the "sinus" regression function $f_2^*(x)$ are demonstrated. Firstly, if we do not consider the oracle rule J_{or} (4.15), the bias-variance trade-off (4.14) performs the best, and the results for generalized cross-validation and Mallow's C_p are (almost) the same. Secondly, the minimum discrepancy principle J^τ is uniformly better than the full-rank rule J_{FR}. Moreover, we remark that as the sample size n increases, it becomes more and more statistically meaningful to use all the variables of the design matrix $\mathbf{X}$. It can be explained as follows: if the sample size n increases, $r/n \to 0$ (r is the constant rank of $\mathbf{X}$), and the rate in Eq. (4.19) becomes close to $\mathcal{O}(1/n)$, which is the fast rate. Thus, it is more reasonable to consider more subsets of $\{1, \ldots, d\}$ (we recall that there were only $r = d$ selected subsets in our simulated experiments). Notice that when the number of variables and sample size are of the same order ($n \leq 150$), the performance of J^τ is close

to that of J_{CP} and J_{GCV}. In addition to that, for $n \geq 300$, one can say that the performance of J^{τ} is getting close to that of J_{CP} and J_{GCV} as the sample size increases. Panel (a), where we reported the results for the "smooth" regression function $f_1^*(x)$, provides us with almost the same arguments regarding the conclusion.

4.4 Real data

Second, we tested the performance of the early stopping rule (4.7) for choosing the parameter k in the k-NN regression estimator on five different data sets, mostly taken from the UCI repository [57].

4.4.1 Data sets description

Let us start with the description of the data sets.

The wine quality data set (Wine Quality) contains 11-dimensional input points corresponding to the physico-chemistry of wine samples, the output points are the wine quality.

The housing data set (Boston Housing Prices) concerns the task of predicting housing values in areas of Boston (USA), the input points are 13-dimensional.

Diabetes data set consists of 10 columns that measure different patient's characteristics (age, sex, body mass index, ...), the output is a quantitative measure of disease progression one year after the baseline.

The Power Plant data set contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the plant was set to work with a full load.

California Houses Prices data set [88] contains information from the 1990 California census. The input variables are "total bedrooms", "total rooms", etc. The output variable is the median house value for households within a block (measured in US Dollars).

Notice that for "California Houses Prices" and "Power Plants" data sets, we take the first 3000 samples in order to speed up the calculations.

4.4.2 Description of the simulation design

Assume that we are given one of the data sets described above. Let us rescale each variable (feature) of this data set $\widetilde{x} \in \mathbb{R}^n$ such that all the components $\widetilde{x}_i,\ i = 1, \ldots, n$, belong to $[0, 1]$:

$$\widetilde{x}_i = \frac{\widetilde{x}_i - \min(\widetilde{x})}{\max(\widetilde{x}) - \min(\widetilde{x})},\ i = 1, \ldots, n,$$

where $\min(\widetilde{x})$ and $\max(\widetilde{x})$ denote the minimum and the maximum components of the vector $\widetilde{x}$.

After that, we split the data set into two parts: one is denoted $S_{\text{train}} = \{x_{\text{train}}, y_{\text{train}}\}$ (70 % of the whole data) and is made for training and model selection (early stopping rules k^{τ}, k_{GCV}, and

k_{HO}), the other one (30 % of the whole data) is denoted $S_{\text{test}} = \{x_{\text{test}}, y_{\text{test}}\}$ and made for making a prediction on it. Then, our experimental design is divided into four parts.

In the beginning, we estimate the noise variance σ^2 from the regression model (3.1). There is a large amount of work on the efficient estimation of σ^2 in nonparametric regression [68, 94]. In our simulated experiments, we take the estimator from [116, Eq. (5.86)], which is a (low-bias model) consistent estimator of σ^2 under an assumption that f^* is sufficiently smooth. This satisfies our simulation experiments' purposes.

$$\widehat{\sigma}^2 = \frac{\|(I_{n_{\text{train}}} - A_k)Y\|^2}{n_{\text{train}}(1 - 1/k)}, \quad \text{with} \quad k = 2 \;\text{ and }\; n_{\text{train}} = \lceil 0.7n \rceil. \tag{4.22}$$

Further, we compute the MDP stopping rule k^τ from Eq. (4.7). To do that, we compute the k-NN estimator (4.1) and the empirical risk $R_k = \|Y - A_k Y\|_n^2$ for $k_{\max} = \lfloor n_{\text{train}}/2 \rfloor$, and at each step of the iteration process we reduce the value of k by one. Remark that one does not have to calculate *explicitly* the neighborhood matrix A_k for each $k \in \{1, \ldots, k_{\max}\}$, since it is sufficient to do only for $k_{\max}$. This procedure is repeated until the empirical risk crosses the threshold $\widehat{\sigma}^2$. Fig. 4.5 provides an illustration of the minimum discrepancy strategy k^τ applying to two data sets: "Boston Houses Prices" and "Diabetes".

After that, the Holdout stopping rule (4.11) and the generalized cross-validation rule k_{GCV} are calculated. Let us describe how we do that in two steps. We start by defining the grid of values for $k : \{1, 2, \ldots, \lfloor n_{\text{train}}/2 \rfloor\}$. Further, one should compute k_{HO} and k_{GCV} from Eq. (4.11) and Eq. (4.10), respectively, over the mentioned grid.

In the final part, given k^τ, k_{HO}, and k_{GCV}, the goal is to make a prediction on the test data set S_{test}. This can be done as follows. Assume that $x_0 \in x_{\text{test}}$, then the prediction of the k-NN estimator on x_0 can be defined as

$$f^k(x_0) = a_k(x_0)^\top y_{\text{train}}, \tag{4.23}$$

where $x_{\text{train}} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_{n_{\text{train}}}^\top \end{pmatrix} \in \mathbb{R}^{n_{\text{train}} \times d}$, and $a_k(x_0) = [a_k(x_0, x_1), \ldots, a_k(x_0, x_{n_{\text{train}}})]^\top$, with $a_k(x_0, x_i) = 1/k$ if $x_i,\ i \in \{1, \ldots, n_{\text{train}}\}$, belongs to the set of indices of the k nearest neighbors of x_0, denoted as $\mathcal{N}_k(0)$, otherwise 0. Further, one can choose k to be equal to k^τ, k_{HO}, or k_{GCV} that are already computed. Combining all the steps together, one is able to assess the prediction error by

$$\|f^k - y_{\text{test}}\| = \sqrt{1/n_{\text{test}} \sum_{j=1}^{n_{\text{test}}} (f^k(x_j) - (y_{\text{test}})_j)^2}.$$

Dataset	n	d	Train	Test	k_{HO}-error	k_{GCV}-error	k^{τ}-error
Wine Quality	4898	12	3428	1470	27.39	27.39	**26.95**
Power Plants	3000	5	2100	900	116.00	**115.65**	117.00
Boston H. P.	506	13	355	151	53.17	**50.60**	54.29
California H. P.	3000	8	2100	900	11.73	**11.47**	11.48
Diabetes	442	10	310	132	**632.38**	649.31	**632.38**

Table 4.1 – The prediction error of the k-NN estimator (4.1) for k chosen from the Hold-out strategy (4.11) and generalized cross-validation (4.10) compared to the minimum discrepancy rule k^{τ} (4.7).

4.4.3 Results of the simulation experiments.

Table 4.1 displays the names of the data sets and the partitions made on the train and test samples. "Train" measures the number of samples for training and model selection of k (our stopping rule k^{τ}, the Hold-out method k_{HO}, and the generalized cross-validation k_{GCV}), whereas "Test" measures the number of samples taken out to make a prediction. The last three columns show the prediction error obtained when choosing k_{HO}, k_{GCV}, and k^{τ}, respectively.

According to the last three columns of Table 4.1, one can deduce that k^{τ} achieves comparable performance w.r.t. k_{HO} or k_{GCV}. In more detail, it performs better or (almost) equally on "Diabetes", "Wine Quality", and "California Houses Prices" data sets, while the performances on "Power Plants" and "Boston Houses Prices" are significantly worse. Remark that in our simulated experiments we estimated the value of σ^2, which can (partially) explain why there are data sets, on which k^{τ} performs worse than its competitors. To support additionally this argument, we move back to Fig. 4.5 (the bottom one), where one can see the value of $\sigma^2 \approx 3000$ that corresponds to an abrupt change in the behavior of the empirical risk. This point is detected by the estimator of the variance (4.22), and the prediction error of k^{τ} is equal to that of k_{HO}. Notice that Arlot and Bach [6] observed a similar presence of a jump around σ^2 for the so-called minimal penalty term, and they used this phenomenon in order to estimate σ^2 and plug in it into the final model selection procedure.

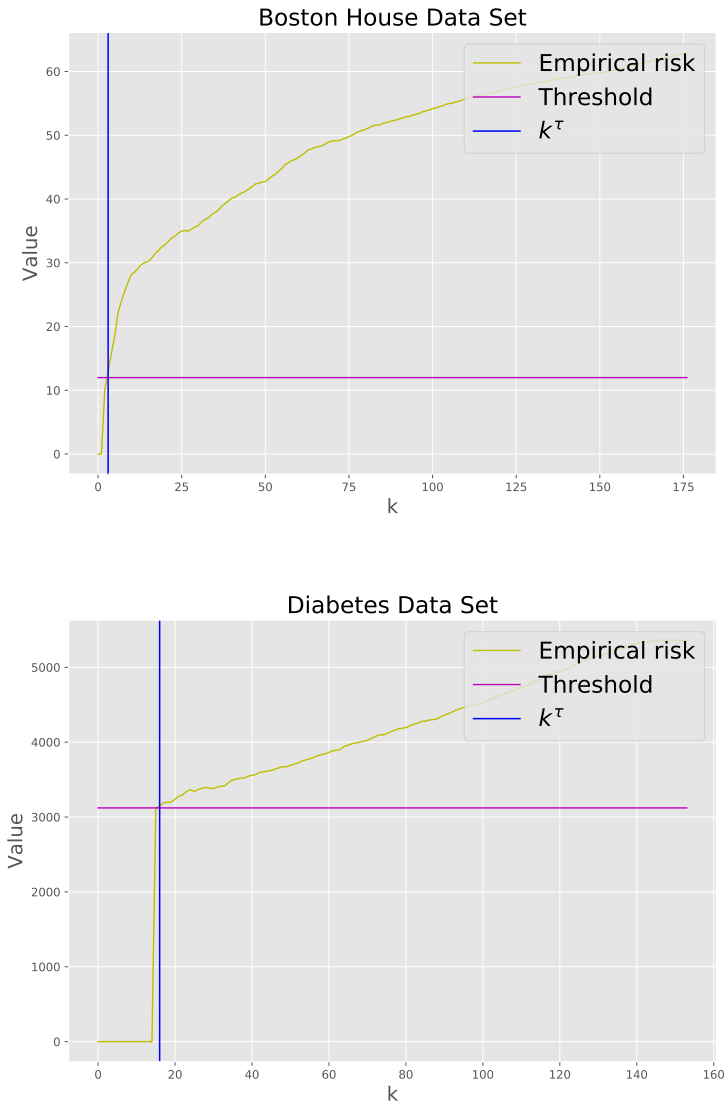


Figure 4.5 – Stopping the learning process based on the rule (4.7) applied to two data sets: "Boston House Prices" and "Diabetes". "Threshold" line corresponds to the estimated variance from Eq. (4.22).

CONCLUSIONS AND PERSPECTIVES

4.5 Summary of the thesis

In this thesis, we focused on constructing statistically optimal early stopping rules for several iterative learning algorithms. More precisely, we started our analysis with some spectral filter algorithms (gradient descent, ridge regression) in the framework of reproducing kernel Hilbert space (RKHS) and further expanded the explored ideas to tuning the parameter in linear estimators.

As our first contribution, we constructed a data-driven early stopping rule for gradient descent and iterative ridge regression in RKHS by means of the so-called minimum discrepancy principle, originally coming from the ill-posed inverse problem literature. The crucial quantity on which the rule is developed is the empirical risk of a functional estimator. It turned out that the original minimum discrepancy stopping rule provided a minimax optimal functional estimator only in the case of finite-rank reproducing kernels. If one considers infinite-rank kernels, the initial stopping rule has to be modified due to a large deviation of the empirical risk around its expectation. For this reason, we proposed using the so-called polynomial smoothing strategy that consists in proper weighting the empirical risk utilizing the eigenvalues of the normalized kernel matrix, which is itself a consequence of weighting the empirical norm. This new strategy, under some (mild) assumptions on the eigenvalues of the normalized kernel matrix, has been proved to achieve minimax optimality over a range of kernel classes, in particular, the one that corresponds to Sobolev spaces. The proof of the mentioned result involved careful analysis of the (smoothed) localized Rademacher complexities and their critical radii. We should emphasize that, to the best of our knowledge, the idea of weighting the empirical norm and connecting it with functional complexity measures such as the localized Rademacher complexities, is novel. Besides that, we established a clear connection between early stopping and kernel approximation with randomized sketches (projections). This connection may strengthen the intuition that there should be an equivalence between different statistical procedures (kernel approximation [14, 121], distributed learning [78, 127], and early stopping [12, 92]) aiming at reducing the computational complexity of a learning algorithm while preserving its optimality. Simulation results in Chapter 2 verified our theory.

The second contribution consists in extending the minimum discrepancy strategy to the task of tuning the parameter in linear estimators. In Chapter 3, we focused on the theoretical analysis of the k-NN regression estimator. We applied the aforementioned strategy for choosing k in order to lower the computational time of the selection procedure. In the end, it turned out that this choice provided a minimax optimal estimator, in particular, over the class of Lipschitz functions defined on a bounded domain. The main reason for the optimality was the fact that the minimax rate of the mentioned class

is quite slow due to the notorious "curse of dimensionality".

In Chapter 4, we were mainly interested in carrying out simulated experiments on artificial and real data sets. It turned out that the minimum discrepancy principle had comparable performance to other model selection procedures (the Hold-out method, generalized cross-validation, Mallows' C_p) for the task of choosing the parameter in linear estimators. However, the theoretical investigation of the performance of, for instance, the Nadaraya-Watson regression estimator should be done in the future to complete the work.

4.6 Perspectives

There are several possible directions to extend the results of this work.

— As it was already mentioned in the previous section, we are interested in the theoretical performance of the Nadaraya-Watson regressor. Apparently, it should be close to the one of the k-NN regression estimator. The main difficulty should come from the fact that, if A_h is the smoothing matrix of the Nadarya-Watson estimator, $\operatorname{tr}\left(A_h^\top A_h\right) \neq \operatorname{tr}\left(A_h\right)$. This fact implies that the expectation of the empirical risk minus the noise variance will not be equal to the difference between the bias and variance terms. Therefore, there should be another concentration result that deals with this problem.

Apart from that, the variable selection estimator was defined only for the "well-behaved" design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, meaning that $d \leq n$. On the other hand, there is plenty of practical applications in biology, medicine, and computer vision when the "high-dimensional" case $d > n$ is of interest. Thus, one should understand the behavior of the MDP rule in this setting.

It turned out (see, e.g., sumulations results on real datasets in Chapter 4) that the MDP strategy to choose the tuning parameter for linear estimators is close to the work [6], where Arlot and Bach explored the idea of minimal penalties. Therefore, we could borrow their strategy to build a plug-in estimator (with an estimated noise variance σ^2).

— We should emphasize that the early stopping rules in this work were estimating the famous bias-variance trade-off [70, Chapter 7]. However, recently, [20, 21] the bias-variance balancing paradigm was rethought by discovering some settings (exact fit to the data), for which a phenomenon of the "double descent" of the risk curve appeared. It would be interesting to understand if early stopping can work for these settings. The interested reader can look at a very recent paper [59] and references therein for another reexamination of the paradigm.

— As it was said in the conclusion section of Chapter 2, computing all eigenvalues of the kernel matrix is prohibitive in large-scale problems. Thus, some kernel approximation techniques [41, 96] could be helpful. Besides that, we are interested in extending the theoretical understanding of the stopping rules to the classification framework. That can be done by changing the square loss to the 0/1 loss or the log-loss.

— Another compelling research direction would be an investigation of the statistical theory of the minimum discrepancy principle stopping rule for the stochastic gradient descent algorithm. This could shed some light on the theoretical understanding of early stopping in (deep) artificial neural networks. A curious reader can take a look at the promising paper [97] to start with the statistical framework of the problem.

BIBLIOGRAPHY

[1] Hirotogu Akaike. « Information theory and an extension of the maximum likelihood principle ». In: *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[2] Hirotugu Akaike. « A new look at the statistical model identification ». In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723.

[3] Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. « A Continuous-Time View of Early Stopping for Least Squares ». In: *arXiv e-prints*, arXiv:1810.10082 (Oct. 2018), arXiv:1810.10082. arXiv: 1810.10082 [stat.ML].

[4] Tomas Angles et al. « NYTRO: When Subsampling Meets Early Stopping ». In: *arXiv e-prints*, arXiv:1510.05684 (Oct. 2015), arXiv:1510.05684. arXiv: 1510.05684 [stat.ML].

[5] Stephan W Anzengruber and Ronny Ramlau. « Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators ». In: *Inverse Problems* 26.2 (2009), p. 025001.

[6] Sylvain Arlot and Francis Bach. « Data-driven calibration of linear estimators with minimal penalties ». In: *arXiv e-prints*, arXiv:0909.1884 (Sept. 2009), arXiv:0909.1884. arXiv: 0909.1884 [math.ST].

[7] Sylvain Arlot and Francis R Bach. « Data-driven calibration of linear estimators with minimal penalties ». In: *Advances in Neural Information Processing Systems*. 2009, pp. 46–54.

[8] Sylvain Arlot, Alain Celisse, et al. « A survey of cross-validation procedures for model selection ». In: *Statistics surveys* 4 (2010), pp. 40–79.

[9] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. « A Kernel Multiple Change-point Algorithm via Model Selection. » In: *Journal of Machine Learning Research* 20.162 (2019), pp. 1–56.

[10] Sylvain Arlot and Pascal Massart. « Data-driven Calibration of Penalties for Least-Squares Regression. » In: *Journal of Machine learning research* 10.2 (2009).

[11] Nachman Aronszajn. « Theory of reproducing kernels ». In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[12] Yaroslav Averyanov and Alain Celisse. « Early stopping and polynomial smoothing in regression with reproducing kernels ». In: *arXiv preprint arXiv:2007.06827* (2020).

[13] Mona Azadkia. *Optimal choice of k for k-nearest neighbor regression*. 2019. arXiv: 1909.05495 [math.ST].

[14] Francis Bach. « Sharp analysis of low-rank kernel matrix approximations ». In: *Conference on Learning Theory*. 2013, pp. 185–209.

[15] Luca Baldassarre et al. « Multi-output learning via spectral filtering ». In: *Machine learning* 87.3 (2012), pp. 259–301.

[16] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. « Local rademacher complexities ». In: *The Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[17] Peter L Bartlett and Mikhail Traskin. « Adaboost is consistent ». In: *Journal of Machine Learning Research* 8.Oct (2007), pp. 2347–2368.

[18] Peter L Bartlett et al. « Benign overfitting in linear regression ». In: *Proceedings of the National Academy of Sciences* (2020).

[19] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. « On regularization algorithms in learning theory ». In: *Journal of complexity* 23.1 (2007), pp. 52–72.

[20] Mikhail Belkin, Daniel Hsu, and Ji Xu. « Two models of double descent for weak features ». In: *arXiv preprint arXiv:1903.07571* (2019).

[21] Mikhail Belkin et al. « Reconciling modern machine-learning practice and the classical bias–variance trade-off ». In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854.

[22] Jon Louis Bentley. « Multidimensional binary search trees used for associative searching ». In: *Communications of the ACM* 18.9 (1975), pp. 509–517.

[23] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[24] PK Bhattacharya and YP Mack. « Weak convergence of k-NN density and regression estimators with varying k and applications ». In: *The Annals of Statistics* (1987), pp. 976–994.

[25] Gérard Biau, Frédéric Cérou, and Arnaud Guyader. « Rates of convergence of the functional k-nearest neighbor estimate ». In: *IEEE Transactions on Information Theory* 56.4 (2010), pp. 2034–2040.

[26] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Vol. 246. Springer, 2015.

[27] Lucien Birgé and Pascal Massart. « Minimal penalties for Gaussian model selection ». In: *Probability theory and related fields* 138.1-2 (2007), pp. 33–73.

[28] Gilles Blanchard, Marc Hoffmann, and Markus Reiß. « Optimal adaptation for early stopping in statistical inverse problems ». In: *arXiv preprint arXiv:1606.07702* (2016).

[29] Gilles Blanchard, Marc Hoffmann, and Markus Reiß. « Optimal adaptation for early stopping in statistical inverse problems ». In: *SIAM/ASA Journal on Uncertainty Quantification* 6.3 (2018), pp. 1043–1075.

[30] Gilles Blanchard, Marc Hoffmann, Markus Reiß, et al. « Early stopping for statistical inverse problems via truncated SVD estimation ». In: *Electronic Journal of Statistics* 12.2 (2018), pp. 3204–3231.

[31] Gilles Blanchard and Nicole Krämer. « Convergence rates of kernel conjugate gradient for random design regression ». In: *Analysis and Applications* 14.06 (2016), pp. 763–794.

[32] Gilles Blanchard and Nicole Krämer. « Optimal learning rates for kernel conjugate gradient regression ». In: *Advances in Neural Information Processing Systems*. 2010, pp. 226–234.

[33] Gilles Blanchard and Peter Mathé. « Conjugate gradient regularization under general smoothness and noise assumptions ». In: *Journal of Inverse and Ill-posed Problems* 18.6 (2010), pp. 701–726.

[34] Gilles Blanchard and Peter Mathé. « Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration ». In: *Inverse problems* 28.11 (2012), p. 115011.

[35] Léon Bottou. « Large-scale machine learning with stochastic gradient descent ». In: *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[36] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[37] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

[39] Peter Bühlmann and Bin Yu. « Boosting with the L 2 loss: regression and classification ». In: *Journal of the American Statistical Association* 98.462 (2003), pp. 324–339.

[40] Florentina Bunea, Alexandre B Tsybakov, Marten H Wegkamp, et al. « Aggregation for Gaussian regression ». In: *The Annals of Statistics* 35.4 (2007), pp. 1674–1697.

[41] Raffaello Camoriano et al. « Nytro: When subsampling meets early stopping ». In: *Artificial Intelligence and Statistics*. 2016, pp. 1403–1411.

[42] Y Cao and Y Golubev. « On oracle inequalities related to smoothing splines ». In: *Mathematical Methods of Statistics* 15.4 (2006), pp. 398–414.

[43] Andrea Caponnetto. « Optimal Rates for Regularization Operators in Learning Theory ». In: (Sept. 2006).

[44] Andrea Caponnetto and Ernesto De Vito. « Optimal Rates for the Regularized Least-Squares Algorithm ». In: *Foundations of Computational Mathematics* 7.3 (2007), pp. 331–368.

[45] Andrea Caponnetto and Yuan Yao. « Cross-validation based adaptation for regularization operators in learning theory ». In: *Analysis and Applications* 8.02 (2010), pp. 161–183.

[46] Rich Caruana, Steve Lawrence, and C Lee Giles. « Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping ». In: *Advances in neural information processing systems*. 2001, pp. 402–408.

[47] Laurent Cavalier et al. « Oracle inequalities for inverse problems ». In: *The Annals of Statistics* 30.3 (2002), pp. 843–874.

[48] Alain Celisse and Tristan Mary-Huard. « Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier ». In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2373–2426.

[49] Alain Celisse and Martin Wahl. « Analyzing the discrepancy principle for kernelized spectral filter learning algorithms ». In: *arXiv preprint arXiv:2004.08436* (2020).

[50] Alain Celisse and Martin Wahl. « Analyzing the discrepancy principle for kernelized spectral filter learning algorithms ». In: *arXiv preprint arXiv:2004.08436* (2020).

[51] E Chernousova and Yu Golubev. « Spectral cut-off regularizations for ill-posed linear models ». In: *Mathematical Methods of Statistics* 23.2 (2014), pp. 116–131.

[52] G Collomb et al. « Estimation de la regression par la méthode des k points les plus proches: propriétés de convergence ponctuelle ». In: (1979).

[53] Peter Craven and Grace Wahba. « Smoothing noisy data with spline functions ». In: *Numerische mathematik* 31.4 (1978), pp. 377–403.

[54] Felipe Cucker and Steve Smale. « On the mathematical foundations of learning ». In: *Bulletin of the American mathematical society* 39.1 (2002), pp. 1–49.

[55] Luc Devroye. « The uniform convergence of nearest neighbor regression function estimators and their application in optimization ». In: *IEEE Transactions on Information Theory* 24.2 (1978), pp. 142–151.

[56] Luc Devroye et al. « On the almost everywhere convergence of nonparametric regression function estimates ». In: *The Annals of Statistics* 9.6 (1981), pp. 1310–1319.

[57] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[58] Nigel Duffy and David Helmbold. « Boosting methods for regression ». In: *Machine Learning* 47.2-3 (2002), pp. 153–200.

[59] Raaz Dwivedi et al. « Revisiting complexity and the bias-variance tradeoff ». In: *arXiv preprint arXiv:2006.10189* (2020).

[60] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*. Vol. 375. Springer Science & Business Media, 1996.

[61] Jerome Friedman and Bogdan E Popescu. « Gradient directed regularization ». In: *Unpublished manuscript, http://www-stat. stanford. edu/˜ jhf/ftp/pathlite. pdf* (2004).

[62] Thomas Gartner. *Kernels for structured data*. Vol. 72. World Scientific, 2008.

[63] Seymour Geisser. « The predictive sample reuse method with applications ». In: *Journal of the American statistical Association* 70.350 (1975), pp. 320–328.

[64] L Lo Gerfo et al. « Spectral algorithms for supervised learning ». In: *Neural Computation* 20.7 (2008), pp. 1873–1897.

[65] Ian Goodfellow et al. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.

[66] Chong Gu. *Smoothing spline ANOVA models*. Vol. 297. Springer Science & Business Media, 2013.

[67] László Györfi et al. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[68] Peter Hall and JS Marron. « On variance estimation in nonparametric regression ». In: *Biometrika* 77.2 (1990), pp. 415–419.

[69] Per Christian Hansen. *Discrete inverse problems: insight and algorithms*. Vol. 7. Siam, 2010.

[70] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[71] Wenxin Jiang et al. « Process consistency for adaboost ». In: *The Annals of Statistics* 32.1 (2004), pp. 13–29.

[72] Vladimir Koltchinskii et al. « Local Rademacher complexities and oracle inequalities in risk minimization ». In: *The Annals of Statistics* 34.6 (2006), pp. 2593–2656.

[73] Samory Kpotufe. « k-NN regression adapts to local intrinsic dimension ». In: *Advances in neural information processing systems*. 2011, pp. 729–737.

[74] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. « Suprema of chaos processes and the restricted isometry property ». In: *Communications on Pure and Applied Mathematics* 67.11 (2014), pp. 1877–1904.

[75] Louis Landweber. « An iteration formula for Fredholm integral equations of the first kind ». In: *American journal of mathematics* 73.3 (1951), pp. 615–624.

[76] Ker-Chau Li. « Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set ». In: *The Annals of Statistics* (1987), pp. 958–975.

[77] Ker-Chau Li et al. « Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing ». In: *The Annals of Statistics* 14.3 (1986), pp. 1101–1112.

[78] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. « Distributed learning with regularized least squares ». In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 3202–3232.

[79] Colin L Mallows. « Some comments on Cp ». In: *Technometrics* 42.1 (2000), pp. 87–94.

[80] Pascal Massart. *Concentration inequalities and model selection*. Vol. 6. Springer, 2007.

[81] Peter Mathé and Sergei V Pereverzev. « Geometry of linear ill-posed problems in variable Hilbert scales ». In: *Inverse problems* 19.3 (2003), p. 789.

[82] Shahar Mendelson. « Geometric parameters of kernel machines ». In: *International Conference on Computational Learning Theory*. Springer. 2002, pp. 29–43.

[83] Shahar Mendelson. « On the performance of kernel classes ». In: *Journal of Machine Learning Research* 4.Oct (2003), pp. 759–771.

[84] Vladimir Alekseevich Morozov. « On the solution of functional equations by the method of regularization ». In: *Doklady Akademii Nauk*. Vol. 167. 3. Russian Academy of Sciences. 1966, pp. 510–512.

[85] Elizbar A Nadaraya. « On estimating regression ». In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142.

[86] Arkadi Nemirovski. « Topics in non-parametric ». In: *Ecole d'Eté de Probabilités de Saint-Flour* 28 (2000), p. 85.

[87] Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

[88] R Kelley Pace and Ronald Barry. « Sparse spatial autoregressions ». In: *Statistics & Probability Letters* 33.3 (1997), pp. 291–297.

[89] Roger Penrose. « A generalized inverse for matrices ». In: *Mathematical proceedings of the Cambridge philosophical society*. Vol. 51. 3. Cambridge University Press. 1955, pp. 406–413.

[90] Mark Semenovich Pinsker. « Optimal filtering of square-integrable signals in Gaussian noise ». In: *Problemy Peredachi Informatsii* 16.2 (1980), pp. 52–68.

[91] Lutz Prechelt. « Early stopping-but when? » In: *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69.

[92] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. « Early stopping and non-parametric regression: an optimal data-dependent stopping rule ». In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 335–366.

[93] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. « Minimax-optimal rates for sparse additive models over kernel classes via convex programming ». In: *Journal of Machine Learning Research* 13.Feb (2012), pp. 389–427.

[94] John Rice et al. « Bandwidth choice for nonparametric regression ». In: *The Annals of Statistics* 12.4 (1984), pp. 1215–1230.

[95] Mark Rudelson, Roman Vershynin, et al. « Hanson-Wright inequality and sub-gaussian concentration ». In: *Electronic Communications in Probability* 18 (2013).

[96] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. « Less is more: Nyström computational regularization ». In: *Advances in Neural Information Processing Systems*. 2015, pp. 1657–1665.

[97] Johannes Schmidt-Hieber et al. « Nonparametric regression using deep neural networks with ReLU activation function ». In: *Annals of Statistics* 48.4 (2020), pp. 1875–1897.

[98] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[99] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[100] Gideon Schwarz et al. « Estimating the dimension of a model ». In: *The annals of statistics* 6.2 (1978), pp. 461–464.

[101] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[102] Ritei Shibata. « An optimal selection of regression variables ». In: *Biometrika* 68.1 (1981), pp. 45–54.

[103] Ritei Shibata. « Asymptotically efficient selection of the order of the model for estimating parameters of a linear process ». In: *The annals of statistics* (1980), pp. 147–164.

[104] Steve Smale and Ding-Xuan Zhou. « Learning theory estimates via integral operators and their approximations ». In: *Constructive approximation* 26.2 (2007), pp. 153–172.

[105] Bernhard Stankewitz. « Smoothed residual stopping for statistical inverse problems via truncated SVD estimation ». In: *arXiv preprint arXiv:1909.13702* (2019).

[106] Charles J Stone et al. « Additive regression and other nonparametric models ». In: *The annals of Statistics* 13.2 (1985), pp. 689–705.

[107] Masashi Sugiyama and Hidemitsu Ogawa. « Theoretical and experimental evaluation of the subspace information criterion ». In: *Machine Learning* 48.1-3 (2002), pp. 25–50.

[108] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[109] Alexandre B Tsybakov. « Optimal rates of aggregation ». In: *Learning theory and kernel machines*. Springer, 2003, pp. 303–313.

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

[111] Grace Wahba. « Practical approximate solutions to linear operator equations when the data are noisy ». In: *SIAM Journal on Numerical Analysis* 14.4 (1977), pp. 651–667.

[112] Grace Wahba. *Spline models for observational data*. Vol. 59. Siam, 1990.

[113] Grace Wahba. « Three topics in ill-posed problems ». In: *Inverse and ill-posed problems*. Elsevier, 1987, pp. 37–51.

[114] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[115] Martin J Wainwright. « Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting ». In: *IEEE Transactions on Information Theory* 55.12 (2009), pp. 5728–5741.

[116] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[117] Marten Wegkamp et al. « Model selection in nonparametric regression ». In: *The Annals of Statistics* 31.1 (2003), pp. 252–273.

[118] Yuting Wei, Fanny Yang, and Martin J Wainwright. « Early stopping for kernel boosting algorithms: A general analysis with localized complexities ». In: *Advances in Neural Information Processing Systems*. 2017, pp. 6065–6075.

[119] Yuhong Yang. « Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation ». In: *Biometrika* 92.4 (2005), pp. 937–950.

[120] Yuhong Yang. « Model selection for nonparametric regression ». In: *Statistica Sinica* (1999), pp. 475–499.

[121] Yun Yang, Mert Pilanci, Martin J Wainwright, et al. « Randomized sketches for kernels: Fast and optimal nonparametric regression ». In: *The Annals of Statistics* 45.3 (2017), pp. 991–1023.

[122] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. « On Early Stopping in Gradient Descent Learning ». In: *Constructive Approximation* 26.2 (Aug. 2007), pp. 289–315. ISSN: 1432-0940. DOI: 10.1007/s00365-006-0663-2. URL: https://doi.org/10.1007/s00365-006-0663-2.

[123] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. « On early stopping in gradient descent learning ». In: *Constructive Approximation* 26.2 (2007), pp. 289–315.

[124] Jianguo Zhang et al. « Local features and kernels for classification of texture and object categories: A comprehensive study ». In: *International journal of computer vision* 73.2 (2007), pp. 213–238.

[125] Tong Zhang. « Learning bounds for kernel regression using effective data dimensionality ». In: *Neural Computation* 17.9 (2005), pp. 2077–2098.

[126] Tong Zhang, Bin Yu, et al. « Boosting with early stopping: Convergence and consistency ». In: *The Annals of Statistics* 33.4 (2005), pp. 1538–1579.

[127] Yuchen Zhang, John Duchi, and Martin Wainwright. « Divide and conquer kernel ridge regression ». In: *Conference on learning theory*. 2013, pp. 592–617.

[128] Puning Zhao and Lifeng Lai. « Minimax Rate Optimal Adaptive Nearest Neighbor Classification and Regression ». In: *arXiv preprint arXiv:1910.10513* (2019).