High-capacity hypothesis spaces in modern statistical learning

Luca Wellmeier

April 21st, 2023

Contents

Introduction

Modern deep learning practice has put statistical learning theory to new challenges. Highly overparameterized neural networks perform outstandingly well in an enormous variety of tasks, while seemingly violating one of the most fundamental ideas of classical statistical learning: the bias-variance tradeoff. With this concept in mind, an estimator complex enough perfectly fit a huge training set should be regularized in order to generalize well to unseen inputs, as one needs to expect overfitting otherwise. However, it is very common practice to train neural nets down to vanishingly low sample errors (i.e. even the noisy or plain-wrong examples!) and still achieve state-of-the-art performance.

The goal of this thesis is provide an overview over the current state of research. In particular, kernel ridge (and ridgeless) regression is developed in much detail as these classical estimators often behave similar to neural nets due to their high-capacity hypothesis spaces and the ability to fit any dataset. Emphasis is put on the role of capacity controls and implicit and explicit regularization techniques. Finally, a simple but powerful study model around kernels on tori is proposed and shown to exhibit phenomena comparable to real problems but with very low computational costs and the ability of full analytic treatment.

All code for experiments and the source of this document itself can be found on GitHub[1].

[1] `https://github.com/lucw0/mscthesis-code`

1 Statistical learning theory

Following Cucker and Smale 2002, Vapnik 2000 (and partially Steinwart and Christmann 2008, Chapter 6), we start off by developing a basic statistical framework, powerful enough to demonstrate the frontiers of modern statistical learning, yet tractable enough to be treated by common analytical tools. First, we develop the probabilistic product space model which is standard in the literature and after that we let the statistics enter by discussing how to learn from data and discuss how to generalize from those limited observations to the true problem. We discuss an example of a simple algorithm to illustrate structural risk minimization. Finally, we dive into the problems with the classical statistical learning theory and discuss modern frontiers.

1.1 Probabilistic model

Throughout, let X be a topological space equipped with its Borel σ-algebra and $Y = \mathbb{K}$. "Measurable" will always refer to Borel measurable with respect to the topological space in question and we denote by $M(X;Y)$ the vector space of all Borel measurable functions $X \to Y$. Consider a Borel probability measure P on $Z \coloneqq X \times Y$ and marginal distribution $\mu = X_*P$ of X. For simplicity of notation, we will also use the symbols X, Y, Z as random variables: $X\colon Z \to X$ and $Y\colon Z \to Y$ are the projections and $Z = (X, Y)$ is their cartesian product. There will never be confusion between space and random variable through context or language.

The variable X models the covariate input. For now we only assume that it lives in some topological space but for our analyses we will later need to add more restrictions. The random input X causes the random response Y. The complexity of the relationship is captured by the measure P: the aim of statistical learning is to understand its behavior through finite samples.

While the response Y is random, we generally assume that given information on X has strong influence on it. In fact, we will assume that we have a "mostly" functional relationship between the variables of the form

$$Y = f^*(X) + \varepsilon, \tag{1.1}$$

where the noise ε is a "controllable" scalar random variable. Instead of modeling the randomness directly, we are looking for functions $f\colon X \to Y$ that are as close as possible to f^* according to a measure of accuracy.

Let us make this more precise.

Definition 1.1. *For a measurable function $f\colon X \to Y$ we quantify the random difference $f(X) - Y$ by the risk $R(f) = R_P(f)$ w.r.t. the squared loss:*

$$R_P(f) \coloneqq \mathbb{E}\left[|f(X) - Y|^2\right] = \int_Z |f(X) - Y|^2 \, dP. \tag{1.2}$$

If such a function f is proposed to solve or provide an approximate solution to the estimation problem $\inf_f R(f)$, we often refer to it as an *estimator*. If the variable is sufficiently regular, there is an essentially unique perfect estimator $f^* = f_P^*$ solving the above problem:

Proposition 1.2 (Regression function)**.** *If the variable Y has finite variance, i.e. $Y \in L^2(Z, P)$, then there exists a solution $f^* = f_P^*\colon X \to Y$ to the problem*

$$\min_{f \in M(X;Y)} R_P(f), \tag{1.3}$$

and any two such solutions are equal μ-a.e. on X. Moreover, we can write

$$Y = f^*(X) + \varepsilon, \tag{1.4}$$

where ε is a random variable with $\mathbb{E}[\varepsilon] = 0$ and $\sigma^2 \coloneqq \operatorname{Var}[\varepsilon] = \mathbb{E}[\varepsilon^2] < \infty$.

Proof. The first statement is a mere translation of the L^2-theory of conditional expectations that are variance minimizers (see for instance Dudley 2002 or Billingsley 1995). The sought object is exactly the random variable $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid \sigma(X)] \in L^2(Z, \sigma(X), P)$. Since, it is $\sigma(X)$-measurable, we can prove that it is almost deterministic given the concrete event $\{X = x\}$.

Write $Y = \lim_n Y_n$ as a point-wise limit of simple functions $Y_n = \sum_{i=1}^{M_n} y_{n,i} 1_{Z_{n,i}}$ with $Z_{n,i} \in \sigma(X)$. We can rewrite the latter sets as $Z_{n,i} = X^{-1}(X_{n,i})$ with Borel sets $X_{n,i} \subseteq X$. Then, $1_{Z_{n,i}} = 1_{X_{n,i}} \circ X$ so that $Y_n = f_n(X)$ if we set $f_n = \sum_{i=1}^{M_n} y_{n,i} 1_{X_{n,i}}$. Consequently,

$$Y = \limsup_n Y_n = \limsup_n f_n(X) =: f^*(X) \tag{1.5}$$

as a point-wise limit and we can write $f^*(X) = \mathbb{E}[Y \mid X]$ or $f^*(x) = \mathbb{E}[Y \mid X = x]$ which holds μ-a.s.

Finally, let $\varepsilon \coloneqq Y - f^*(X)$. Then, by the standard properties of conditional expectations we find

$$\mathbb{E}[\varepsilon] = \mathbb{E}[\mathbb{E}[\varepsilon \mid X]] = \mathbb{E}[\mathbb{E}[Y \mid X] - \mathbb{E}[f^*(X) \mid X]] = 0, \tag{1.6}$$

and it has finite variance as $\sqrt{\operatorname{Var}[\varepsilon]} = \sqrt{\mathbb{E}[|\varepsilon|^2]} = \|Y - f^*(X)\|_{L^2}$ is the norm of a difference of two L^2 functions. □

From now on we will reference an arbitrary but fixed choice of the regression function f^*.

Since $f^*(x) = \mathbb{E}[Y \mid X = x]$ is the optimal *least squares* estimator, we can express the risk of another proposed estimator by comparing it to the regression function instead of the random variable Y.

Proposition 1.3 (Risk)**.** *Let $f\colon X \to Y$ be measurable and suppose ε is independent of X. Then,*

$$R(f) = \int_X |f - f^*|^2 \, d\mu + \sigma^2. \tag{1.7}$$

The independence assumption is standard in the statistical learning literature as it seems natural (think about imprecisions in measurements) and it allows for nice decompositions of the risk as we will see later.

Proof. Let $z = (x, y)$ and recall that the variables X and Y are mere projections with $X(z) = x$ and $Y(z) = y$. Since $\varepsilon(z) = y - f^*(x)$, we have

$$\begin{aligned}
|f(x) - y|^2 &= \big|(f(x) - f^*(x)) + (f^*(x) - y)\big|^2 && (1.8)\\
&= \big|(f(x) - f^*(x)) - \varepsilon\big|^2 && (1.9)\\
&= \big((f(x) - f^*(x)) - \varepsilon\big)\overline{((f(x) - f^*(x)) - \varepsilon)} && (1.10)\\
&= |f(x) - f^*(x)|^2 + |\varepsilon|^2 + 2\Re[\varepsilon(f(x) - f^*(x))]. && (1.11)
\end{aligned}$$

Now, for any $x \in X$ we let $P(\cdot \mid x)$ be the conditional distribution of Y given the event $X = x$. These are random probability measures that admit a version of the Fubini theorem (we refer to Dudley 2002 for a review) allowing to conceptually split dP into $dP(\cdot \mid x)$ and μ:

$$\begin{aligned}
R(f) &= \int_Z |f(x) - y|^2 \, dP(z) && (1.12)\\
&= \int_X \int_Y |f(x) - y|^2 \, P(dy \mid x) d\mu(x) && (1.13)\\
&= \int_X \int_Y \Big(|f(x) - f^*(x)|^2 + |\varepsilon|^2 + 2\Re[\varepsilon(f(x) - f^*(x))]\Big) \, P(dy \mid x) d\mu(x) && (1.14)\\
&= \int_X |f(x) - f^*(x)|^2 \, d\mu(x) + \sigma^2 + 2\Re\big[\underbrace{\mathbb{E}[\varepsilon]}_{=0} \mathbb{E}[f(X) - f^*(X)]\big] && (1.15)\\
&= \int_X |f - f^*|^2 \, d\mu + \sigma^2 && (1.16)
\end{aligned}$$

where the last equalities follows from the independence and zero-mean property of the noise ε. □

1.2 Empirical risk minimization

There is a big, practical problem with the model: the underlying distribution P of $Z = (X, Y)$ is usually unknown or untractable making even the evaluation of the risk R_P impossible (let alone its minimization). We need a way to circumvent this. Statistical learning is all about estimation of the complicated relationship between the variables X and Y by *learning* it from a finite sample.

Definition 1.4. *Let $Z_1 = (X_1, Y_1), Z_2 = (X_2, Y_2), \dots$ be IID copies of Z. The sequences $\hat{Z}^n = (Z_i)_{i=1}^n$ are called samples of Z and we denote by $\hat{P}^n = \sum_{i=1}^n \delta_{Z_i}$ the (random) empirical distribution and by $\hat{\mu}^n$ the empirical marginal.*

Any quantitity that depends on a sample will be carrying a hat. The fundamental object is the following: if a suitable sample of the variable Z is obtained, we can assess the risk of an estimator in an approximate fashion.

Definition 1.5. *If $f \in M(X;Y)$ and $\hat{Z}^n$ is a sample we can define its empirical risk as the risk with respect to the empirical distribution $\hat{P}$ as*

$$\hat{R}(f) = \hat{R}^n(f) \coloneqq R_{\hat{P}^n}(f) = \int_Z |f(X) - Y|^2 \, d\hat{P}^n = \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2. \tag{1.17}$$

This quantity $\hat{R}(f)$ is completely computable. It gives rise to a principled way of finding learning methods: instead of minimizing R_P directly, we instead attempt to minimize $\hat{R}$, which we can evaluate explicitly. This process is called *empirical risk minimization* (ERM).

The question is: is $\hat{R}$ actually a good minimization target? Is the empirical risk of an estimator close to the true risk? The ability of a learning method to learn the true distribution while it only ever sees a finite sample is called *generalization*. By the IID assumption $\mathbb{E}[\hat{R}^n(f)] = R(f)$, so that the strong law of large numbers (SLLN) yields first postive evidence for a fixed estimator f: almost surely $\lim_n \hat{R}^n(f) = R(f)$. We can quantify the rate of convergence:

Proposition 1.6 (Defect bound)**.** *Let $f \in M(X;Y)$ be an estimator and assume that there is some $M > 0$ such that a.s. $|f(X) - Y| \leq M$. Then, for any $\varepsilon > 0$*

$$|\hat{R}^n(f) - R(f)| \leq \varepsilon \tag{1.18}$$

with probability (on the draw of $\hat{Z}^n$ according to P^n) at least

$$1 - 2\exp\left(-\frac{n\varepsilon^2}{2(\mathrm{Var}[f(X) - Y] + M^2\varepsilon/3)}\right). \tag{1.19}$$

Proof. This is a simple application of Bernstein's concentration inequality for bounded variables. See for instance Cucker and Smale 2002, Theorem A, or Vershynin 2018, Theorem 2.8.4. □

The RHS of the bound is the confidence that the *defect* is small. However, while it becomes exponentially better in n, the whole bound is applicable only to a particular choice of f. This is problematic: it would require us to somehow reiterate the bound for any proposed estimator and any sample size (keeping in mind that the estimator depends substantially on the sample) rendering this result useless. Instead, we would like a uniform bound that holds for the whole range of possible estimators that a particular ERM-based learning method can produce.

Such a result is pointless on an overly complicated space like $M(X;Y)$: no learning machine would ever have the full class of measurable functions as range. Thus, we restrict ourselves to wisely chosen subspaces with more structure. These are usually referred to as *hypothesis spaces*, and usually denoted by $\mathcal{H}$.

The following is an illustrative example of a uniform risk bound.

Proposition 1.7 (Uniform defect bound)**.** *Let X be a compact set and $\mathcal{H}$ a compact subset of $C(X)$ endowed with the uniform convergence topology. If there is an $M > 0$ such that for all $f \in \mathcal{H}$ $|f(X) - Y| \leq M$ holds P-almost surely, then for any $\varepsilon > 0$*

$$\sup_{f \in H} |\hat{R}(f) - R(f)| \leq \varepsilon \tag{1.20}$$

with probability (on the draw of $\hat{Z}^n$ according to P^n) at least

$$1 - 2C_1 \exp\left(-\frac{n\varepsilon^2}{4(C_2 + M^2\varepsilon/3)}\right), \tag{1.21}$$

where $C_1 = C_1(H, \varepsilon, M)$ is the minimum number of balls of radius $\varepsilon/(8M)$ needed to cover H, and $C_2 = C_2(H) \coloneqq \sup_{f \in H} \operatorname{Var}[f(X) - Y]$.

Proof. Skipped. Can be found in Cucker and Smale 2002, Theorem B. □

Observe that the confidence in this uniform bound is essentially the same up to the expected change of taking the supremum of the variances over all functions and the new appearence of the covering number. The latter will scale in an exponential fashion with the "volume" of the hypothesis space.

This bound also yields first evidence of a tradeoff that comes with the size of the chosen hypothesis space. First, if we shrink the hypothesis space, then covering number and the variance term will shrink as well, making the confidence stronger. Simultaneously, however, a smaller hypothesis space will increase the defect between empirical and true risk, thus increasing ε. Summarizing shrinking the hypothesis space will increase the confidence of a worse bound. We are faced with an instance of a *bias-variance tradeoff*. Bias here refers to the capacity of the hypothesis space (low bias generally means to have a complicated high-capacity hypothesis space) and variance refers to the lack of confidence. Let us make this more precise.

1.3 Bias-variance tradeoff and structural risk minimization

We study components of the risk that illustrate this tradeoff. If an estimator is obtained only through a finite sample (e.g. through ERM) we denote it by $\hat{f}_n \coloneqq f(\cdot; \hat{Z}^n)$. They are random variables encoding the underlying learning method (the n is not mentioned explicitly). Recall the model $Y = f^*(X) + \varepsilon$.

Proposition 1.8 (Expected risk)**.** *Let $\hat{f}$ be an estimator learned from (or "dependent only on") the random sample of size n. If the noise ε is independent of X, then we have the decomposition*

$$\mathbb{E}_{\hat{Z}^n \sim P^n}[R(\hat{f})] = \mathbf{B}^2 + \mathbf{V} + \sigma^2, \tag{1.22}$$

where

$$\mathbf{B}^2 = \mathbb{E}_{X \sim \mu}\left[\left|\mathbb{E}_{\hat{Z}^n \sim P^n}[\hat{f}(X)] - f^*(X)\right|^2\right], \tag{1.23}$$

$$\mathbf{V} = \mathbb{E}_{X \sim \mu, \hat{Z}^n \sim P^n}\left[\left|\hat{f}(X) - \mathbb{E}_{\hat{Z}^n \sim P^n}[\hat{f}(X)]\right|^2\right]. \tag{1.24}$$

Here, the subscript in the expectation means that we integrate only over the indicated variables while we condition on the remaining ones.

Proof. Starting from proposition 1.3, we write

$$|f^*(X) - \hat{f}(X)|^2 = |(f^*(X) - \mathbb{E}_{\hat{Z}^n}[\hat{f}(X)]) + (\mathbb{E}_{\hat{Z}^n}[\hat{f}(X)] - \hat{f}(X))|^2. \tag{1.25}$$

Then we write out the terms of the square and obtain (after taking expectations) exactly $\mathbf{B}^2$ and $\mathbf{V}$ plus the following cross-term:

$$2\Re\big[(\hat{f}(X) - \mathbb{E}_{\hat{Z}^n}[\hat{f}(X)])\overline{(\mathbb{E}_{\hat{Z}^n}[\hat{f}(X)] - f^*(X))}\big]. \tag{1.26}$$

After multiplying these out we see that taking expection in X makes the cross-term vanish. □

The term $\mathbf{B}^2$ is called the *bias*. It measures pointwise how far the expected estimator will be from the value of the regression function. Note that it is highly dependent on the (effective) hypothesis space that the underlying learning method has used. The larger this space, the lower will be the bias. *Underfitting* refers to large bias paired with low variance.

The term $\mathbf{V}$ is called the *variance*. It quantifies how much the estimators oscillate around their mean when the concrete sample is changing. Note that it is a true variance. The smaller the effective space, the more deterministic the output of the learning method, the lower will be the variance. *Overfitting* refers to having high variance but low bias.

This classical belief of the two risk components working against each other with changing capacity of the hypothesis space is known as the mentioned *bias-variance tradeoff* and lead to an extension of ERM as design principle for learning methods: *structural risk minimization* (SRM). Let $\Omega \subset \mathbb{R}$ be a set that parameterizes a family of learning methods $\hat{f}_\omega$ together with their hypothesis spaces $\mathcal{H}_\omega$ such that if $\omega_1 \leq \omega_2$, then $\mathcal{H}_{\omega_1} \subseteq \mathcal{H}_{\omega_2}$. The tradeoff suggests that there should be a "sweet spot" ω^* that corresponds to the right capacity given the data, i.e. the one that minimizes the expected risk $\omega \mapsto \mathbb{E}[R(\hat{f}_\omega)]$.

Designing an algorithm following the SRM principle practically means to expose *hyper-parameters* that control the capacity of the effective hypothesis space of the method. They can then be used to perform *model selection* (which shall not be part of this work). There are many different ways to achieve this control. Probably, the most natural examples are found in *parametric methods*, where a parameter space is used to effectively characterize the hypothesis space and the learning is performed with an algorithm that acts on the parameter space instead of the function space. Neural nets with hyper-parameters per-layer-width, depth, activation and so on with weights trained via a gradient-based method belong to this class. However, many interesting methods like k-nearest neighbors or kernel methods (including the protagonist of this work: kernel ridge(less) regression) are *non-parametric*. Here, other, less obvious ways of capacity control are required. Throughout, we will be interested in the classical norm regularization method which we introduce now.

Definition 1.9. *Let $\mathcal{H}$ be a normed hypothesis space and $\alpha \geq 0$. The regularized empirical risk of $f \in \mathcal{H}$ is defined as*

$$\hat{R}_\alpha(f) = \hat{R}(f) + \frac{\alpha}{n}\|f\|_{\mathcal{H}}^2. \tag{1.27}$$

Note that we recover $\hat{R}(f) = \hat{R}_0(f)$ and we will use this alternative notation for unified presentation of results.

Minimizing $\hat{R}_\alpha$ instead of $\hat{R}$, indeed, implements a capacity control into the ERM principle. The norm-penalty term drives output estimtors to be "small" in terms of the notion of distance of their respective hypothesis space. This has manifold interpretations: in case of parameterized methods, common choices include the 1-norm or 2-norm in parameter space, resulting in *sparsity* and *weight decay*, respectively. If the penalty is applied to function spaces, one often sees L^p norms, Sobolev norms or total variation. In any case, α offers the capacity control. Note that, while the hypothesis space apparently doesn't change, the effective reach of the learning method is reduced with large α.

1.4 Example: least squares polynomial regression

Next, we will see simple examples of structural risk minimization in action, both as direct hypothesis space parameterization and via norm penalty.

Let $X = [-1, 1]$, $Y = \mathbb{R}$ and let $f^* \colon X \to Y$ be some smooth function. Choose the model distribution P so that μ is the half Lebesgue measure on X (i.e. a uniform distribution) and recall that the final model $(X, Y) = Z \sim P$ is such that $Y = f^*(X) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ is Gaussian noise indepdent of X.

For an integer $p \geq 0$ let $\mathcal{H}_p$ be the space of polynomial functions on X of degree less or equal p and consider the natural parametrization

$$\mathbb{R}^{p+1} \to \mathcal{H}_p, \quad \theta \mapsto f_\theta = \langle \theta, \Phi_p(\cdot) \rangle = \theta^T \Phi_p(\cdot) \tag{1.28}$$

with $\Phi_p(x) \coloneqq [x^0, \dots, x^p]^T \in \mathbb{R}^p$. This is a vector space isomorphism allowing us to use the induced norm $\|f_\theta\|_{\mathcal{H}_p} \coloneqq \|\theta\|_{\mathbb{R}^{p+1}}$.

Take a sample $\hat{Z}^n = \{(X_i, Y_i)\}_{i=1}^n$ and write out the regularized empirical risk into matrix form using $\hat{\Phi}_p = [\Phi_p(X_1) \mid \cdots \mid \Phi_p(X_n)]^T \in \mathbb{R}^{n \times p}$ and $\hat{Y} = [Y_1, \dots Y_n]^T \in \mathbb{R}^n$:

$$n\hat{R}_\alpha(f_\theta) = \sum_{i=1}^n (f_\theta(X_i) - Y_i)^2 + \alpha\|f_\theta\|^2 \tag{1.29}$$

$$= \sum_{i=1}^n \big((\hat{\Phi}_p\theta - \hat{Y})_i\big)^2 + \alpha\|\theta\|^2 = \|\hat{\Phi}_p\theta - \hat{Y}\|_2^2 + \alpha\|\theta\|^2 \tag{1.30}$$

for any $\alpha \geq 0$. Therefore, the map $\theta \mapsto \hat{R}(f_\theta)$ is convex (not necessarily strictly) and the global minima are stationary points. The gradient in θ is given by

$$n\nabla_\theta \hat{R}(f_\theta) = n\nabla_\theta\big[\theta^T\hat{\Phi}_p^T\hat{\Phi}_p\theta - 2\hat{Y}^T\hat{\Phi}_p\theta + \hat{Y}^T\hat{Y} + \alpha\theta^T\theta\big] \tag{1.31}$$

$$= 2\big(\theta^T(\hat{\Phi}_p^T\hat{\Phi}_p + \alpha I) - \hat{Y}^T\hat{\Phi}_p\big) \tag{1.32}$$

Since the gradient is linear in θ, the map has constant Hessian so that, reversely, all stationary points are global minima. They are the solutions of $0 = \nabla_\theta \hat{R}(f_\theta)$ giving rise to the normal equations

$$(\hat{\Phi}_p^T\hat{\Phi}_p + \alpha I)\theta = \hat{\Phi}_p^T\hat{Y}. \tag{1.33}$$

Now, observe that $\hat{\Phi}_p \in \mathbb{R}^{n \times p}$ is nothing but a rectangular *Vandermonde matrix*. For the deterministic case it is well known that these matrices have maximal rank if $X_i \neq X_j$ whenever $i \neq j$. In our setup of a uniform distribution on $X = [0, 1]$, it is clear that this holds at least μ-almost surely. Thus, we have three possibilities:

1. If $\alpha = 0$ and $p \leq n$ we are in the *under-parameterized regime* where $\hat{\Phi}_p^T\hat{\Phi}_p \in \mathbb{R}^{p \times p}$ has full-rank and is hence invertible. We obtain the *least squares estimator* from
$$\hat{\theta}_0 = (\hat{\Phi}_p^T\hat{\Phi}_p)^{-1}\hat{\Phi}_p^T Y. \tag{1.34}$$

2. If $\alpha = 0$ and $p > n$ we are in the *over-parameterized regime* where $\operatorname{rank}\hat{\Phi}_p^T\hat{\Phi}_p = n < p$. Here $\hat{\Phi}_p$ has full range, so that existence of (whole subspaces of) solutions to equation (1.33) is guaranteed. In that case we choose the *minimum-norm least squares estimator*
$$\hat{\theta}_0 = (\hat{\Phi}_p^T\hat{\Phi}_p)^\dagger\hat{\Phi}_p^T Y \coloneqq \operatorname{argmin}_{\theta:\, \hat{\Phi}_p^T\hat{\Phi}_p\theta = \hat{\Phi}_p^T\hat{Y}} \|\theta\|_2. \tag{1.35}$$

3. If $\alpha > 0$ for any n and p, the matrix $\hat{\Phi}_p^T\hat{\Phi}_p + \alpha I$ is invertible and we can define the *regularized least squares estimator*
$$\hat{\theta}_\alpha = (\hat{\Phi}_p^T\hat{\Phi}_p + \alpha I)^{-1}\hat{\Phi}_p^T Y. \tag{1.36}$$

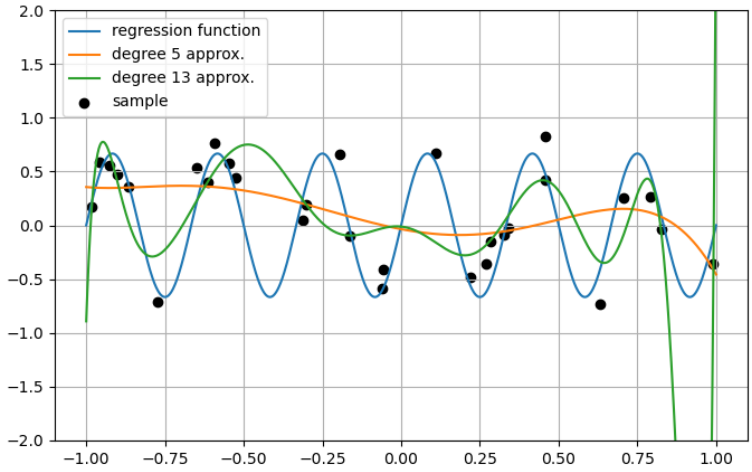

Figure 1.1: Least squares estimation of noisy observations of a sine wave ($n = 30$ and $\sigma = 1/5$) by polynomials of degree 5 and 13 and norm penalty strength $\alpha = 0$.

Note. The notation $A^\dagger$ and how to compute minimum-norm least squares solutions of minimal norm will be the topic of chapter 2 where we introduce the *pseudoinverse*. In fact, all of the above estimators will be generalized in said chapter, so keep this part in mind.

Note that both α and p are capacity controls. To demonstrate their effects we conduct simple numerical experiments. Let us first demonstrate how produced estimators could look like. In figure 1.1, we can see that while both the degree-5 and the degree-13 polynomial fail to capture the full complexity of the sine wave, a higher degree helps seems to help understanding the nature of the function (keep in mind the Taylor series of the sine function). However, we also see that higher degrees can become unstable: even though the degree-13 polynomial fits the two most-right points of the sample, the behavior in-between is wild. Perhaps more sample points can help, and if they are not available one can pass to other regularization methods. This is consistent with our idea of structural risk minimization keeping in mind that $\mathcal{H}_5 \subset \mathcal{H}_{13}$.

Next, let us see how the norm penalty acts. Figure 1.2 demonstrates that the regression function (a polynomial of degree 7) is comparatively not well understood by a degree-4 polynomial as expected one might expect. A degree-7 estimator does achieve low error, but the estimation gets worse as α increases. This might be attributed to the fact that the hypothesis space already has the right capacity and increasing α limits the effective reach of the least squares estimation. This is confirmed by the last the degree-9 estimation: the hypothesis space is more capable than necessary. We can see that there is

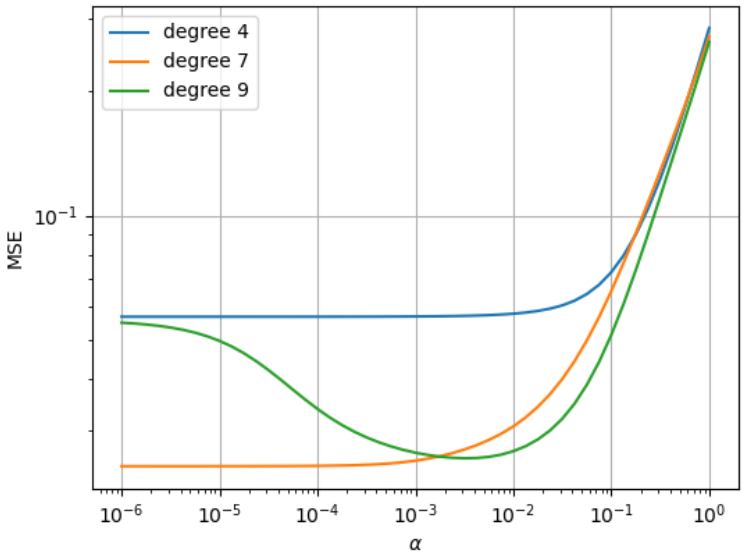

Figure 1.2: The regression function is a degree-7 polynomial, which is observed through $n = 100$ noisy samples with $\sigma = 1/2$. The mean-squared error (MSE) $\frac{1}{N}\sum_i(\hat{y}_i - y_i)^2$ of true values y_i against predictions $\hat{y}_i = \hat{f}_\alpha(x_i)$ on a large test set $\{(x_i, y_i)\}_{i=1}^N$ of true values is plotted against the regularization strength α. These estimators are polynomials of varying degrees.

a sweet-spot α^* that imposes just enough regularization for the error to be almost equal to the one of the true hypothesis space.

1.5 Modern research: understanding learning with high-capacity hypothesis spaces

The experiments in the last examples have confirmed the classical intuition about the bias-variance tradeoff. Note, however, how simple the setups where: low data dimensionality, few parameters, controllable noise, "nice" regression functions. The advent of deep learning has put these foundational ideas on trial. There, we are talking highly over-parameterized ($p \gg n$) models that are able to fit very high-dimensional and very noisy data (images, videos, text, genome expression profiles, ...). This has sparked new statistial learning research in many directions. We attempt to give a brief overview.

The fundamental question that leads these new fields is why highly over-parameterized deep neural nets perform as well as they do given that the usual established model, the bias-variance trade-off, is seemingly not capuring this regime at all. The idea is that in order to find a good estimator one should

choose the complexity of the hypothesis space to be high enough to avoid *underfitting* but simultaneously low enough to avoid *overfitting*. However, modern, practically successful, deep neural nets tend to have a parameter count that exceeds the training sample count by orders of magnitude. This is not captured by classical theory as most risk bounds, for instance the uniform defect bound are based on a complexity measure of the used hypothesis space. If it is too big (as for very complicated neural nets), we obviously don't expect any got rates anymore. Sp the question is two-fold: is overparameterization the reason for their performance and what exactly happens in this *interpolating regime*.

Researchers have ever since started to explore the regime with models that are easier to deal with and, in fact, found that even estmators as simple as least squares can perform well when they are perfectly fitting the data! This means that interpolating estimators are (under some condtions) apparently *not* prone to variance error (overfitting).

Bartlett et al. 2020 have proved this, pointing out that the main necessary condition for so-called *benign overfitting* to happen in high-dimensional and noisy settings, is low "effective" dimension of the data. This is in line with the "folklore" manifold hypothesis, stating that even technically is high-dimensional (think about images or videos), the examples will in reality live on lower-dimensional submanifold-like structure. Benign overfitting scenarios weer ever since discovered and studied in other models (e.g. Tsigler and Bartlett 2020; Pagliana et al. 2020). A related phenomenon is *double descent* (Belkin et al. 2019; Belkin, Hsu, and Xu 2020), in which under similar assumptions, one can see a second descent in out-of-sample error when increasing the parameter count (complexity) of the model after the critical point that classical theory would predict overfitting for. An interesting classification was also done by Mallinar et al. 2022

In particular, there are many indications that kernel learning methods behave analgously to deep learning from many different perspectives. Much evidence was provided in Belkin, Ma, and Mandal 2018. First, many kernels as for instance Laplacian and Gaussian are always able to interpolate the training data perfectly, providing a prime example of interpolating estimators. Unlike deep neural nets, kernel methods are easy to handle not only on paper but also practically due to the very small amount of hyper-parameters and there are many tasks in which their performance is comparable to the latter or even exceed it. In Liang and Rakhlin 2020, the authors provide experimental and theoretic evidendence that it is better to interpolate instead of the classically suggested norm penalties (*à la* Tikhonov) under certain circumstances. Now, these circumstances have not been fully understood but some necessary conditions seem to have pinned down: high-dimensionality (Rakhlin and Zhai 2019; Bartlett et al. 2020) with low effective dimension (Bartlett et al. 2020) and properly chosen curvature of the kernel w.r.t. the nature of the data (Liang and Rakhlin 2020). Moreover, there seem to be strong links (Chen and Xu

2020) between Laplacian kernels and ReLU feed-forward neural nets through another kernel called the *neural tangent kernel* that is able to capture the learning dynamics of gradient descent there.

These considerations have also sparked new interest in the other direction: due to all the analogies between kernels and neural nets researchers have started to scale up kernel methods to deal with big data (see for instance Rudi, Carratino, and Rosasco 2017).

2 Least squares estimation and spectral filtering

This chapter is devoted to the formal development and generalization of the least squares estimators from section 1.4. We develop the minimum-norm estimators for general bounded operators on Hilbert spaces and then specialize to compact operators where can discuss the Tikhonov spectral filter, which gives us a solid interpretation of the norm penalty in ERM. Finally, we demonstrate our findings through experiments. We follow Clason 2021, Chapters 3 to 6, for the first two sections.

Throughout, we fix Hilbert spaces H_1 and H_2. Let $A\colon H_1 \to H_2$ be a linear operator. We develop the theory around solving a deterministic *linear inverse problem*

$$Ax = y \tag{2.1}$$

in $x \in H_1$ for a datum $y \in H_2$. Such a problem is called *well-posed* if three conditions are satisfied: a solution $x \in H_1$ exists, it is unique and the dependence of x on y is continuous (if $Ax_n \to y$ then $x_n \to x$). For instance, in finite dimensional Euclidean space, well-posedness is equivalent to $\det A \neq 0$ where the matrix inverse is a continuous linear operator.

Most interesting cases we will encounter are *ill-posed* or "barely well-posed" in the sense of high condition numbers. In these cases we instead attempt to solve them approximately by finding $x \in H_1$ that minimizes the distance $\|Ax - y\|_{H_2}$. Such solutions usually exist, but they may still not be unique or may not depend continuously on y.

2.1 Minimum-norm least squares solutions and pseudoinverse

We fix a linear bounded operator $A \in B(H_1, H_2)$.

Definition 2.1. *An element $x^* \in H_1$ is called a least squares (LS) solution of the inverse problem $Ax = y$ if*

$$\|Ax^* - y\|_{H_2} = \min_{z \in X} \|Az - y\|_{H_2}. \tag{2.2}$$

If x^* is a least squares solution and has minimal norm among all others, we call it a minimum-norm least squares (MNLS) solution and usually write $x^\dagger$.

In the following we construct an operator, the *pseudoinverse*, whose core is defined on the domain of true invertibility and then extended to the maximal domain in which LS solutions exist. We will see that it maps any y to the corresponding MNLS solution.

Proposition 2.2 (Construction of pseudoinverse)**.** *Set*

$$\widetilde{A} \coloneqq A|_{\ker(A)^\perp} \colon \ker(A)^\perp \to \operatorname{ran}(A). \tag{2.3}$$

Then, $\widetilde{A}^{-1}$ *extends uniquely to a linear operator* $A^\dagger$ *such that*

$$\operatorname{dom}(A^\dagger) = \operatorname{ran}(A) \oplus \operatorname{ran}(A)^\perp, \quad \ker(A^\dagger) = \operatorname{ran}(A)^\perp. \tag{2.4}$$

Proof. $\widetilde{A}$ is bijective and linear, so that $A^\dagger$ exists on $\operatorname{ran}(A)$ and is linear there. Now let $y \in \operatorname{dom}(A^\dagger)$ be uniquely decomposed as $y = y_1 + y_2$ with $y_1 \in \operatorname{ran}(A)$ and $y_2 \in \operatorname{ran}(A)^\perp$ according to the orthogonal sum. We must have $A^\dagger y_2 = 0$ since $y_2 \in \ker(A^\dagger) = \operatorname{ran}(A)^\perp$. Thus, defining

$$A^\dagger y \coloneqq A^\dagger y_1 + A^\dagger y_2 = A^\dagger y_1 = \widetilde{A}^{-1} y_1 \tag{2.5}$$

yields the only possible linear extension. □

Lemma 2.3 (Properties of pseudoinverses)**.** *Let* $P_{\ker}$ *and* $P_{\overline{\operatorname{ran}}}$ *be the orthogonal projections onto* $\ker(A)$ *and* $\overline{\operatorname{ran}}(A)$*, respectively. Then:*

1. $\operatorname{ran}(A^\dagger) = \ker(A)^\perp$ *and* $\operatorname{dom}(A^\dagger)$ *is dense in* H_2
2. $A^\dagger A = I - P_{\ker}$ *and* $AA^\dagger A = A$*,*
3. $AA^\dagger = P_{\overline{\operatorname{ran}}}|_{\operatorname{dom}(A^\dagger)}$ *and* $A^\dagger A A^\dagger = A^\dagger$*.*

Proof. 1. For all $y \in \operatorname{dom}(A^\dagger) = \operatorname{ran}(A) \oplus \operatorname{ran}(A)^\perp$ we have $P_{\overline{\operatorname{ran}}} y \in \operatorname{ran}(A)$ (and not only in its closure!) by the direct sum decomposition so that by construction $A^\dagger y = A^\dagger P_{\overline{\operatorname{ran}}} y = \widetilde{A}^{-1} P_{\overline{\operatorname{ran}}} y$. Therefore, $A^\dagger y \in \operatorname{ran}(\widetilde{A}^{-1}) = \ker(A)^\perp$ and $\operatorname{ran}(A^\dagger) \subseteq \ker(A)^\perp$. Conversely, if $x \in \ker(A)^\perp$, then $A^\dagger A x = \widetilde{A}^{-1}\widetilde{A} x = x$ so that $\ker(A)^\perp \subseteq \operatorname{ran}(A^\dagger)$.

The density of $\operatorname{dom}(A^\dagger)$ follows:

$$\overline{\operatorname{dom}(A^\dagger)} = \overline{\operatorname{ran}(A) \oplus \operatorname{ran}(A)^\perp} = \overline{\operatorname{ran}}(A) \oplus \operatorname{ran}(A)^\perp = \ker(A^*)^\perp \oplus \ker(A^*) = H_2. \tag{2.6}$$

2. If $x \in H_1$,

$$A^\dagger A x = \widetilde{A}^{-1} A x = \widetilde{A}^{-1} A (P_{\ker} x + (I - P_{\ker}) x) \tag{2.7}$$

$$= \underbrace{\widetilde{A}^{-1} A P_{\ker} x}_{=0} + \underbrace{\widetilde{A}^{-1}\widetilde{A}}_{=I} (I - P_{\ker}) x = (I - P_{\ker}) x. \tag{2.8}$$

Hence,

$$AA^\dagger A = A(I - P_{\ker}) = A - A P_{\ker} = A.$$

3. If $y \in \mathrm{dom}(A^\dagger)$, then by the same arguments as in item 1

$$AA^\dagger y = A\widetilde{A}^{-1}P_{\overline{\mathrm{ran}}}y = \widetilde{A}\widetilde{A}^{-1}P_{\overline{\mathrm{ran}}}y = P_{\overline{\mathrm{ran}}}y. \tag{2.9}$$

Moreover,

$$A^\dagger AA^\dagger y = A^\dagger P_{\overline{\mathrm{ran}}}y = \widetilde{A}^{-1}y = A^\dagger y. \tag{2.10}$$

□

Now we make sure that the pseudoinverse is exactly the right object to compute for finding MNLS solutions as claimed.

Proposition 2.4 (Characterization of ls solutions)**.** *If $y \in \mathrm{dom}(A^\dagger)$, then the problem $\min_{x\in H_1}\|y - Ax\|$ has solutions which, in turn, are exactly the solutions of the equation*

$$Ax = P_{\overline{\mathrm{ran}}}y. \tag{2.11}$$

Among these, there is a unique solution $x^\dagger$ of minimum norm given by $x^\dagger = A^\dagger y$ and the set of solutions to the least squares problem is $x^\dagger + \ker(A)$.

Proof. If $y \in \mathrm{dom}(A^\dagger)$. Since $P_{\overline{\mathrm{ran}}}y \in \mathrm{ran}(A)$ there is at least one solution z of equation (2.11). By the optimality of orthogonal projections we have

$$\|y - Az\| = \|y - P_{\overline{\mathrm{ran}}}y\| = \min_{w\in\overline{\mathrm{ran}}(A)}\|y - w\| \le \|y - Ax\| \tag{2.12}$$

for all $x \in H_1$ and z is a LS solutions.

On the other hand, if z is a least squares solution then

$$\|y - P_{\overline{\mathrm{ran}}}y\| \le \|y - Az\| = \min_{x\in H_1}\|y - Ax\| = \min_{w\in\mathrm{ran}(A)}\|w - y\| \le \|y - P_{\overline{\mathrm{ran}}}y\| \tag{2.13}$$

so we have an equality. Hence the ls solutions are exactly those to equation (2.11).

Each ls solution x can be decomposed uniquely as $x = \overline{x}+x_0 \in \ker(A)^\perp \oplus \ker(A)$. Let $x' = \overline{x}' + x_0'$ be another ls solution. Then, $A\overline{x} = Ax = P_{\overline{\mathrm{ran}}}y = Ax' = A\overline{x}'$ but A is injective on $\ker(A)^\perp$, so $\overline{x} = \overline{x}'$. By orthogonality of the two components

$$\|x\|^2 = \|\overline{x} + x_0\|^2 = \|\overline{x}\|^2 + \|x_0\|^2 \ge \|\overline{x}\|^2, \tag{2.14}$$

which means that $x^\dagger := \overline{x}$ is the unique mnls solution and the affine subspace of all ls solutions is $x^\dagger + \ker(A)$. Finally, by lemma 2.3

$$x^\dagger = \overline{x} = \overline{x} - P_{\ker}\overline{x} = (I - P_{\ker})\overline{x} = A^\dagger A\overline{x} = A^\dagger P_{\overline{\mathrm{ran}}}y = A^\dagger AA^\dagger y = A^\dagger y. \tag{2.15}$$

□

The previous characterization states that $x \in H_1$ is a Ls solution iff $Ax = P_{\overline{\mathrm{ran}}}y$, that is, iff $Ax \in \overline{\mathrm{ran}}(A)$ (empty condition) and $Ax - y \in (\overline{\mathrm{ran}}(A))^\perp$. Taking into account that $(\overline{\mathrm{ran}}(A))^\perp = \ker(A^*)$, the second condition means that ls solutions to $Ax = y$ are exactly those that solve $A^*(Ax - y) = 0$: We have recovered the normal equations from the introduction. Let us reframe the characterization for later reference.

Theorem 2.5 (Ordinary least squares)**.** If $y \in \mathrm{dom}(A^\dagger)$, then the linear inverse problem $Ax = y$ has ls solutions and these are exactly the solutions of

$$A^*Ax = A^*y, \tag{2.16}$$

They are all of the form $x^\dagger + \ker(A)$, where $x^\dagger$ is the unique mnls solution

$$x^\dagger = A^\dagger y = (A^*A)^\dagger A^*y. \tag{2.17}$$

Of particular importance is the fact that equation (2.17) reduces the computation of the mnls solution to the pseudoinverse of the selfadjoint operator (A^*A). This gives us access to spectral methods that will be explored in the next section.

Before that, though, we are going to tackle the question of the third condition of well-posed problems: continuous dependence of the solution on the data y. We will see that this is closely linked to the domain of the pseudoinversed operator.

Proposition 2.6 (Continuity of pseudoinverse)**.** *The pseudoinverse $A^\dagger\colon \mathrm{dom}(A^\dagger) \to H_1$ is a bounded operator iff $\mathrm{ran}(A)$ is closed (i.e. iff $\mathrm{dom}(A^\dagger) = H_2$ by lemma 2.3).*

Proof.

- Let us first suppose that $T^\dagger$ is bounded, which implies that it maps Cauchy sequences to Cauchy sequences. Since $\mathrm{dom}(A^\dagger)$ is dense in H^2 we can pick a sequence $y_n \to y$ contained in the domain for a given y and define a continuous extension $\overline{A^\dagger}y \coloneqq \lim A^\dagger y_n$. If we pick $y \in \overline{\mathrm{ran}}(A)$ and $(y_n) \subset \mathrm{ran}(A)$ we have by lemma 2.3

$$y = P_{\overline{\mathrm{ran}}}y = \lim P_{\overline{\mathrm{ran}}}y_n = \lim AA^\dagger y_n = A\overline{A^\dagger}y \in \mathrm{ran}(A), \tag{2.18}$$

  and thus, $\overline{\mathrm{ran}}(A) = \mathrm{ran}(A)$.

- Let now, conversely $\overline{\mathrm{ran}}(A) = \mathrm{ran}(A)$ i.e. $\mathrm{dom}(A^\dagger) = H_2$. In order to prove that $A^\dagger\colon H_2 \to H_1$ is continuous, we invoke the Closed Graph theorem stating that a linear operator between Banach spaces (note that the assumption enters here for the completeness) is continuous iff it has a closed graph. In particular, we suppose that we have a sequence $y_n \to y$ in H_2 so that $A^\dagger y_n \to x \in H_1$ and we have to show that $A^\dagger y = x$. With the help of lemma 2.3 and the continuity of A we find

$$P_{\overline{\mathrm{ran}}}y = \lim P_{\overline{\mathrm{ran}}}y_n = \lim AA^\dagger y_n = AA^\dagger y = Ax, \tag{2.19}$$

  which means that x is a LS solution of $Ax = y$, e.g. of the form $x = x^\dagger + x_0$ with $x_0 \in \ker(A)$. Finally, since $\mathrm{ran}(A^\dagger) = \ker(A)^\perp$ is closed, we have $A^\dagger y_n \to x \in \ker(A)^\perp$, so that x_0 must be 0 and thus $A^\dagger y = x^\dagger = x$.

□

Note. Compact operators with infinite-dimensional range can never have a pseudoinverse according to the previous result.

2.2 Regularization via Tikhonov spectral filtering

In this section we restrict outselves to compact operators $K \in K(H_1, H_2)$; all of the for us relevant operators are. Let $(v_n, \lambda_n)_n$ be the eigensystem of the positive operator K^*K, We fix a singular system $(\sigma_n, u_n, v_n)_{n\in\Lambda}$ of K where $\Lambda = \{n \mid \lambda_n \neq 0\}$. Recall that $\sigma_n = \sqrt{\lambda_n}$ and $u_n = \sigma_n^{-1} K v_n$. Moreover, we can decompose

$$Kx = \sum_n \sigma_n \, (v_n \otimes u_n)(x) = \sum_n \sigma_n \langle x, v_n \rangle_{H_1} u_n, \tag{2.20}$$

$$K^*y = \sum_n \sigma_n \, (u_n \otimes v_n)(y) = \sum_n \sigma_n \langle y, u_n \rangle_{H_2} v_n. \tag{2.21}$$

The pseudoinverse of a compact operator can be conveniently computed when its SVD is known in the most natural way thinkable: by inverting nonzero singular values. Following up on proposition 2.6 and the discussion above, such an expression is only possible on the "continuous" part of the maximal domain. The next result characterizes this using the so-called *Picard condition*.

Theorem 2.7 (Picard condition)**.** Let $y \in H_2$. Then, $y \in \mathrm{ran}(K)$ iff

$$\sum_n \sigma_n^{-2} \, |\langle y, u_n \rangle_{H_2}|^2 < \infty \tag{2.22}$$

with the convention that $\langle y, v_n \rangle = 0$ if $\sigma_n = 0$. In that case, the pseudoinverse can be written as

$$K^\dagger y = \sum_n \sigma_n^{-1} \, (u_n \otimes v_n)(y). \tag{2.23}$$

It is understood that the sum is actually taken over all n such that $\sigma_n > 0$.

Proof. • If $y \in \mathrm{ran}(K)$, there is $x \in H_1$ such that $Kx = y$. By the properties of singular values

$$\langle y, u_n \rangle_{H_2} = \langle x, K^* u_n \rangle_{H_1} = \sigma_n \langle x, v_n \rangle_{H_1}. \tag{2.24}$$

Hence, by the Bessel inequality:

$$\sum_n \sigma_n^{-2} \, |\langle y, u_n \rangle_{H_2}|^2 = \sum_n |\langle x, v_n \rangle_{H_1}|^2 < \infty. \tag{2.25}$$

- Conversely, if $y \in \overline{\operatorname{ran}}(K)$ satisfies the Picard condition, the partial sums of the rhs of equation (2.22) form a Cauchy sequence. Then, also the sequence defined by

$$x_N = \sum_{n=1}^{N} \sigma_n^{-1} \langle y, u_n \rangle_{H_2} v_n \tag{2.26}$$

is a Cauchy sequence since $\|x_N - x_M\|^2$ is easily bounded by the former. It convergences to the series x in the closed space $\overline{\operatorname{ran}}(K^*) = \ker(K)^\perp$. Therefore, using the fact that the u_n are an onb for $\overline{\operatorname{ran}}(K)$,

$$Kx = \sum_n \sigma_n^{-1} \langle y, u_n \rangle_{H_2} K v_n = \sum_n \langle y, u_n \rangle_{H_2} u_n = P_{\overline{\operatorname{ran}}} y = y, \tag{2.27}$$

i.e. $y \in \operatorname{ran}(K)$.

Equation (2.23) now follows from the fact that $Kx = P_{\overline{\operatorname{ran}}} y$ is equivalent to $x = K^\dagger y$ for $x \in \ker(K)^\perp$. □

Let us rearrange equation (2.23) a bit. Write

$$K^\dagger y = (K^*K)^\dagger K^* y = \sum_n \sigma_n^{-1} \, (u_n \otimes v_n)(y) \tag{2.28}$$

$$= \sum_n \sigma_n^{-2} \sigma_n \, (u_n \otimes v_n)(y) \tag{2.29}$$

$$= \sum_n \psi_0(\sigma_n^2) \sigma_n \, (u_n \otimes v_n)(y), \tag{2.30}$$

$$= \psi_0(K^*K) K^* y \tag{2.31}$$

where we used the spectral function $\psi_0(\sigma) = 1/\sigma$. We will now argue that the unboundedness of $K^\dagger$ in case of $\dim \operatorname{ran}(K) = \infty$ (i.e. the set Λ from above has cardinality ∞) is reflected by the fact that $\psi_0(\sigma) \to \infty$ when $\sigma \to 0$. In the same lines, this will open us an intuitive way of regularizing the pseudoinverse by paying the price of controllable inaccuracies as an operator that should reflect the inverse.

Definition 2.8. *Let $\kappa = \|K^*K\| = \max_n \lambda_n$ be the spectral radius of K^*K. The Tikhonov regularizing filter of $(K^*K)^\dagger$ is the family $\{\psi_\alpha\}_{\alpha>0}$ of spectral functions $\psi_\alpha \colon (0, \kappa] \to \mathbb{R}$ defined by*

$$\psi_\alpha(\sigma) \coloneqq \frac{1}{\sigma + \alpha} \tag{2.32}$$

Moreover, we define the Tikhonov regularization as the linear operator

$$T_\alpha \coloneqq \psi_\alpha(K^*K) K^*. \tag{2.33}$$

Note that $\psi_\alpha(\sigma)$ is bounded for all $\sigma > 0$ if $\alpha > 0$. Similar to equation (2.30), we apply the filter as

$$T_\alpha(y) = \sum_n \psi_\alpha(\sigma_n^2) \sigma_n \, (u_n \otimes v_n)(y) = \sum_n \frac{\sigma_n}{\sigma_n^2 + \alpha} \, (u_n \otimes v_n)(y). \tag{2.34}$$

This, indeed, yields a bounded operator that, when applied, approximates the MNLS solution to the inverse problem.

Proposition 2.9. *For all $\alpha > 0$, we have that $T_\alpha \colon H_2 \to H_1$ is a bounded operator. Moreover, if $y \in \operatorname{dom}(K^\dagger)$, the regularized solution $x_\alpha \coloneqq T_\alpha y$ approximates the MNLS solution $x_0 \coloneqq T_0 y \coloneqq K^\dagger y$, i.e. $\lim_{\alpha \to 0} \|x_\alpha - x_0\| = 0$.*

Proof. The boundedness follows from the boundedness of the applied spectral function. We have

$$\|x_\alpha - x_0\|^2 = \sum_n \left(\frac{1}{\sigma_n} - \frac{\sigma_n}{\sigma_n^2 + \alpha} \right)^2 |\langle y, u_n \rangle|^2 = \sum_n \left(\frac{\alpha}{\sigma_n^2 + \alpha} \right)^2 \frac{1}{\sigma_n^2} |\langle y, u_n \rangle|^2. \tag{2.35}$$

Now, since $\alpha/(\sigma_n^2 + \alpha) \leq 1$, converges to 0 for $\alpha \to 0$, and $\sum_n \sigma_n^{-2} |\langle y, u_n \rangle|^2 < \infty$, we can conclude by dominated convergence. □

Finally, we finish our discussion of least squares by proving a sibling to theorem 2.5 which holds for regularized least squares (rls) solutions.

Theorem 2.10 (Regularized least squares)**.** If $\alpha > 0$ and $y \in H_2$, then $x_\alpha = R_\alpha y$ iff

$$(K^*K + \alpha I) x_\alpha = K^* y. \tag{2.36}$$

Moreover, x_α is the unique solution to the minimization problem

$$\min_{x \in H_1} \|Kx - y\|_{H_2}^2 + \alpha \|x\|_{H_1}^2. \tag{2.37}$$

Proof. We have the following eigendecompositions

$$\alpha x_\alpha = \sum_n \alpha \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle y, u_n \rangle_Y v_n, \tag{2.38}$$

$$K^*K x_\alpha = \sum_n \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle y, u_n \rangle K^*K v_n = \sum_n \sigma_n^2 \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle y, u_n \rangle v_n. \tag{2.39}$$

Thus,

$$(K^*K + \alpha I) x_\alpha = \sum_n \sigma_n \langle y, u_n \rangle v_n = K^* y. \tag{2.40}$$

and the first implication is shown.

Now let $x \in H_1$ be a solution of equation (2.36). We plug

$$x = \sum_n \langle x, v_n \rangle v_n + P_{\ker} x \tag{2.41}$$

in and find

$$\sum_n (\sigma_n^2 + \alpha) \langle x, v_n \rangle v_n + \alpha P_{\ker} x = (K^*K + \alpha I) x_\alpha = K^* y = \sum_n \sigma_n \langle y, u_n \rangle v_n. \tag{2.42}$$

We immediately see that $P_{\ker} x = 0$ since the v_n are an ONB for $\ker(K)^\perp$, so that by comparing coefficients we find

$$\langle x, v_n \rangle = \frac{\sigma_n}{\sigma_n^2} \langle y, u_n \rangle. \tag{2.43}$$

Thus,

$$x = \sum_n \langle x, v_n \rangle v_n = \sum_n \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle y, u_n \rangle v_n = x_\alpha. \tag{2.44}$$

Finally we need to prove that x_α is the unique minimizer. Let $J_\alpha(X) := \|Kx - y\|^2 + \alpha \|x\|^2$. Then, for all x

$$J_\alpha(x) - J_\alpha(x_\alpha) = \langle Kx - y, Kx - y \rangle + \alpha \langle x, x \rangle - \langle Kx_\alpha - y, Kx_\alpha - y \rangle - \alpha \langle x_\alpha, x_\alpha \rangle \tag{2.45}$$

$$= \|Kx - Kx_\alpha\|^2 + \alpha \|x - x_\alpha\|^2 + \langle K^*(Kx_\alpha - y) + \alpha x_\alpha, x - x_\alpha \rangle \tag{2.46}$$

$$= \|Kx - Kx_\alpha\|^2 + \alpha \|x - x_\alpha\|^2 \geq 0, \tag{2.47}$$

where we used the normal equations from before. Thus, x_α is a minimizer. If there was another minimizer $\tilde{x}$, then also $J_\alpha(x) - J_\alpha(\tilde{x}) \geq 0$ for all x. Let $x = \tilde{x} + tz$ for arbitrary z and $t > 0$. Then,

$$0 \leq J_\alpha(\tilde{x} + tz) - J_\alpha(\tilde{x}) = t^2 \|Kz\|^2 + t^2 \alpha \|z\|^2 + t \langle K^*(K\tilde{x} - y) + \alpha \tilde{x}, z \rangle. \tag{2.48}$$

Dividing by t and then letting $t \to 0$ yields

$$\langle K^*(K\tilde{x} - y) + \alpha \tilde{x}, z \rangle \geq 0. \tag{2.49}$$

But z was arbitrary, so that this can only hold if $K^*K\tilde{x} + \alpha \tilde{x} = K^*y$. By previously established uniqueness we get $\tilde{x} = x_\alpha$. □

Finally for this section, let us see the effect of spectral filtering in the polynomial least squares experiment from section 1.4. Figure 2.1 shows how increasing regularization strength "eats up" the eigenvalues: indeed, α acts as a lower bound for the minimal eigenvalue. The prediction ability of the estimator depends on having "enough spectrum" available that is not regularized away. Indeed, if we compare the two right-most graphs, we see that the sweet spot is attained exactly when we regularize just so that the lower spectral bands coincide with the one of the right hypothesis space.

2.3 Random design and regularization strength

We now study how MNLS estimators perform aginst RLS ones in a simple test case within our probabilistic learning setup. The great advantage of the feature map approach of section 1.4 was that we could apply linear estimation

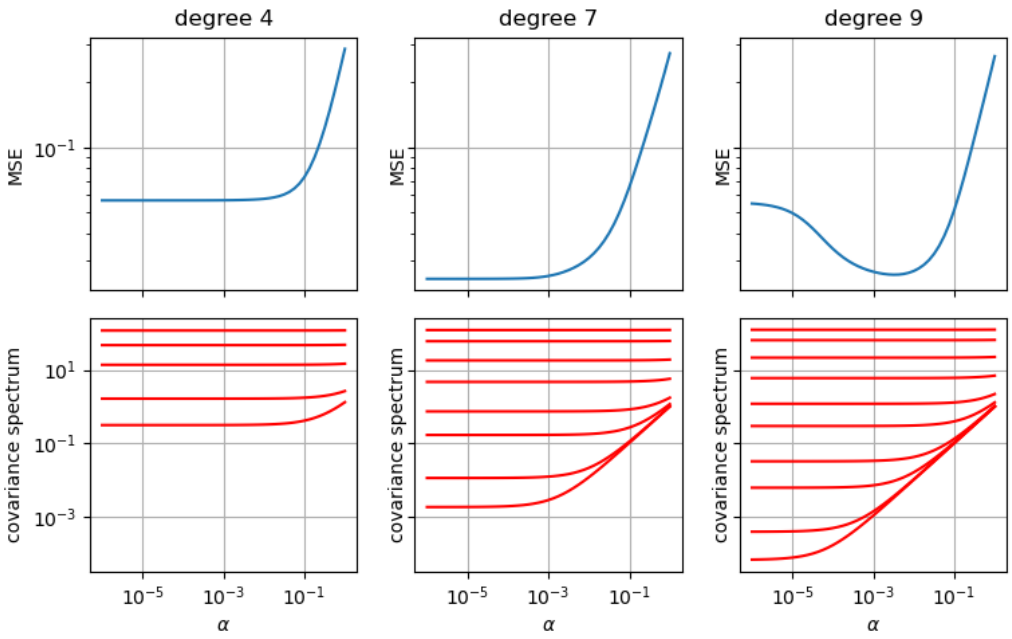

Figure 2.1: For the same three error curves of figure 1.2 we plot the spectra of the empirical covariance matrices $n^{-1}\hat{\Phi}_d^T\hat{\Phi}_d$. The n-th curve from the top indicates the development of the n-largest eigenvalue in α.

techniques with non-linear estimators. Let $X = \mathbb{R}^d$, $\Phi\colon \mathbb{R}^d \to \mathbb{R}^D$ any map, and $Y = \mathbb{R}$. We specialize our probabilistic model form section 1.1 to the case of linear estimators:

$$Y = \langle \theta^*, X \rangle + \varepsilon, \tag{2.50}$$

where the regression functions is now represented by a fixed optimal parameter vector $\theta^* \in \mathbb{R}^D$. Let $\hat{Z}^n = \{(X_i, Y_i)\}_{i=1}^n$ be a sample where $Y_i = \langle \theta^*, X_i \rangle + \varepsilon_i$. Write $\hat{Y} = [Y_1, \ldots, Y_n]^T \in \mathbb{R}^n$ and $\hat{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^T \in \mathbb{R}^n$.

Assume that $\mathbb{E}[X] = 0$. Then, recall that the covariance operator of X is defined by $\Sigma = \mathbb{E}[XX^T] \in \mathbb{R}^{D\times D}$. Similarly, if we let $\hat{X} \in \mathbb{R}^{D\times n}$ be the matrix in which the columns are the input samples $X_i \in \mathbb{R}^D$, we can write the empirical covariance as $\hat{\Sigma} = n^{-1}\hat{X}\hat{X}^T$.

Let us call $\hat{\theta}_\alpha$ the RLS solution of the linear inverse problem $\hat{X}^T\theta = \hat{Y}$ Then, we can obtain a tractable bias-variance risk decomposition for the MNLS/RLS estimator $\hat{f}_\alpha := \langle \hat{\theta}_\alpha, \cdot \rangle$.

Proposition 2.11. *Assume that ε is independent of X. In the above notation and for any $\alpha > 0$ we have the decomposition*

$$\mathbb{E}[R(\hat{f}_\alpha) \mid \hat{X}] = \mathbb{E}[\mathbf{B} + \mathbf{V} \mid \hat{X}] + \sigma^2, \tag{2.51}$$

where bias and variance are given by

$$\mathbf{B} = \alpha^2 \langle \hat{\Sigma}(\hat{\Sigma} + \alpha I)^{-1} \Sigma (\hat{\Sigma} + \alpha I)^{-1} \hat{\theta}^*, \theta^* \rangle, \tag{2.52}$$

$$\mathbf{V} = \frac{\sigma^2}{n} \operatorname{trace}(\hat{\Sigma}(\hat{\Sigma} + \alpha I)^{-1} \Sigma (\hat{\Sigma} + \alpha I)^{-1}). \tag{2.53}$$

Note that the expectation is taken only in the noise $\hat{\varepsilon}$. For a more detailed account of this composition with elementary bounds we refer to Mourtada and Rosasco 2022.

Proof. We have by theorems 2.5 and 2.10 and by the decomposition of $\hat{Y}$ that

$$\hat{\theta}_\alpha = (n^{-1} \hat{X} \hat{X}^T + \alpha I)^{-1} n^{-1} \hat{X} \hat{Y} \tag{2.54}$$

$$= n^{-1} (\hat{\Sigma} + \alpha I)^{-1} \hat{X} (\hat{X}^T \theta^* + \hat{\varepsilon}) \tag{2.55}$$

$$= (\hat{\Sigma} + \alpha I)^{-1} \hat{\Sigma} \theta^* + n^{-1} (\hat{\Sigma} + \alpha I)^{-1} \hat{X} \hat{\varepsilon}. \tag{2.56}$$

Therefore,

$$\hat{\theta}_\alpha - \theta^* = -\alpha (\hat{\Sigma} + \alpha I)^{-1} \hat{\Sigma} \theta^* + n^{-1} (\hat{\Sigma} + \alpha I)^{-1} \hat{X} \hat{\varepsilon}. \tag{2.57}$$

Now let is write out the risk with the expression obtained from proposition 1.3 (now as expectation in the variable X):

$$R(\hat{f}_\alpha) = \mathbb{E}_X[(\hat{f}^\alpha(X) - f^*(X))^2] \tag{2.58}$$

$$= \mathbb{E}_X[(X^T(\hat{\theta}_\alpha - \theta^*))^2] \tag{2.59}$$

$$= \mathbb{E}_X[\langle X X^T (\hat{\theta}_\alpha - \theta^*), \hat{\theta}_\alpha - \theta^* \rangle] \tag{2.60}$$

$$= \langle \Sigma(\hat{\theta}_\alpha - \theta^*), \hat{\theta}_\alpha - \theta^* \rangle =: \|\hat{\theta}_\alpha - \theta^*\|_\Sigma^2 \tag{2.61}$$

If we split the inner product at the terms of the defect of the estimator and write out each term individually inserting the previous results, then we see immediately that the two cross-terms vanish after taking expectation only in ε (it has zero mean). We are left with the two square terms. The bias term is

$$\mathbb{E}[\alpha^2 \|(\hat{\Sigma} + \alpha I)^{-1} \hat{\Sigma} \theta^*\|_\Sigma^2 \mid \hat{X}] \tag{2.62}$$

and the variance is given by

$$\mathbb{E}[n^{-2} \|(\hat{\Sigma} + \alpha I)^{-1} \hat{X} \varepsilon\|_\Sigma^2 \mid \hat{X}] \tag{2.63}$$

$$= n^{-2} \mathbb{E}[\hat{\varepsilon}^T \hat{X}^T (\hat{\Sigma} + \alpha I)^{-1} \Sigma (\hat{\Sigma} + \alpha I)^{-1} \hat{X} \hat{\varepsilon} \mid \hat{X}] \tag{2.64}$$

$$= \frac{\sigma^2}{n^2} \mathbb{E}[\operatorname{trace}(\hat{X}^T (\hat{\Sigma} + \alpha I)^{-1} \Sigma (\hat{\Sigma} + \alpha I)^{-1} \hat{X}) \mid \hat{X}] \tag{2.65}$$

$$= \frac{\sigma^2}{n} \mathbb{E}[\operatorname{trace}(\hat{\Sigma}(\hat{\Sigma} + \alpha I)^{-1} \Sigma (\hat{\Sigma} + \alpha I)^{-1}) \mid \hat{X}], \tag{2.66}$$

where the pre-last equation follows form the well-known fact

$$\mathbb{E}[\hat{\varepsilon}^T S \hat{\varepsilon}] = \operatorname{trace}[S \mathbb{E}[\hat{\varepsilon} \hat{\varepsilon}^T]] + \mathbb{E}[\hat{\varepsilon}]^T A \mathbb{E}[\hat{\varepsilon}], \tag{2.67}$$

and the last equation from the cyclic property of the trace. □

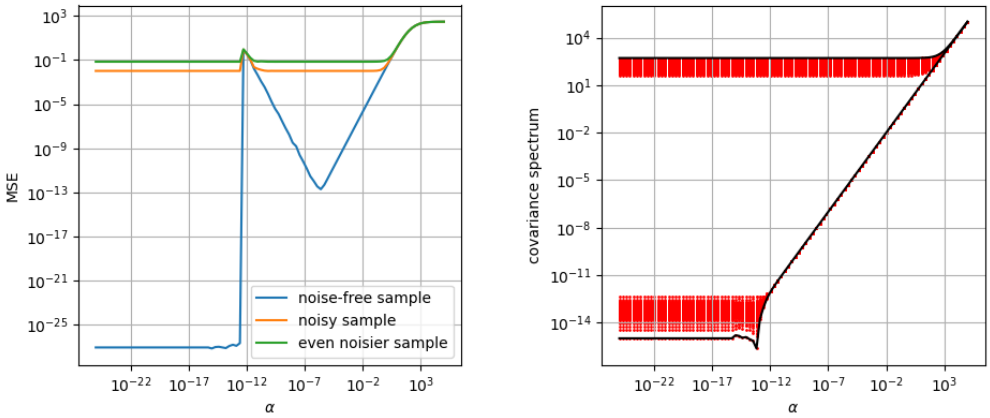

Figure 2.2: The $n = 200$ inputs are sampled with standard normal coefficients from a hyperplane of dimension $d = 50$ in the ambient space of dimension $D = 400$. The regression function is some linear function that is observed with increasing noise levels $\sigma = 0, 1/5, 1/2$.

We now put the above into action. Let $V = \text{span}\{v_1, \ldots, v_d\} \leq \mathbb{R}^D$ be a fixed d-dimensional hyperplane ($d \leq D$) and let the marginal μ be such that the input points take the form $X_i = \sum_{j=1}^{d} \lambda_j v_j$ with standard normal coefficients $\lambda_j \sim N(0, 1)$.

With this setup we aim at simulating a real setting in which the actual data dimension is huge but the effective dimension is comparatively low. Here, we expect the covariance spectrum to have two clusters: one for the hyperplane signal and one for the "useless" dimensions. In this setting, Bartlett et al. 2020 predict benign overfitting, so that we expect to see that very low to vanishing regularization strengths perform best.

Indeed, figure 2.2 shows that even if the sample is noisy, no regularization performs at least as good as regularization. In the noise-free case we see a "snapping" behavior in the transition from RLS to MNLS estimator leading to a perfect fit as soon as regularization is low enough to catch the low band of the covariance spectrum. This confirms the mentioned paper's findings: high-dimensionality but low effective rank can be learned well if we give the method access to the dimensions that have low to no significance.

On the other hand, if the observations are noisy, we do not get better than the regularized, but not regularizing has the big advantage that there is no model selection necessary. Observe as well that in both cases we have error bumps during the "eating phase".

Let us compare our findings to the risk decomposition. Both terms capture some sort of effective dimension. The variance in terms of the trace, which is a straightforward way to measure it and the bias more implicitly by computing the residuals of covariance and empirical covariance each multiplied by the regularized empirical inverse covariance. While we assumed $\alpha > 0$, the snap-

ping behavior in the noiseless case with $\alpha = 0$ is somewhat resembled. In the next chapters we provide a risk decomposition explicitly designed to capture both cases.

3 Reproducing kernel Hilbert spaces

In this chapter we give a complete mathematical account of the random-design kernel ridge(less) regression algorithm and we determine the fundamental quantities and operators that govern its behavior. Finally, we provide a blueprint risk decomposition that will provide intuition of how to assess the generalization performance in concrete problems and models.

To motivate the upcoming theory, let $X = [x_1 \mid \cdots \mid x_n]^T \in \mathbb{R}^{n \times d}$ be the matrix of inputs $x_i \in R^d$ and $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ the vector of real outcomes y_i according to some unknown law $f^* \colon \mathbb{R}^d \to \mathbb{R}$. Assume for now that X and Y are deterministic and that we expect f^* to be well-approximatable by a linear function: any estimator in our model will be of the form $f_\theta(x) = \langle \theta, x \rangle = \theta^T x$ for a parameter vector θ.

So far, we have explained the term "ridge/ridgeless regression", which generally refers to linear estimators found through least squares from the last chapter. The next step is to explore non-linear estimators obtained via "kernels".

The idea is similar to what we've done with polynomial regression in section 1.4. In a nutshell, the idea is to consider *feature maps* $\Phi \colon X \to H$ mapping the input data $x_i \in X$ (now coming from an arbitrary space) in a non-linear fashion to more useful features $\Phi(x_i) \in H$ into some convenient Hilbert space. Then, one can conceptually run all of the above derivations in H instead of $\mathbb{R}^d$, but a quick check to the matrix dimensions reveals that the problem scales in $d = \dim H$, severely limiting the actual usefulness of this method. However, there is a dual version that will turn out to change the roles of sample size n and feature space dimension d, enabling the practical use of infinite dimensional feature spaces.

Kernel methods work by replacing the euclidean inner products $\langle x_i, x_j \rangle_2$ between inputs by the inner product $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_H$ in feature space. The function k is called the *kernel*, to which there can be associated a unique Hilbert space of functions on X, in which all members are *reproducible* through the kernel function. Section 3.2 dives into these underlying spaces, the class of *reproducing kernel Hilbert spaces* (rkhs, pl. rkhss). We will see that they simultaneously act as feature space and as hypothesis space.

This special reproducing structure lets us prove the so-called *representer theorem* in section 3.3 which establishes the mentioned duality: when minimizing the empirical risk $\hat{R}_\alpha(f)$ in H it suffices to search in the subspace of

functions of the form

$$f(x) = \sum_{i=1}^{n} c_i k(x, x_i). \tag{3.1}$$

Similarly as before, one can write down the empirical risk, and find the coefficients c_i via ls:

$$\hat{f}_\alpha(x) = \sum_{i=1}^{n} \left((\hat{K} + \alpha I_n)^\dagger Y \right)_i k(x, x_i) \tag{3.2}$$

Thus, instead of finding a ($\dim H$)-dimensional we can instead look for an n-dimensional "parameter" vetcor. The quotation marks are necessary because not only does the number of parameters depend on the sample but also what these parameters actually parameterize: such kernel methods are *non-parametric*. Despite this, they allow for an efficient search of predictors living in infinite-dimensional spaces which would otherwise be intractable.

Finally, we introduce certain operators that will allow us to study kernel ridge(less) estimators in the probabilistic setup of Chapter 1. These will turn out to be favorably compact and admit spectral decompositions. The risk can be decomposed in terms of these operators and can, consequently, be analyzed via their spectra. We explain intuitively what the components of the decomposition are.

In this section we will mostly follow the standard reference Steinwart and Christmann 2008, Chapter 4. Throughout, let $X \neq \emptyset$ be a non-empty set. In view of Chapter 1, think about the space that the data originates from.

3.1 Kernels and feature maps

Recalling what was said in the introduction, we now formalize the concept of sending inputs to more expressive features in a non-linear fashion. Often, the relevant properties of such mappings can be encoded in so-called *kernels*. Let, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

Definition 3.1. *A function $k\colon X \times X \to \mathbb{K}$ is called a kernel if there is a $\mathbb{K}$-Hilbert space F and a map $\Phi\colon X \to F$ such that*

$$k(x, x') = \langle \Phi(x'), \Phi(x) \rangle_F \tag{3.3}$$

for all $x, x' \in X$. In that case, Φ is called the feature map and F the feature space.

Given a concrete problem, feature maps should be constructued, so that the associated kernel acts as a measure of similarity. We can, for instance, consider data from the sphere $\mathbb{S}^{n-1}$: choose the feature maps to be the inclusion $\mathbb{S}^{n-1} \to \mathbb{R}^n$. Then, the kernel is given by $k(x, x') = \langle x', x \rangle_2$ and similarity is measured as spherical distance. Indeed, the kernel assumes its maximum

value 1 if the inputs are equal and its minimum value -1 if they are antipodal. Let us make these thoughts more precise.

A kernel k naturally induces a pseudometric on the underlying space X: if $\Phi\colon X \to F$ is a feature map, then the *kernel metric* d_k is defined by

$$d_k(x, x') \coloneqq \|\Phi(x) - \Phi(x')\|_H = \sqrt{k(x,x) + k(x',x') - 2k(x,x')}. \tag{3.4}$$

Note that this definition is independent of the feature map and that d_k is only a proper metric if Φ is injective but, nevertheless, generates a topology on X. Our first result characterizes continuity.

Lemma 3.2. *If $(X, \mathcal{T})$ is a topological space and k a kernel on X with feature map $\Phi\colon X \to F$, then tfae:*

1. *k is continuous w.r.t. the product space topology,*
2. *k is continuous separately in both arguments and $x \mapsto k(x,x)$ is continuous,*
3. *$\iota\colon (X, \mathcal{T}) \to (X, d_k)$ is continuous, and*
4. *$\Phi\colon (X, \mathcal{T}) \to F$ is continuous.*

Proof. (1) immediately implies (2).

If (2) holds, then $d_k(\cdot, x)\colon (X, \mathcal{T}) \to \mathbb{R}$ is continuous for each x. Therefore, the "open" ε-ball $\{x' \in X \mid d_k(x', x) < \varepsilon\}$ around x is truly open w.r.t. $\mathcal{T}$ and (3) follows.

Let (3) hold. It is easy to see that the map $\Phi'\colon (X, d_k) \to F$ is continuous by definition of the kernel metric. Therefore, so is $\Phi = \Phi' \circ \iota$ and (4) follows.

Finally, let (4) hold. In order to show (1) and close the equivalence, it suffices to observe that the product $\Phi \times \Phi\colon X \times X \to F \times F$ is continuous and $k = \langle \cdot, \cdot \rangle_F \circ (\Phi \times \Phi)$. □

We will mostly be interested in kernels with that satisfy the assumptions of the following easy result.

Lemma 3.3 (Real-valued kernels)**.** *Let $k\colon X \times X \to \mathbb{C}$ a kernel with feature map $\Phi\colon X \to F$. If k assumes only real values, then $F_0 \coloneqq F$ equipped with the inner product $\langle f, f' \rangle_{F_0} \coloneqq \Re \langle f, f' \rangle_F$ makes it a real Hilbert space and $\Phi\colon X \to F_0$ is a feature map for the $k_0 \coloneqq X \times X \to \mathbb{R}$.*

Real kernels can be nicely characterized via the spectra of their Gram matrices.

Lemma 3.4 (Matrix characterization)**.** *A function $k\colon X \times X \to \mathbb{R}$ is a real kernel iff the Gram matrices $[k(x_i, x_j)]_{ij}$ for all choices of n and $x_1, \ldots, x_n \in X$ are symmetric and psd.*

Proof. If k is a real-valued kernel, then all Gram matrices coming from it are psd simply by defintion as a real inner product in equation (3.3).

To the other end, let $F_{\text{pre}} \subseteq F$ be the vector subspace of linear combinations space of functions $\sum_{i=1}^{n} c_i k(\cdot, x_i)$ with complex coefficients for any n, and any choices of $c_i \in \mathbb{R}$ and x_i. Let $f \coloneqq \sum_{i=1}^{n} c_i k(\cdot, x_i)$ and $g \coloneqq \sum_{j=1}^{m} d_j k(\cdot, x'_j)$. Then,

$$\langle f, g \rangle \coloneqq \sum_i \sum_j c_i d_j k(x'_j, x_i) = \sum_j d_j f(x'_j) = \sum_i c_i g(x_i).$$

is independent of the concrete representations of f and g by the last two equalities. Furthermore, it is clearly bilinear and psd by assumption. We need to show positive definiteness. Take f as above but assume that $\langle f, f \rangle = 0$. For any x we have

$$|f(x)|^2 = |\langle f, k(\cdot, x) \rangle|^2 \leq \langle k(\cdot, x), k(\cdot, x) \rangle \langle f, f \rangle = 0 \tag{3.5}$$

by Cauchy-Schwarz (which holds in the psd case already) proving that $\langle \cdot, \cdot \rangle$ makes F_{pre} a pre-Hilbert space. Now, simply take the isometric embedding into the completion $\iota \colon F_{\text{pre}} \to F$ and observe that

$$k(x, x') = \langle k(\cdot, x'), k(\cdot, x) \rangle_{F_{\text{pre}}} = \langle \iota k(\cdot, x'), \iota k(\cdot, x) \rangle_F,$$

i.e. $x \mapsto \iota k(\cdot, x)$ is a feature map and, consequently, k a kernel. □

Practically speaking, one would construct a feature map either directly specific to the problem, or implicitly by constructing a kernel according to intuition. The latter way is to be preferred in most cases, as we will see that a concrete, easily computable expression for kernels is highly beneficial. Constructing a feature map usually involves computations in very high- or infinite-dimensional spaces. Therefore, we briefly summarize methods to construct new kernels from simple ones (without details).

- any map $f \colon \widetilde{X} \to X$ gives rise to a new kernel $\widetilde{k} = k(f, f)$ on $\widetilde{X}$ with the feature map $\widetilde{\Phi} = \Phi \circ f$,
- in particular, if $\widetilde{X} \subseteq X$, then $k|_{\widetilde{X} \times \widetilde{X}}$ is a kernel,
- finite conic combinations (i.e. linear combinations with non-negative coefficients) of kernels are kernels,
- the product $k_1 \cdot k_2$ of kernels k_1, k_2 on X_1, X_2 with feature maps $\Phi_1 \colon X_1 \to F_1, \Phi_2 \colon X_2 \to F_2$ is a kernel on $X_1 \times X_2$ with feature map $\Phi_1 \otimes \Phi_2 \colon X_1 \times X_2 \to F_1 \otimes F_2$, where $F_1 \otimes F_2$ is the Hilbert space tensor product,
- pointwise limits of real kernels are real kernels.

3.2 The reproducing property

In the proof of lemma 3.4 we constructed our feature space as a Hilbert space of functions so that the relation in equation (3.5) held: we were able to express function evaluation of *any* function of the space in terms of its inner product with the underlying kernel. We will now proceed to generalize this as the *reproducing property*.

Definition 3.5. *Let H be a $\mathbb{K}$-Hilbert space of functions $X \to \mathbb{K}$ where sum and scalar multiplication are defined as the usual pointwise ones. A function $k\colon X \times X \to \mathbb{K}$ is called a reproducing kernel of H if $k_x \coloneqq k(\cdot, x) \in H$ for any x and it satisfies the reproducing property*

$$f(x) = \langle f, k_x \rangle \tag{3.6}$$

for any x and $f \in H$. In that case k_x is called canonical feature map. If a Hilbert function space H has a reproducing kernel, we call it a reproducing kernel Hilbert spaces (RKHS, pl. RKHSs).

Proposition 3.6. *A reproducing kernel k is a kernel and satisfies the following version of Cauchy-Schwarz:*

$$|k(x, x')|^2 \leq k(x, x)k(x', x'). \tag{3.7}$$

Proof. A reproducing kernel on a space H is easily seen to be a kernel by the reproducing property (equation (3.6)) with the canonical feature map $x \mapsto k_x$:

$$k(x, x') = k_{x'}(x) = \langle k_{x'}, k_x \rangle_H. \tag{3.8}$$

The estimate follows immediately:

$$|k(x, x')|^2 = |\langle k_{x'}, k_x \rangle|^2 \leq \|k_{x'}\|^2 \|k_x\|^2 = k(x, x)k(x', x'). \tag{3.9}$$

□

RKHSs have the special property that norm convergence already implies pointwise convergence as functions. This property conveniently avoids many issues found for instance in L^p spaces, and makes them a precious tool theory and practice.

Proposition 3.7. *Let H be a Hilbert space of functions on X and for each $x \in X$ let $L_x\colon H \to \mathbb{K}$ be the evaluation functional defined by $L_x(f) = f(x)$. Then, H is an rkhs iff L_x is bounded. In that case, norm convergence implies pointwise convergence.*

Proof. Let first H be an rkhs with reproducing kernel k. Then,

$$|L_x(f)| = |f(x)| = |\langle f, k_x\rangle_H| \leq \|k_x\|_H \|f\|_H, \tag{3.10}$$

so $\|L_x\| \leq \|k_x\|_H < \infty$. In this case, if $f_n \to f$ converges in H-norm, then

$$|f_n(x) - f(x)| = |L_x(f_n - f)| \leq \|L_x\|_H \|f_n - f\|_H \to 0, \tag{3.11}$$

and the function f_n converges pointwise to f.

Conversely, let H be a Hilbert function space with bounded evaluation functionals L_x. Consider the function $k(x, x') = \langle L_x, L'_x\rangle_{H^*}$ defined as the inner product on the Hilbert space dual. Let $R\colon H^* \to H$ be the isometric, (anti-)linear Riesz isomorphism. Then

$$k_{x'}(x) = k(x, x') = \langle L_x, L_{x'}\rangle_{H^*} = \langle RL_{x'}, RL_x\rangle_H = L_x(RL_{x'}) = (RL_{x'})(x) \tag{3.12}$$

or simply $k(\cdot, x') = RL_{x'}$. The reproducing property follows:

$$f(x') = L_{x'} f = \langle f, RL_{x'}\rangle_H = \langle f, k_{x'}\rangle_H.$$

Thus, k is a reproducing kernel and H a rkhs. □

In the proof, we constructed a kernel using the boundedness of the evaluation functional and the Riesz (anti-)isomorphism. Surprisingly, this construction turns out to be unique.

Proposition 3.8. *Let H be a rkhs with reproducing kernel k and pick an onb $(e_i)_i$ of H. Then,*

$$k(x, x') = \sum_i \overline{e_i(x')} e_i(x). \tag{3.13}$$

In particular, if a Hilbert space of functions has a reproducing kernel, it must be unique.

Proof. A straightforward application of Parseval's identity and reproducing property:

$$k_{x'} = \sum_i (e_i \otimes e_i)(k'_x) = \sum_i \langle k_{x'}, e_i\rangle_H e_i = \sum_i \overline{e_i(x')} e_i. \tag{3.14}$$

This expression converges in H-norm, thus, by the previous result also pointwise the claim follows. □

The main reason that rkhss are a useful tool in machine learning and many other applied fields, is that many computations can be broken down to the reproducing kernel, which, usually, will be a reasonably simple and tractable function. The following statement gives a theoretical justification which will be made extensive use of in section 3.3.

Proposition 3.9. *Let H be a rkhs with reproducing kernel k. Define the vector subspace*

$$H_{pre} \coloneqq \left\{ f = \sum_{i=1}^{\infty} c_i k_{x_i} \colon x_i \in X, c_i \in \mathbb{K}, \#\{c_i \neq 0\} < \infty \right\} \subseteq H, \tag{3.15}$$

of all finite linear combinations of the canonical feature map. If we equip it with the inner product

$$\langle f, g \rangle_{pre} = \left\langle \sum_i c_i k_{x_i}, \sum_j d_j k_{y_j} \right\rangle_{pre} \coloneqq \sum_{i,j} c_i \overline{d_j} k(x_i, y_j), \tag{3.16}$$

then H is the completion of the pre-Hilbert space H_{pre}.

Proof. Observe that $\langle \cdot, \cdot \rangle_{\text{pre}}$ is nothing but the restriction of $\langle \cdot, \cdot \rangle_H$. Therefore, it suffices to prove density. If this wasn't the case, H_{pre} has a non-trivial orthogonal complement in H. This would imply that there are $f \in H_{\text{pre}}^{\perp}$ and $x \in X$ such that $f(x) \neq 0$. But since $k_x \in H_{\text{pre}}$

$$0 = \langle f, k_x \rangle = f(x) \neq 0. \tag{3.17}$$

□

We have already seen that reproducing kernels exist and are unique if the evaluation functional is bounded. We also know that reproducing kernels are kernels. The last result of this section establishes the other direction: a kernel has a unique rkhs.

Theorem 3.10 (Rkhss and kernels are 1-to-1)**.** Each rkhs H has the unique reproducing kernel k given by

$$k(x, x') = \sum_i \overline{e_i(x')} e_i(x), \tag{3.18}$$

where $(e_i)_i$ is any onb of H. Conversely, every kernel k with feature map Φ and feature space F gives rise to a unique (up to isomorphism) rkhs H for which it is a reproducing kernel. This space is given by

$$H = \{V(g) \coloneqq (x \mapsto \langle g, \Phi(x) \rangle_F) \colon g \in g\} \tag{3.19}$$

equipped with the norm

$$\|f\|_H \coloneqq \inf_{g \in V^{-1}(f)} \|g\|_F. \tag{3.20}$$

Proof. The first part of the theorem has already been proven in proposition 3.8.

We now prove that the space H as defined above is indeed a Hilbert space by establishing an isometry with a subspace of F via the operator $V\colon F \to H$. First, it is easily seen that $\ker V$ is closed (take a convergent sequence and notice that $V(\cdot)(x)$ is continuous) so that we have an orthogonal sum $H = \ker V \oplus F'$. The restriction $V|_{F'}$ is, therefore, injective and also surjective: if $f = V(g)$, decompose $g = g_0 + g'$, thus, $f = V|_{F'} g'$. Moreover, we find (with the same technique) that $\|f\|_H^2 = \|(V|_{F'})^{-1} f\|_{F'}^2$, which implies that $V|_{F'}$ is an isometric isomorphism and H is a Hilbert space.

Next, we show that k is a reproducing kernel of H. Since Φ is its feature map we have

$$k_x = \langle \Phi(x), \Phi(\cdot) \rangle_F = V\Phi(x) \in H. \tag{3.21}$$

Moreover, $\langle g, \Phi(x) \rangle = 0$ for any $g \in \ker V$ so that $\Phi(x) \in (\ker V)^\perp = F'$. The reproducing property follows by using the isometry of $V|_{F'}$

$$f(x) = \langle (V|_{F'})^{-1} f, \Phi(x) \rangle_F = \langle f, V|_{H'} \Phi(x) \rangle_H = \langle f, k_x \rangle_H. \tag{3.22}$$

Finally, we prove uniqueness. Observe first that the inner product on the dense subspace H_{pre} of H (from proposition 3.9) is only dependent on the kernel k so that it is a dense for all rkhss with this reproducing kernel. Now let H_1 and H_2 be two rkhss with reproducing kernel k. It clearly is enough to show that $H_1 \subseteq H_2$ is an isometric inclusion as the other inclusion will work the same way. Both contain the dense subset H_{pre}. We fix $f \in H_1$ together with a sequence $f_n \in H_{\text{pre}}$ such that $f_n \to f$ in H_1. This sequence is also contained in H_2 and is Cauchy also there, so that $f_n \to g$ in H_2 for some g. But norm convergence already implies point-wise convergence which implies $f = g \in H_2$. Finally, since norms coincide on the dense subset

$$\|f\|_{H_1} = \lim \|f_n\|_{H_{\text{pre}}} = \|f\|_{H_2}.$$

□

Note. Let us repass what happened: for any feature map we can find a surjective isometry from the associated feature space onto the unique RKHS, which, itself, is a feature space. Hence, RKHSs are the smallest feature spaces in this sense.

3.3 Empirical risk minimization via representer theorem

In this section we present the result that establishes the usefulness of RKHSs as hypothesis spaces in the statistical framework of Chapter 1. Roughly speaking, if we are given a finite number n of examples the *representer theorem* allows us to break down an *a priori* intractable least squares estimation of a function in an infinite-dimensional RKHS to simply finding n scalar weights.

While we keep all results in this section deterministic, the result will be the key to erm and random design analysis.

Throughout this section, fix a deterministic set $\{(x_i, y_i)\}_{i=1}^n \subseteq X \times \mathbb{K}^n$ of examples and a rkhs H with reproducing kernel k.

Next, we introduce two operators that will turn out to have very intuitive meanings and allow us to write much of the theory in a very concise way.

Proposition 3.11 (Sampling and realization)**.** *Let $\hat{S}\colon H \to \mathbb{K}^n$ be the sampling operator defined component-wise by $(\hat{S}f)_i \coloneqq f(x_i)$. Then, $\ker(\hat{S}) = \hat{H}^\perp$ where*

$$\hat{H} \coloneqq \operatorname{ran}(\hat{S}) = \operatorname{span}\{k_{x_i} \colon i = 1, \dots, n\} \subseteq H. \tag{3.23}$$

Moreover, its Hilbert space adjoint $\hat{S}^*\colon \mathbb{K}^n \to H$, the realization operator, satisfies $\operatorname{ran}(\hat{S}^*) = \hat{H}$ as it acts on $c = (c_i)_{i=1}^n \in \mathbb{K}^n$ by

$$\hat{S}^* c = \sum_{i=1}^n c_i k_{x_i} \tag{3.24}$$

Proof. We compute the kernel of $\hat{S}$. If $\hat{S}f = 0$, then for all components $i = 1, \dots, n$ we have $0 = f(x_i) = \langle f, k_{x_i}\rangle$ by the reproducing property. Thus, $f \in \hat{H}^\perp$ and the other direction is immediate.

The adjoint is found easily via the reproducing property as well:

$$\langle \hat{S}f, c\rangle_2 = \sum_i \langle f, k_{x_i}\rangle_H \overline{c_i} = \left\langle f, \sum_i c_i k_{x_i} \right\rangle_H = \langle f, \hat{S}^* c\rangle_H. \tag{3.25}$$

□

This notation allows us to write the empirical risk $\hat{R}_\alpha\colon H \to \mathbb{R}$ ($\alpha \geq 0$) in a very compact form:

$$n\hat{R}_\alpha(f) = \|\hat{S}f - y\|_2^2 + \alpha\|f\|_H^2, \tag{3.26}$$

where y is the column vector with entries y_i. The core result of this chapter can now be proven.

Proposition 3.12 (Representer theorem)**.** *Any solution f to the ERM problem $\inf_f \hat{R}_\alpha(f)$ admits the explicit form $f = \hat{S}^* c$ for some $c \in \mathbb{K}^n$.*

Proof. Decompose $f = \hat{f} + \hat{f}^\perp \in \hat{H} \oplus \hat{H}^\perp$. Since $\ker(\hat{S}) = \hat{H}^\perp$,

$$\hat{S}f = \hat{S}\hat{f} + \underbrace{\hat{S}\hat{f}^\perp}_{=0} = \hat{S}\hat{f}. \tag{3.27}$$

In particular, $\hat{R}_0(f) = \hat{R}_0(\hat{f})$ and the theorem is already proven in case $\alpha = 0$. Now, if $\hat{P}\colon H \to H$ is the orthogonal projection onto $\hat{H}$, we find

$$\|\hat{P}f\|_H^2 = \|\hat{f}\|_H^2 \leq \|\hat{f}\|_H^2 + \|\hat{f}^\perp\|_H^2 = \|f\|_H^2. \tag{3.28}$$

Therefore,

$$\inf_{g\in H} \hat{R}_\alpha(g) \leq \inf_{g\in \hat{H}} \hat{R}_\alpha(g) \tag{3.29}$$

$$= \inf_{g\in H} \hat{R}_0(\hat{P}g) + \alpha\|\hat{P}g\|_H^2 \tag{3.30}$$

$$\leq \inf_{g\in H} \hat{R}_0(g) + \alpha\|g\|_H^2 = \inf_{f\in H} \hat{R}_\alpha(g), \tag{3.31}$$

so that all of the above are actually equalities and $f \in \hat{H}$. □

From least squares estimation (Theorems 2.5 and 2.10) we already knew in theory how to solve the erm problem $\inf_{f\in H} R_\alpha(f)$: solutions are given by $\hat{f}_\alpha = (\hat{S}^*\hat{S} + \alpha I)^{-1}\hat{S}^* y$, where we exchange inverse by pseudoinverse if $\alpha = 0$ (this convention will hold throughout the work). However, while the operator $\hat{S}^*\hat{S}\colon H \to H$ is finite-rank, it is still operating in the infinite-dimensional space and, thus, not usable for implementation. The representer theorem resolves this issue and we are finally able to give a concrete description of the estimators computed by the kernel ridge(less) regression algorithm.

Theorem 3.13 (Kernel ridge(less) regression (krr))**.** For any $\alpha \geq 0$, mnls / rls solutions $\hat{f}_\alpha \in H$ to

$$\inf_{f\in H} \hat{R}_\alpha(f) \tag{3.32}$$

exist, are unique and can be expressed in the following two ways:

$$\hat{f}_\alpha = (\hat{\Sigma} + \alpha I)^\dagger \hat{S}^* y \tag{3.33}$$

$$= \hat{S}^*(\hat{K} + \alpha I)^\dagger y. \tag{3.34}$$

Here, $\hat{K} \coloneqq \hat{S}\hat{S}^*\colon \mathbb{K}^n \to \mathbb{K}^n$ and $\hat{\Sigma} \coloneqq \hat{S}^*\hat{S}\colon H \to H$.

The operators $\hat{K}$ (the kernel matrix) and $\hat{\Sigma}$ (the empirical, non-centered feature space covariance operator), as already $\hat{S}$ and $\hat{S}^*$, are empirical specializations of more meaningful probabilistic operators that will be studied and explored in-depth in the next section. For now the following fact, will be enough.

Lemma 3.14. *Let $\hat{P}\colon H \to H$ be the orthogonal projection onto $\hat{H}$. We have $\hat{\Sigma}\hat{\Sigma}^\dagger = \hat{P}$.*

Proof. Recall that $\hat{\Sigma}$ is self-adjoint. Let $f = f_1 + f_2$ where $f_1 \in \ker\hat{\Sigma} = \operatorname{ran}(\hat{\Sigma})^\perp$ and $f_2 \in \ker(\hat{\Sigma})^\perp = \operatorname{ran}\hat{\Sigma}$. Then,

$$\hat{\Sigma}\hat{\Sigma}^\dagger f = \hat{\Sigma}\hat{\Sigma}^\dagger(f_1 + f_2) = \hat{\Sigma}\hat{\Sigma}^{-1} f_1 = f_2 = \hat{P}f. \tag{3.35}$$

□

Proof. Equation (3.33) is immediate by our least squares discussion as discussed above and we only need to prove equation (3.34). By the representer theorem, we have that $\hat{f}_\alpha = \hat{P}\hat{f}_\alpha$. The rest is a computation:

$$\hat{f}_\alpha = \hat{P}\hat{f}_\alpha = \hat{\Sigma}\hat{\Sigma}^\dagger(\hat{\Sigma} + \alpha I)^\dagger \hat{S}^* y \tag{3.36}$$

$$= \hat{\Sigma}\left((\hat{\Sigma} + \alpha I)\hat{\Sigma}\right)^\dagger \hat{S}^* y \tag{3.37}$$

$$= \hat{\Sigma}(\hat{\Sigma}^2 + \alpha\hat{\Sigma}))^\dagger \hat{S}^* y \tag{3.38}$$

$$= \hat{\Sigma}\left(\hat{S}^*(\hat{K} + \alpha I)\hat{S}\right)^\dagger \hat{S}^* y \tag{3.39}$$

$$= \hat{S}^* \underbrace{(\hat{S}\hat{S}^\dagger)}_{=I}(\hat{K} + \alpha I)^\dagger \underbrace{(\hat{S}^*)^\dagger \hat{S}^*}_{=I} y \tag{3.40}$$

$$= \hat{S}^*(\hat{K} + \alpha I)^\dagger y. \tag{3.41}$$

The last equality follows from the linear independence assumption. □

3.4 Learning operators and risk

In order to incorporate rkhss as hypothesis spaces into the statistical learning setup from chapter 1, we introduce and study operators to facilitate working with the squared loss. The principal one acts as a restriction of functions defined on the whole space onto L^2-equivalence classes with support defined by a probability measure; think about the unknown marginal distribution or the empirical distribution. The notions introduced in this section can be also be found in Steinwart and Christmann 2008 but vast generalizations are available (see e.g. Carmeli, De Vito, and Toigo 2006).

Lemma 3.15. *If X is a separable space and k a continuous kernel then the associated rkhs is separable.*

Proof. By the previous lemma we know that the feature map Φ associated to k is continuous. This implies that $\Phi(X)$ is separable and, therefore, also H_{pre} as the span. The former space is dense and the claim follows. □

Lemma 3.16. *Let k be a kernel on a measurable space X and let H be the associated rkhs. Then k_x is measurable for all x iff all $f \in H$ are.*

Proof. Since $k_x \in H$, the sufficiency is immediate. If, conversely, all k_x are measurable, then so are all functions in H_{pre} (as in the proof of theorem 3.10). Fix now $f \in H$ and a sequence $(f_n)_n \subset H_{\text{pre}}$ such that $f_n \to f$ in H. Since this implies point-wise convergence, and all f_n are measurable, so is f. □

From now on we fix a measurable space X equipped with a probability measure μ (think about the true or the empirical marginal from chapter 1). and a separable rkhs H with measurable, real-valued kernel $k\colon X \times X \to \mathbb{R}$. By $L^2(\mu)$ we denote $L^2(X, \mu)$.

Proposition 3.17 (Restriction operator)**.** *Assuming that*

$$\|k\|_{L^2(\mu)} \coloneqq \left(\int_X k(x,x)\, d\mu(x) \right)^{1/2} < \infty$$

is finite, all members of H are square-integrable and the canonical inclusion $H \to L^2(\mu)$ sending functions to their equivalence classes is continuous with operator norm bounded by $\|k\|_{L^2(\mu)}$.

Proof. Fix an $f \in H$. Recall that $\|k_x\|_H = \sqrt{k(x,x)}$. Both claims follow from a simple application of Hoelder's inequality:

$$\begin{aligned} \|f\|^2_{L^2(\mu)} &= \int_X |f(x)|^2 \, d\mu(x) = \int_X |\langle f, k_x \rangle_H|^2 \, d\mu(x) \\ &\leq \|f\|^2_H \int_X k(x,x)\, d\mu(x) = \|f\|^2_H \, \|k\|_{L^2(\mu)}. \end{aligned}$$

□

The intuitive meaning of this operator links our rkhss closely to the probabilistic model from chapter 1. Indeed, it allows to consider functions $f \in H$ as random variables $y = S_\mu f$ modeling the response.

The following definition introduces four operators that will be fundamental for our analysis.

Definition 3.18. *Let $S_\mu \colon H \to L^2(\mu)$ be the inclusion operator from the last proposition.*

- *$S_\mu \colon H \to L^2(\mu)$ will be referred to as the restriction operator.*
- *Its (Hilbert space) adjoint $S^*_\mu \colon L^2(\mu) \to H$ is called the extension operator.*
- *The operator $K_\mu \coloneqq n S_\mu S^*_\mu \colon L^2(\mu) \to L^2(\mu)$ is called the kernel operator.*
- *Finally, $\Sigma \coloneqq S^*_\mu S_\mu \colon H \to H$ is the (not centered) covariance operator in feature space.*

The remaining part of this section will establish essential properties of these operators, mainly compactness. Note that the restriction operator S_μ looks like an innocent inclusion but can be highly non-injective: it sends everywhere defined functions to equivalence classes defined only almost everywhere even on the support of μ. Suppose, for example, that μ is an empirical measure of n observations. In that case $L^2(\mu) \cong \mathbb{R}^n$ can be naturally identified, so that the linear operator $S_\mu \colon H \to \mathbb{R}^n$ sends from a usually inifinite dimensional Hilbert space onto a finite dimensional one (keep this example in mind for section 3.3). Indeed, it is a well-known fact that S_μ is injective iff its adjoint $S^*_\mu \colon L^2(\mu) \to H$ has dense range. Conversely, S^*_μ is injective if and only if S_μ has dense image, i.e. H is dense in $L^2(\mu)$.

Let us now explicitly compute the extension operator.

Proposition 3.19 (Extension operator)**.** *Let H be separable. The extension operator $S_\mu^*\colon L^2(\mu) \to H$ is bounded and given pointwise by*

$$S_\mu^* g(x') \coloneqq \int_X k(x', x)\, g(x)\, d\mu(x). \tag{3.42}$$

We use basic properties of Bochner integration in the following proof. See for instance the relevant appendix in Steinwart and Christmann 2008.

Proof. First, the adjoint S_μ^* is bounded since S_μ is. Hoelder and the kernel version of Cauchy-Schwarz (see equation (3.7) above) together yield

$$\begin{aligned}\int_X |k(x', x)g(x)|\, d\mu(x) &\leq \sqrt{k(x', x')} \int_X \sqrt{k(x, x)}|g(x)|\, d\mu(x) \\ &\leq \sqrt{k(x', x')}\, \|k\|_{L^2}\, \|g\|_{L^2}.\end{aligned}$$

Thus, the rhs of equation (3.42) exists pointwise. Before showing that this is really the adjoint, we need to establish first that it is an element of H employing Bochner integration theory. Note that Cauchy-Schwarz shows as well that $x \mapsto \|\Phi(x)g(x)\|_H$ is integrable:

$$\begin{aligned}\int_X \|\Phi(x)g(x)\|_H\, d\mu(x) = \int_X g(x)\|\Phi(x)\|_H\, d\mu(x) &= \int_X g(x)\sqrt{k(x, x)}\, d\mu(x) \\ &\leq \|k\|_{L^2}\, \|g\|_{L^2}.\end{aligned}$$

By Bochner's integration criterion we know that the above function is Bochner integrable, so $\int_X \Phi g\, d\mu \in H$. Since Bochner integration commutes with bounded linear operators (APPENDIX???), in our case with $\langle \cdot, \Phi(x)\rangle$, we find

$$S_\mu^* g(x) = \int_X k_x\, g\, d\mu = \left\langle \int_X \Phi g\, d\mu, \Phi(x) \right\rangle$$

and, consequently, $S_\mu^* g \in H$ since evaluation happens in the lines of the reproducing property, i.e. the construction equation (3.42) is well-defined.

Finally, we show that this is in fact the adjoint of restriction. Let $g \in L^2(\mu)$ and $f \in H$. Then,

$$\begin{aligned}\langle g, S_\mu f\rangle_{L^2} &= \int_X g\, S_\mu f\, d\mu = \int_X g(x)\, \langle f, k_x\rangle_H\, d\mu(x) \\ &= \Bigg\langle f, \underbrace{\int_X k_x\, g(x)\, d\mu(x)}_{S_\mu^* g} \Bigg\rangle_H = \langle f, S_\mu^* g\rangle_H.\end{aligned}$$

□

These operators admit useful compactness properties that will be crucial to our analysis.

Proposition 3.20 (Compactness of extension and restriction)**.** *We assume that both H and $L^2(X,\mu)$ are separable. and $\|k\|_{L^2(\mu)} < \infty$. Then, both restriction S_μ and restriction S_μ^* are Hilbert-Schmidt operators with $\|S_\mu\|_{HS} = \|k\|_{L^2(\mu)} = \|S_\mu^*\|_{HS}$. Moreover, both the kernel operator $K_\mu \colon L^2(\mu) \to L^2(\mu)$ and the feature space covariance $\Sigma_\mu \colon H \to H$ are positive, self-adjoint and trace-class with $\operatorname{tr} K_\mu, \operatorname{tr} \Sigma_\mu \leq \|k\|^2_{L^2(\mu)}$.*

Proof. We start with the extension operator S_μ. Fix an orthonormal basis $(e_i)_{i\in I}$ (I countable by assumption) of H. Then,

$$\begin{aligned}\|S_\mu\|_{\mathrm{HS}}^2 &= \sum_i \|S_\mu e_i\|_{L^2(\mu)}^2 \\ &= \int_X \sum_i |S_\mu e_i(x)|^2 \, d\mu(x) = \int_X \sum_i |e_i(x)|^2 \, d\mu(x) = \|k\|^2_{L^2(\mu)},\end{aligned}$$

where the last equation follows from equation (3.13). Thus, S_μ is Hilbert-Schmidt. Recalling that a compact operator is Hilbert-Schmidt iff its adjoint is, the same holds for S_μ^* with equal norm.

Since K_μ and Σ_μ are compositions of Hilbert-Schmidt operators, they are trace-class themselves and, in particular, compact. Note that they are clearly positive and self-adjoint as the compositions of a compact operator with its adjoint. By the previous results both $\operatorname{tr} K_\mu$ and $\operatorname{tr} \Sigma_\mu$ are bounded by $\|S_\mu\|\|S_\mu^*\| = \|k\|^2_{L^2(\mu)}$. □

4 An all-inclusive study model

This final chapter proposes a simple model that allows to study all many capacity controls in a unified fashion:

- arbitrarily many features via cut-off Fourier coefficients
- function space norm penalty via Tikhonov
- intrinsic kernel smoothness via Sobolev spaces

in problem setups that allow for regression functions from very specific hypothesis space, noise, dimensionality and submanifolds. All of these are computationally tractable even on weak machines, and there are computable risk decompositions available. Thus, this model seems to provide a solid foundation to study the phenomena of the modern regime. The idea comes from De Vito 2022 and our contribution is only to provide implementations and experiments as proof of concepts. A related theoretical work was published by Potts and Schmischke 2021.

Content-wise we will briefly discuss one more capacity parameter: the number of parameters. We will observe that even in 1 dimension with this simple model we can find the phenomenon of *double descent*.

4.1 Mercer kernels on the torus

We denote by $\mathbb{T}^d \coloneqq \mathbb{R}^d/\mathbb{Z}^d$ the d-dimensional torus parameterized by $[0,1]^d$ and equipped with the induced Lebesgue measure dx so that the measure of the torus is 1. We recall from Fourier theory that the family of functions

$$e_{\mathbf{k}} \colon \mathbb{T}^d \to \mathbb{C}, \quad x \mapsto \exp(2\pi i \mathbf{k} \cdot x) \tag{4.1}$$

with $\mathbf{k} \in \mathbb{Z}^d$ is an ONB of $L^2(\mathbb{T}^d)$ (see for instance Chavel, Randol, and Dodziuk 1984, Chapter II.2). Any function $f \in L^2(\mathbb{T}^d)$ can be represented as $f = \sum_{\mathbf{k}} \omega_{\mathbf{k}} e_{\mathbf{k}}$ with

$$\omega_{\mathbf{k}} = \int_{\mathbb{T}^d} f(x) e_{\mathbf{k}}(x)\, dx. \tag{4.2}$$

We study a class of kernels on the torus.

Definition 4.1. *A toral Mercer kernel is any kernel k of the form*

$$k(x, x') = \sum_{\mathbf{k}} \lambda_{\mathbf{k}} e_{\mathbf{k}}(x') \overline{e_{\mathbf{k}}(x)}. \tag{4.3}$$

such that the sequence of coefficients satisfies

$$\lambda_{\mathbf{k}} \geq 0, \quad \lambda_{\mathbf{k}} = \lambda_{-\mathbf{k}}, \quad \sum_{\mathbf{k}} \lambda_{\mathbf{k}} < \infty. \tag{4.4}$$

The convergence of the sum is absolute and uniform by the conditions on the coefficients. Note that these "kernels" are indeed kernels (by independence of the $e_{\mathbf{k}}$ and lemma 3.4) and that they are *stationary* in the sense that $k(x, x') = k_0(x - x')$. For this reason we sometimes use $k = k_0$ as a function of a single variable. By the symmetry, we see immediately that k only assumes real values. A simple consequence of equation (3.7) is that $k(x, x') \leq k(0) = \sum_{\mathbf{k}} \lambda_{\mathbf{k}}$.

Their RKHSs can be characterized very precisely using *Mercer's theorem*.

Theorem 4.2 (Formulation from Theorem 4.49 in Steinwart and Christmann 2008)**.** Let X be a compact metric space with a Borel measure μ with full support and k a continuous kernel on X. Let $(e_i, \lambda_i)_{i \in I}$ be the eigensystem of the kernel operator $K_\mu \colon L^2(X, \mu) \to L^2(X, \mu)$ Then, the functions $\widetilde{e}_i \coloneqq \lambda_i^{-1} S_\mu^* e_i$ give rise to an ONB $\{\sqrt{\lambda_i}\widetilde{e}_i\}_{i \in I}$ of H.

The proof is skipped but can be found in the reference and references therein. Note that we have almost surely $e_i = \lambda^{-1} K_\mu e_i = \widetilde{e}_i$, so that we identify the two with slight abuse of notation.

Proposition 4.3. *The RKHS H associated to a toral Mercer kernel k is a space of continuous functions on $\mathbb{T}^d$ and the inclusion $H \to C(\mathbb{T}^d)$ is compact. Letting*

$$\Lambda \coloneqq \{\mathbf{k} \mid \lambda_{\mathbf{k}} \neq 0\}, \tag{4.5}$$

we find that the family $\{\sqrt{\lambda_{\mathbf{k}}} e_{\mathbf{k}}\}_{\mathbf{k} \in \Lambda}$ is an ONB for H. In particular, for $f \in H$

$$\|f\|_H^2 = \sum_{\mathbf{k} \in \Lambda} \frac{|\langle f, e_{\mathbf{k}} \rangle_{L^2}|^2}{\lambda_{\mathbf{k}}}. \tag{4.6}$$

Proof. Since the k is continuous, we know by lemma 3.2 that so is the canonical feature map $\Phi : x \mapsto k_x$. Therefore, the image $\Phi(\mathbb{T}^d)$ is compact and, hence, so is the space $(\mathbb{T}^d, d_k)$ with the kernel pseudometric. If $C(X, d_k)$ is the space of functions that are continuous w.r.t. the pseudometric topology, we have

$$|f(x) - f(x')| = |\langle f, \Phi(x) - \Phi(x') \rangle| \leq \|f\|_H d_k(x, x'), \tag{4.7}$$

which means that f is Lipschitz continous on $(\mathbb{T}^d, d_k)$. Therefore, the unit ball B_H of H is equicontinuous and also $\|\cdot\|_\infty$-bounded ($\mathbb{T}^d$ is compact), by which it follows that $\overline{B_H}$ is compact in $C(X, d_k)$ by Arzela-Ascoli. By composing with the continuous identity, we see that the inclusion $H \to C(\mathbb{T}^d)$ is compact. The rest of the statement is Mercer's theorem. □

4.2 Unified risk decomposition

Let us come to the random design risk analysis for kernelized MNLS and RLS estimators. While the result will mainly feel similar to proposition 2.11, the main difference is that we know allow for *misspecification*, i.e. regression function that are not part of the hypothesis space giving rise to additional terms. Fix a toral Mercer kernel k and the associated RKHS H. We adopt most of the notation of the

Proposition 4.4. *Let the noise be indepdent of X (as usual) and let P_Λ be the projection from L^2 onto the closed subspace $\overline{\operatorname{span}}\{e_{\mathbf{k}} \mid \mathbf{k} \in \Lambda\}$. Suppose there is $f_H \in H$ such that $Sf_H = P_H f^*$, and let $\hat{M}$ be such that $\hat{Y} = \hat{S}f_H + \hat{M} + \hat{\varepsilon}$. Then, we obtain a risk decomposition for the MNLS/RLS estimator $\hat{f}_\alpha$, $\alpha \geq 0$*

$$\mathbb{E}[R(\hat{f}_\alpha) \mid \hat{X}^n] = \hat{\mathbf{B}} + \hat{\mathbf{V}} + \hat{\mathbf{M}} + \sigma^2 \tag{4.8}$$

with bias term

$$\hat{\mathbf{B}} = \|S\hat{Q}(\alpha)f_H\|_{L^2}^2, \qquad \hat{Q}(\alpha) = I - (\hat{\Sigma} + \alpha I)^{-1}\hat{\Sigma}, \tag{4.9}$$

(as usual inverse is replaced by pseudoinverse if $\alpha = 0$), variance term

$$\hat{\mathbf{V}} = \frac{\sigma^2}{n} \operatorname{trace}\left(\Sigma(\hat{\Sigma} + \alpha I)^{-2}\hat{\Sigma}\right), \tag{4.10}$$

and the new misspecification term

$$\hat{\mathbf{M}} = \|(I - P_H)f^*\|_{L^2}^2 + n^{-2}\|S(\hat{\Sigma} + \alpha I)^{-1}\hat{S}^*\hat{M}\|_{L^2}^2 \tag{4.11}$$

$$- \frac{2}{n}\Re\langle S\hat{Q}(\alpha)f_H, S(\hat{\Sigma} + \alpha I)^{-1}\rangle\hat{S}^*\hat{M}\rangle_{L^2} \tag{4.12}$$

The proof is completely analogous to those of proposition 2.11 and we skip it. With our trick of using the empirical kernel instead of the empirical covariance we can reformulate the whole risk decomposition into a completely computable finite dimensional form.

Note that misspecification term is another form of bias: in some sense we can see the term $\hat{\mathbf{B}}$ as the bias due to explicit regularization (least squares and norm penalty) while $\hat{\mathbf{M}}$ is the bias due to implicit regularization (choice of hypothesis space, e.g. kernel parameters).

4.3 Double descent

Finally, we demonstrate experimentally that the model is powerful enough to display the phenomena of modern statistical learning.

In the following experiments we are using three types of kernels in case $d = 1$ (i.e. on the circle) that we define here first:

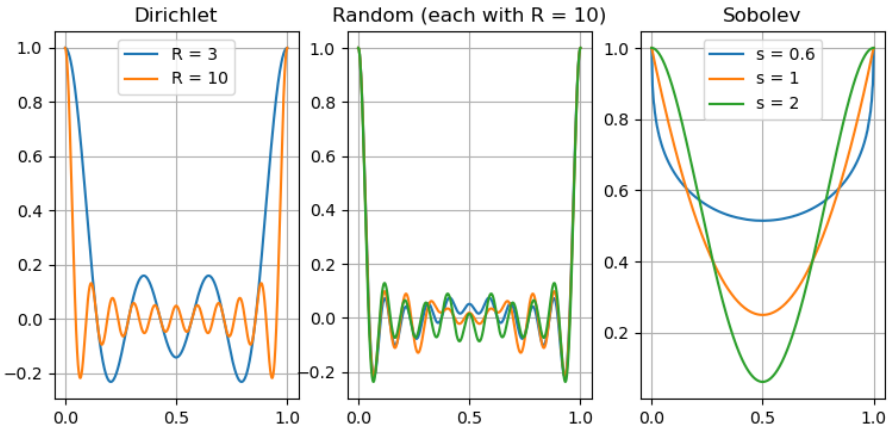

Figure 4.1: Visualizations of toral Mercer kernels with $d = 1$ on plotted against the parameterization $[0, 1]$.

1. **Dirichlet kernel**: defined by $\lambda_k = 1$ for all $k \in \Lambda = [-R, R] \cap \mathbb{Z}$. The kernel parameter R controls the complexity directly as fitting with this kernels offers $N = 2R + 1$ features to be trained.

2. **Random kernel**: Similar to the Dirichlet kernel with $\Lambda = [-R, R] \cap \mathbb{Z}$ but with random λ_k sampled from some uniform distribution ($k \geq 0$ only, since $k < 0$ must be symmetric)

3. **Sobolev kernel**: defined via $\lambda_k = |k|^{-2s}$ for real $s > \frac{1}{2}$ and $k \in \Lambda = \mathbb{Z}$. The naming is not by coincidence: they do correspond to true Sobolev spaces with real smoothness index. A detailed derivation can be found in De Vito, Mücke, and Rosasco 2021.

See figure 4.1 for visualizations.

As discovered in Belkin et al. 2019; Belkin, Hsu, and Xu 2020, the *double descent* phenomenon appears with toral kernels. Here, we simply pick the Dirichlet kernel and we plot in the number of available features $N = 2R + 1$. Take a look at figure 4.2. Since there is no noise, the variance term will be zero, and we are left with implicit and explicit bias. Classical theory predicts the regime $n < p$. As soon as we approach the "critical line" $n = p$ the error explodes: this is the classical understanding of overfitting. However, as we further increase the parameter count p the error starts decreasing again even for the misspecifed Dirichlet kernel. A look at the bias terms reveals that this is due to the decrease of the implicit. Then, however, the explicit bias takes over and seems to converge. In case of the Sobolev kernels the second descent strength is even stronger to the point that we get better as the parameter count increases. This is in strong resemblence to Neural nets that perform best when they are highly overparameterized, keeping in mind that there are theoretical

connections between ReLU neural nets and Sobolev spaces as hinted to in the introduction.

In the future, this model shall be further developed (high dimensions, subtori, faster computations, ...) and should undergo an in-depth theoretical study which seems to be very feasible due to the simple structure of the involved components.

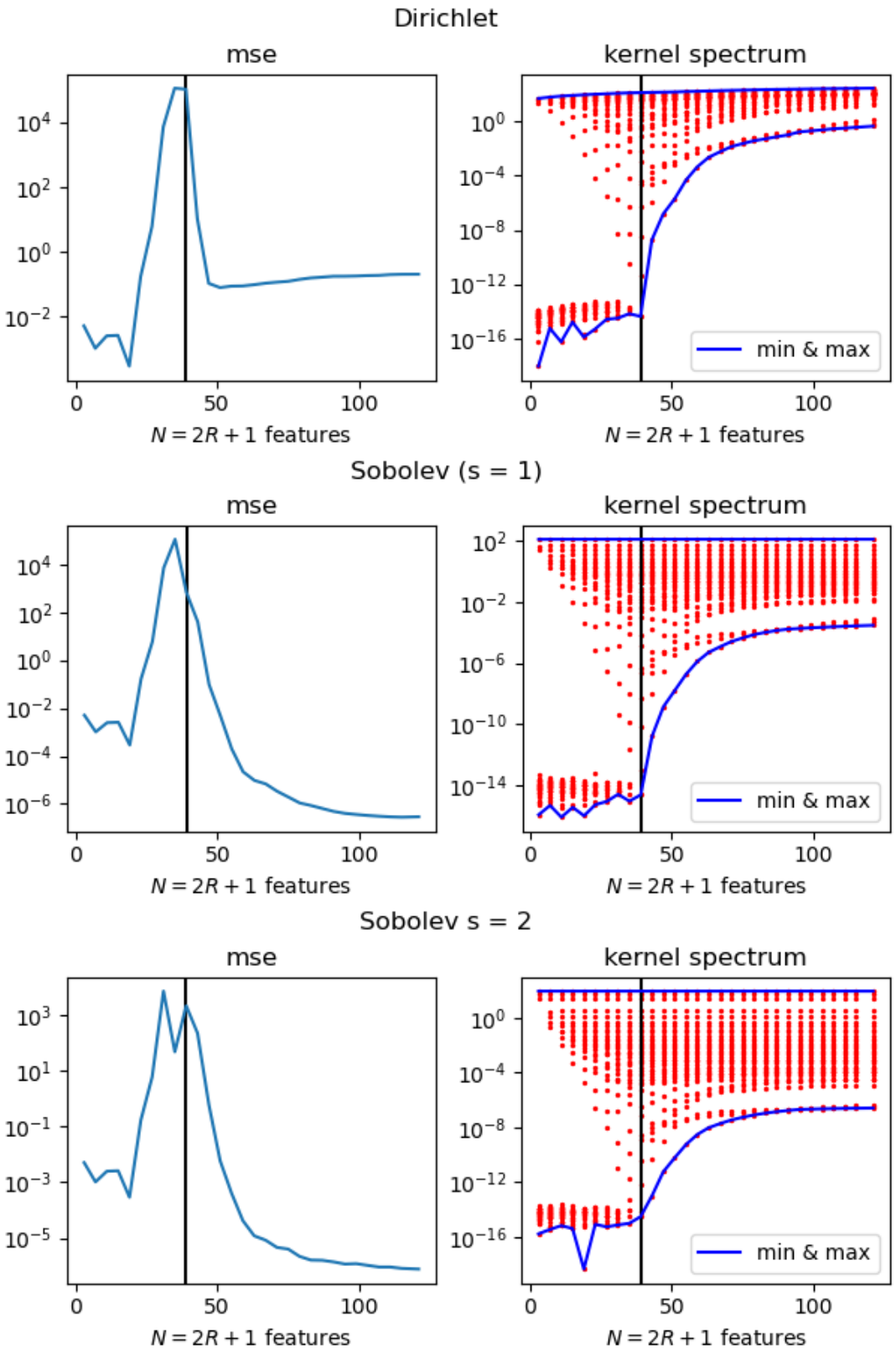

Figure 4.2: Toral Mercer kernel regression with $N = 2R + 1$ features. The first row is a Dirichlet kernel and the next two are band-limited Sobolev kernels of smoothnesses $s = 1, 2$ cut at frequency N. Tested on $n = 39$ training points with no noise and $\alpha = 0$. The regression function is Sobolev kernel with $s = 1$ itself. The two left plots are showing MSE while the right one shows the spectrum of the kernel matrix. The black line indicates the critical line where the number of parameters equals the number of training points.

References

Bartlett, Peter L et al. (2020). "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48, pp. 30063–30070.

Belkin, Mikhail, Daniel Hsu, and Ji Xu (2020). "Two models of double descent for weak features". In: *SIAM Journal on Mathematics of Data Science* 2.4, pp. 1167–1180.

Belkin, Mikhail, Siyuan Ma, and Soumik Mandal (2018). "To understand deep learning we need to understand kernel learning". In: *International Conference on Machine Learning*. PMLR, pp. 541–549.

Belkin, Mikhail et al. (2019). "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854.

Billingsley, Patrick (1995). *Probability and measure*. Third. Wiley Series in Probability and Mathematical Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, pp. xiv+593. ISBN: 0-471-00710-2.

Carmeli, Claudio, Ernesto De Vito, and Alessandro Toigo (2006). "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem". In: *Analysis and Applications* 4.04, pp. 377–408.

Chavel, I., B. Randol, and J. Dodziuk (1984). *Eigenvalues in Riemannian Geometry*. ISSN. Elsevier Science. ISBN: 9780080874340. URL: https://books.google.it/books?id=0v1VfTWuKGgC.

Chen, Lin and Sheng Xu (2020). "Deep neural tangent kernel and laplace kernel have the same rkhs". In: *arXiv preprint arXiv:2009.10683*.

Clason, Christian (Feb. 8, 2021). *Regularization of Inverse Problems*. DOI: 10.48550/arXiv.2001.00617. arXiv: 2001.00617[cs,math]. URL: http://arxiv.org/abs/2001.00617 (visited on 11/15/2022).

Cucker, Felipe and Steve Smale (2002). "On the mathematical foundations of learning". In: *Bull. Amer. Math. Soc. (N.S.)* 39.1, pp. 1–49. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-01-00923-5. URL: https://doi.org/10.1090/S0273-0979-01-00923-5.

De Vito, Ernesto (2022). personal communication.

De Vito, Ernesto, Nicole Mücke, and Lorenzo Rosasco (2021). "Reproducing kernel Hilbert spaces on manifolds: Sobolev and diffusion spaces". In: *Anal. Appl. (Singap.)* 19.3, pp. 363–396. ISSN: 0219-5305. DOI: 10.1142/S0219530520400114. URL: https://doi.org/10.1142/S0219530520400114.

Dudley, R. M. (2002). *Real analysis and probability*. Vol. 74. Cambridge Studies in Advanced Mathematics. Revised reprint of the 1989 original. Cambridge

University Press, Cambridge, pp. x+555. ISBN: 0-521-00754-2. DOI: 10.1017/CBO9780511755347. URL: https://doi.org/10.1017/CBO9780511755347.

Liang, Tengyuan and Alexander Rakhlin (2020). "Just interpolate: Kernel "ridgeless" regression can generalize". In.

Mallinar, Neil et al. (2022). "Benign, tempered, or catastrophic: A taxonomy of overfitting". In: *arXiv preprint arXiv:2207.06569*.

Mourtada, Jaouad and Lorenzo Rosasco (2022). "An elementary analysis of ridge regression with random design". In: *Comptes Rendus. Mathématique* 360.G9, pp. 1055–1063.

Pagliana, Nicolò et al. (2020). "Interpolation and learning with scale dependent kernels". In: *arXiv preprint arXiv:2006.09984*.

Potts, Daniel and Michael Schmischke (2021). "Approximation of high-dimensional periodic functions with Fourier-based methods". In: *SIAM Journal on Numerical Analysis* 59.5, pp. 2393–2429.

Rakhlin, Alexander and Xiyu Zhai (2019). "Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon". In: *Conference on Learning Theory*. PMLR, pp. 2595–2623.

Rudi, Alessandro, Luigi Carratino, and Lorenzo Rosasco (2017). "Falkon: An optimal large scale kernel method." In: *Advances in neural information processing systems 30*.

Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. Information Science and Statistics. Springer New York. ISBN: 9780387772424. URL: https://books.google.it/books?id=HUnqnrpYt4IC.

Tsigler, Alexander and Peter L Bartlett (2020). "Benign overfitting in ridge regression". In: *arXiv preprint arXiv:2009.14286*.

Vapnik, Vladimir N. (2000). *The nature of statistical learning theory*. Second. Statistics for Engineering and Information Science. Springer-Verlag, New York, pp. xx+314. ISBN: 0-387-98780-0. DOI: 10.1007/978-1-4757-3264-1. URL: https://doi.org/10.1007/978-1-4757-3264-1.

Vershynin, Roman (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics 47. Cambridge University Press.