Theory of nonparametric regression

Alex H Williams
Center for Neural Science, New York University
Center for Computational Neuroscience, Flatiron Institute

Last Updated: February 8, 2024

These notes are from an informal lecture series being delivered in January 2024. They provide only a high-level overview of a sophisticated and well-developed literature, but I am sharing them in the hope that they may be a useful entry point to others. The notes should be accessible to anyone who has taken a full introductory course on machine learning or statistical modeling.

Contents

1 Overview of nonparametric regression

1.1 Problem setup and assumptions

In these notes we will consider the problem of predicting a scalar random variable Y from a multi-dimensional random variable X taking values on some space $\mathcal{X}$. We call Y the dependent variable and X the independent variable(s). In a typical setting, we observe n paired observations drawn from some joint distribution $P_{X,Y}$ over $\mathcal{X} \times \mathbb{R}$. Formally,

$$(X_i, Y_i) \sim P_{X,Y} \qquad \text{independently for } i = 1, \ldots, n. \tag{1}$$

Given these data, we would like to predict the dependent variable, given some value of the independent variable. That is, we want to build a model of the conditional distribution $P(Y \mid X = \boldsymbol{x})$, for any chosen value of $\boldsymbol{x} \in \mathcal{X}$.

We have set up the problem by assuming the inputs $X_1, \ldots, X_n$ are random. Statisticians call this *random design regression*. It is also possible to set up the problem by assuming the inputs are fixed ahead of time by the experimentalist (i.e. not random). Statisticians call this *fixed design regression*. This can change aspects of the theory, but we will mostly not pay attention to these details as they are rather subtle. A nice aspect of random design regression is that there is a natural way to measure how the model performs on future datapoints, since we can sample $(X_*, Y_*) \sim P_{X,Y}$.

Throughout, we will assume that the true conditional distribution $P(Y \mid X = \boldsymbol{x})$ always has a well-defined and finite mean and variance. Thus, we can define:

$$f(\boldsymbol{x}) = \mathbf{E}[Y \mid X = \boldsymbol{x}] \tag{2}$$

which we call the *target function*; it is a bounded function mapping $\mathcal{X} \mapsto \mathbb{R}$. Furthermore, we can define:

$$\epsilon_i = Y_i - \mathbf{E}[Y \mid X_i] \qquad \text{for } i = 1, \ldots, n \tag{3}$$

which are, by construction, mean-zero random variables. Putting these together, we can reformulate eq. (1) as equivalent to:

$$\begin{aligned} X_i &\sim P_X && \text{independently for } i = 1, \ldots, n \\ \epsilon_i \mid X_i &\sim P_\epsilon(X_i) && \text{independently for } i = 1, \ldots, n \\ Y_i &= f(X_i) + \epsilon_i && \text{for } i = 1, \ldots, n \end{aligned} \tag{4}$$

Note that we have allowed for the possibility that noise distribution P_ϵ can vary as a function of X. This is sometimes called a "heteroschedastic" model. If we assume the noise is independent of X, the second line could be replaced with $\epsilon_i \sim P_\epsilon$ and called a "homoschedastic" model. While this assumption greatly simplifies analysis, it is often not justified in practical circumstances.

We have several modeling goals. Our primary goal is to estimate the target function f from data. This can be done from a frequentist or Bayesian perspective,[1] and we will cover both

[1] I recommend Michael I. Jordan's lecture "Bayesian or Frequentist, Which Are You?" for a clear and succinct overview of the distinction, which is sadly muddled by a lot of subpar online content. The lecture is available at: https://www.youtube.com/watch?v=HUAE26lNDuE

approaches in due course. A second goal will be to quantify our uncertainty in our estimate of f, either through confidence intervals (frequentist) or posterior credible intervals (Bayesian). Finally, we may be interested in estimating higher-order moments of the conditional distribution, in particular the variance $\mathbf{Var}[\epsilon \mid X = \boldsymbol{x}]$.

Note that we have made very minimal assumptions about the underlying data distribution. Indeed, we have not assumed that f has any parametric form (e.g. linear or polynomial). We similarly have not assumed that the "noise" terms $\epsilon_1, \ldots, \epsilon_n$ have any parametric form (e.g. Gaussian). In a nutshell, this is why statisticians refer to our problem of interest as **nonparametric regression**.

We will mostly consider the space of independent variables, $\mathcal{X}$, to be a bounded subset of d-dimensional Euclidean space, $\mathbb{R}^d$. However, in some sections we will consider this space to be a non-Euclidean manifold, but in all cases we will stick to the convention that $\dim(\mathcal{X}) = d$.

1.2 Smoothness assumptions on f

Although we do not assume that our target function, f, has a parametric form, we clearly need to make *some* assumptions about the problem for it to be tractable. Indeed, we can come up with many discontinuous functions that are effectively impossible to estimate. For example, suppose $\boldsymbol{x} \in [0, 1]$ and choose N arbitrary constants $c_1, \ldots, c_N$ and define:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} c_i I\left[\frac{i-1}{N} \leq \boldsymbol{x} < \frac{i}{N}\right] \tag{5}$$

where $I[\cdot]$ is the indicator function. Thus, we've constructed a function with N discontinuous "steps" between arbitrary values. If the number of steps is much larger than the number of samples, n, our prediction error on heldout data can be very large since it is likely that our unseen data will fall into an interval that was not observed in the training set.

While many functions are intractable to estimate, it is intuitively possible to estimate $f(\boldsymbol{x})$ when it is "smooth." If the output value $f(\boldsymbol{x})$ changes very slowly as we vary $\boldsymbol{x}$, it will suffice to sample the input space coarsely—the problem in this case is "easy" and we can get away with a small number of samples, n. If, on the other hand $f(\boldsymbol{x})$ changes rapidly for small perturbations, then we need to draw many samples to cover $\mathcal{X}$ with a fine grid. This motivates us to formalize the degree of "smoothness" in a function, as this will enable us to quantify the difficulty of the problem.

The reader may have previously encountered the concept of *Lipschitz smoothness* or *Lipschitz continuous* functions. A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be C-Lipschitz if the following inequality holds:

$$\|f(\boldsymbol{x}) - f(\boldsymbol{z})\|_2 \leq C\|\boldsymbol{x} - \boldsymbol{z}\|_2 \tag{6}$$

for some value of $C > 0$ and for all choices of $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$. If we think about f as being everywhere differentiable, then intuitively C acts as an upper bound on the first directional derivative (seen by taking $\boldsymbol{x}$ and $\boldsymbol{z}$ to be arbitrarily close together). Thus for smaller C, the function is constrained to change more slowly as we move away from a point $\boldsymbol{x}$.

In the statistical literature, it is common to generalize this notion of smoothness as follows.

Definition 1.1: Hölder smoothness

For any integer $k \geq 0$ and constant $0 \leq \gamma \leq 1$, we will say that a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is $(k + \gamma, C)$-Hölder smooth when:

$$\left| \frac{\partial^k f(\boldsymbol{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_N^{\alpha_N}} - \frac{\partial^k f(\boldsymbol{z})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_N^{\alpha_N}} \right| \leq C \|\boldsymbol{x} - \boldsymbol{z}\|_2^{\gamma} \tag{7}$$

holds for some constant C, for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ and for all $\alpha_1 + \alpha_2 + \cdots + \alpha_d = k$.

The reader should verify that $(1, C)$-smooth functions are Lipschitz with constant C. Further, $(1 + k, C)$-smooth functions are k-times differentiable functions where the kth-order partial derivatives are all Lipschitz with constant C.

The interpretation of γ is trickier. It can help to rearrange eq. (7) as follows:

$$\frac{\left| \frac{\partial^k f(\boldsymbol{x})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_N^{\alpha_N}} - \frac{\partial^k f(\boldsymbol{z})}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_N^{\alpha_N}} \right|}{\|\boldsymbol{x} - \boldsymbol{z}\|_2^{\gamma}} \leq C \tag{8}$$

from which we see that the smoothness condition fails when the left hand side blows up to infinity. If the function is everywhere $(k + 1)$-times differentiable, the left hand side will remain finite when $\gamma = 1$; this being the same intuition we applied when interpreting Lipschitz functions. Now, let us consider the limit where $\gamma \to 0$. In this limit, unless $\|\boldsymbol{x} - \boldsymbol{z}\|_2$ is chosen to be much smaller than γ, the denominator is approximately equal to one (*the reader should check this!*). Intuitively then, the left hand side of eq. (8) will be finite as long as the numerator is finite—that is, the left hand side will be finite if and only if the function is everywhere k-times differentiable.

So, without getting too hung up on the details, we can see that the largest value of $p = k + \gamma$ satisfying eq. (7) is a nice measure of smoothness. We can roughly interpret p as the number of times f is everywhere differentiable, with γ interpolating between integral values of differentiation. For any $q < p$, we it is easy to see that a (p, C_1)-smooth function is also (q, C_2)-smooth for some constant C_2, but it may be the case that $C_1 > C_2$.

1.3 Loss functions, risk of an estimation procedure

In the frequentist setting, we treat the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ as a sequence of random variables. The *estimation procedure* or *estimator* is a deterministic mapping[2] of these random variables to a function,

$$(X_1, Y_1), \dots, (X_n, Y_n) \mapsto \hat{f}. \tag{9}$$

where $\hat{f} : \mathcal{X} \mapsto \mathbb{R}$ is a point estimate of the target function $f : \mathcal{X} \mapsto \mathbb{R}$. It is important to understand that $\hat{f}$ is a random function—it inherits randomness from the training data. Thus,

[2] In principle, we could allow for randomized estimators. However, under most common situations it can be shown that deterministic estimation procedures outperform any non-deterministic procedure. Thus, for simplicity, we will only consider estimators as deterministic mappings from training data to estimated function. See theorem 3.28 (Rao-Blackwell theorem) in Keener [6] for more details.

even if the fitting procedure to obtain $\hat{f}$ is deterministic, the output of the estimator is *not deterministic* unless we condition on a particular realization of the training data.

Once we fix an estimation procedure, we then would like to ask: How well is it expected to perform? It should be noted that even asking this question commits oneself to a frequentist perspective, in which we are interested in understanding the long-run performance of a method over hypothetical future datasets.

A natural way to answer this question is to compute the expected discrepancy between f and $\hat{f}_n$ on *heldout data* or *test data* drawn from the same distribution as the training data. We measure discrepancy using a *loss function*, $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ and the expected loss incurred by an estimator on heldout data is called the *risk*, which we denote $\mathcal{R}(f, \hat{f})$ and discuss in detail below in definition 1.2.

There are many potential loss functions we could choose. We will see that it is convenient to use the *quadratic loss*:

$$\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \tag{10}$$

but other choices are possible including the absolute error $\ell(y, \hat{y}) = |y - \hat{y}|$, which is more robust to outliers. In general, we will assume that the loss is a convex function with respect to $\hat{y}$ for any fixed value of y.

We will soon see that it is useful to interpret the loss function as the negative log-likelihood of observing $Y_i = y$ under the conditional distribution centered at $\hat{y} = \hat{f}(X_i)$. That is, we would choose $\ell(y, \hat{y}) = -\log p_\epsilon(y - \hat{y})$, where $p_\epsilon(\cdot)$ is the probability density (or probability mass) function associated to the distribution of the noise variable. Under this interpretation of the loss function, the quadratic loss in eq. (10) is, up to an additive constant, equal to the negative log likelihood computed from a Gaussian noise model. Choosing the loss function in this way will connect our approach to Bayesian approaches and to maximum likelihood estimation. Although maximum likelihood estimates have optimal frequentist properties in the limit of infinitely many observations ($n \to \infty$), they may not be optimal for finite sample sizes. Thus, it is best to not be dogmatic—it may be reasonable to choose a quadratic loss for convenience even when the true distribution of noise is non-Gaussian. One should never imply that choosing a quadratic loss necessarily "assumes a Gaussian noise model."

The loss function is a deterministic mapping that measures the discrepancy of our estimated function $\hat{f}_n(\boldsymbol{x})$ and the target function $f(\boldsymbol{x})$ for any specified $\boldsymbol{x} \in \mathcal{X}$. The frequentist views the observed data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and future observations (X, Y) as random variables, and seeks to characterize how the loss incurred by an estimation procedure behaves in expectation. The resulting measure of estimator performance the *risk*, which we now define.

Definition 1.2: Risk

Let $\mathcal{D}^n \mapsto \hat{f}$ denote an estimator of the unknown target function f. Furthermore, let $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ be a loss function which measures discrepancy between $\hat{f}$ and f. Then the *risk*, $\mathcal{R}$, is the expected loss of $\hat{f}$ on a heldout datapoint:

$$\mathcal{R}(f, \hat{f}) = \mathbf{E}\left[\ell(f(X), \hat{f}(X))\right] \tag{11}$$

where the expectation is taken jointly over the samples used for training $\mathcal{D}^n \sim P^n_{X,Y}$ and an independent heldout sample $X \sim P_X$ used for evaluation.

Before proceeding we must apologize for an abuse of notation in the above definition. We have written the risk as a function of the target function f and an estimated function $\hat{f}$. In reality, the risk is really a function of two things: the ground truth data generating distribution, $P_{X,Y}$, and the estimation procedure, $\mathcal{D}^n \mapsto \hat{f}$. Thus, it would be more accurate to write $\mathcal{R}(P_{X,Y}, \mathcal{D}^n \mapsto \hat{f})$ in our definition. However, our preferred notation of $\mathcal{R}(f, \hat{f})$ is less cumbersome and also captures the core intuition behind risk—i.e., we want to compare how close $\hat{f}$ is to f in expectation. The reader should understand that the risk is implicitly dependent on $P_{X,Y}$ (since the expectation in eq. (11) is taken with respect to this distribution), and is really a measure of the overall estimation procedure mapping $\mathcal{D}^n \mapsto \hat{f}$ rather than $\hat{f}$ itself, per se.

With apologies out of the way, let us introduce some additional notation and terminology that will come in useful later.

Definition 1.3: Out-of-sample Expected Loss

Due to the Law of Total Expectation (a.k.a. Tower Rule) we can express the risk as:

$$\mathcal{R}(f, \hat{f}) = \mathbf{E}_{\mathcal{D}_n}\left[\mathbf{E}_{X \sim P_X}\left[\ell(f(X), \hat{f}(X)) \mid \mathcal{D}_n\right]\right] \tag{12}$$

We can interpret the inner expectation as the **expected out-of-sample loss**, $\mathcal{L}_{P_X}$, of the estimated function. This can actually be defined as a general measure of discrepancy between any two functions $g : \mathcal{X} \mapsto \mathbb{R}$ and $h : \mathcal{X} \mapsto \mathbb{R}$, as follows:

$$\mathcal{L}_{P_X}(g, h) = \mathbf{E}_{X \sim P_X}[\ell(g(X), h(X))] \tag{13}$$

We can express the risk in terms of this quantity by additionally taking the expectation over the random training data:

$$\mathcal{R}(f, \hat{f}) = \mathbf{E}_{\mathcal{D}_n}\left[\mathcal{L}_{P_X}(f, \hat{f}) \mid \mathcal{D}_n\right] \tag{14}$$

In addition to the risk of an estimator, will also make use of a closely related quantity called the predictive risk.

Definition 1.4: Predictive Risk

The *predictive risk*, $\mathcal{R}^*$, is defined similarly to the risk, but measures the discrepancy between $\hat{f}_n(X)$ and Y instead of the discrepancy to the target function.

$$\mathcal{R}^*(f, \hat{f}) = \mathbf{E}\left[\ell(Y, \hat{f}(X))\right] \tag{15}$$

where the expectation is taken jointly over the samples used for training $\mathcal{D}^n \sim P^n_{X,Y}$ and an independent sample $(X, Y) \sim P_{X,Y}$ used for evaluation.

Note that the risk and predictive risk are deterministic (non-random) quantities since we have taken the expectation over all random variables.

In practice, we can never directly compute the risk or predictive risk. Doing so would require knowing the true data generating distribution, $(X, Y) \sim P_{X,Y}$, to compute the expectations appearing in eqs. (11) and (15). Given any particular estimate of the target function, $\hat{f}$, we can estimate the predictive risk with the *empirical risk*.

Definition 1.5: Empirical Risk

The *empirical risk*, E^*, is defined similarly to the risk, but measures the discrepancy between $\hat{f}(X)$ and Y instead of the discrepancy to the target function:

$$E^*(\hat{f}, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \hat{f}(X_i)) \tag{16}$$

Intuitively, the empirical risk simply approximates the expectation in eq. (15) with an empirical average over the n datapoints. The law of large numbers tells us that as $n \to \infty$, the empirical risk will converge to the true predictive risk. However, we will see that the empirical risk is often a biased due to overfitting, motivating us to develop cross-validation procedures.

The following two exercises show that the predictive risk, $\mathcal{R}^*$, is closely related to the risk, $\mathcal{R}$, and thus it usually suffices to have an empirical estimate of the former.

Exercise 1.1: Predictive risk vs. risk with quadratic loss

Show that if we use a quadratic loss $\ell(y, \hat{y}) = (y - \hat{y})^2$, then the risk is equal to the predictive risk plus a constant. Specifically,

$$\mathcal{R}^*(f, \hat{f}) = \mathcal{R}(f, \hat{f}) + \mathbf{E}[\epsilon^2] \tag{17}$$

where ϵ is a mean-zero random variable describing "noise" as in eq. (4). For simplicity, you can assume that the noise is constant as a function of $x \in \mathcal{X}$ (homoschedastic).

Exercise 1.2: Predictive risk vs. risk with absolute error

Show that if we use an absolute error criterion $\ell(y, \hat{y}) = |y - \hat{y}|$ the risk is upper bounded by the predictive risk plus a constant.

$$\mathcal{R}(f, \hat{f}) \geq \mathcal{R}^*(f, \hat{f}) - \mathbf{E}[|\epsilon|] \tag{18}$$

where ϵ is the "noise" term as in exercise 1.1.

Before closing this section, we will remark on a useful reformulation of the risk called the *bias-variance decomposition*. Unfortunately, this decomposition only applies if we assume a quadratic loss, although analogues can be developed for other loss functions (e.g. [7]). First define $\bar{f}_n : \mathcal{X} \mapsto \mathbb{R}$ to be the expected outcome of the estimation procedure:

$$\bar{f}_n = \mathbf{E}_{\mathcal{D}_n}[\hat{f}_n] \tag{19}$$

where the expectation is taking over the training data $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. It is important to understand that $\bar{f}_n$ is not a random object, in contrast to $\hat{f}_n$ which depends on the random training data (as we mentioned before).

Result 1.1: Bias-Variance Decomposition

When using a quadratic loss function, the risk of a nonparametric regression estimator can be written:

$$\mathcal{R}(f, \hat{f}) = \underbrace{\mathbf{E}_X[(f(X) - \bar{f}_n(X))^2]}_{\text{"bias"}} + \underbrace{\mathbf{E}_{X,\mathcal{D}_n}[(\hat{f}_n(X) - \bar{f}_n(X))^2]}_{\text{"variance"}} \tag{20}$$

where the first expectation is taken with respect to a heldout data point $X \sim P_X$ and the second expectation is taken jointly over $X \sim P_X$ and the training data.

Proof. Due to linearity of expectation:

$$\begin{aligned} R(f) &= \mathbf{E}\left[(f(X) - \hat{f}_n(X))^2\right] && (21)\\ &= \mathbf{E}\left[f^2(X) + \hat{f}_n^2(X) - 2f(X)\hat{f}_n(X)\right] && (22)\\ &= \mathbf{E}\left[f^2(X)\right] + \mathbf{E}\left[\hat{f}_n^2(X)\right] - 2\mathbf{E}[f(X)\hat{f}_n(X)]. && (23) \end{aligned}$$

First, we have:

$$\begin{aligned} \mathbf{E}\left[\hat{f}_n^2(X)\right] &= \mathbf{E}\left[\hat{f}_n^2(X) + \bar{f}_n^2(X) - \bar{f}_n^2(X) - 2\hat{f}_n(X)\bar{f}_n(X) + 2\hat{f}_n(X)\bar{f}_n(X)\right] \\ &= \mathbf{E}\left[\left(\hat{f}_n(X) - \bar{f}_n(X)\right)^2 - \bar{f}_n^2(X) + 2\hat{f}_n(X)\bar{f}_n(X)\right] \\ &= \mathbf{E}\left[\left(\hat{f}_n(X) - \bar{f}_n(X)\right)^2\right] - \mathbf{E}\left[\bar{f}_n^2(X)\right] + 2\mathbf{E}\left[\hat{f}_n(X)\bar{f}_n(X)\right] \\ &= \mathbf{E}\left[\left(\hat{f}_n(X) - \bar{f}_n(X)\right)^2\right] - \mathbf{E}\left[\bar{f}_n^2(X)\right] + 2\mathbf{E}\left[\bar{f}_n^2(X)\right] \\ &= \mathbf{E}\left[\left(\hat{f}_n(X) - \bar{f}_n(X)\right)^2\right] + \mathbf{E}\left[\bar{f}_n^2(X)\right] \end{aligned}$$

Second, we have:

$$\mathbf{E}\left[f(X)\hat{f}_n(X)\right] = \mathbf{E}_X\left[f(X)\mathbf{E}_{\mathcal{D}}\left[\hat{f}_n(X)\right]\right] = \mathbf{E}_X\left[f(X)\bar{f}_n(X)\right]$$

Plugging these in for the second two terms of eq. (23), we arrive at the desired result:

$$\begin{aligned} R(f) &= \mathbf{E}\left[\left(\hat{f}_n(X) - \bar{f}_n(X)\right)^2\right] + \mathbf{E}\left[\bar{f}_n^2(X)\right] + \mathbf{E}\left[f^2(X)\right] - 2\mathbf{E}\left[f(X)\bar{f}_n(X)\right]\Big] \\ &= \mathbf{E}\left[\left(\hat{f}_n(X) - \bar{f}_n(X)\right)^2\right] + \mathbf{E}\left[\left(f(X) - \bar{f}_n(X)\right)^2\right] \end{aligned}$$

□

1.4 Comparing estimators, minimax risk

We have introduced the risk, $\mathcal{R}$, of an estimator as a way to quantify performance. In this section we will show how this quantity can be used to compare different estimators. As before, the estimators we are interested in are mappings from training data to functions $\mathcal{X} \mapsto \mathbb{R}$:

$$(X_1, Y_1), \ldots, (X_n, Y_n) \mapsto \hat{f}_1 \quad \text{and} \quad (X_1, Y_1), \ldots, (X_n, Y_n) \mapsto \hat{f}_2 \tag{24}$$

where $\hat{f}_1$ and $\hat{f}_2$ are estimates of the target function f. To condense notation, we will again use $\mathcal{D}^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ to denote the training data. Thus, we can equivalently write

$$\mathcal{D}^n \mapsto \hat{f}_1 \quad \text{and} \quad \mathcal{D}^n \mapsto \hat{f}_2 \tag{25}$$

As before, $\hat{f}_1$ and $\hat{f}_2$ are random objects that inherit randomness from the training data. The risks associated with these estimators, $\mathcal{R}(f, \hat{f}_1)$ and $\mathcal{R}(f, \hat{f}_2)$, are non-random quantities, since we have taken the expectation over $\mathcal{D}^n$.

We are now in a position to quantify which estimator is "better" in the sense of having lower risk of a set of possible target functions $\mathcal{F}$. You can think of $\mathcal{F}$ as being the set of all (p, C)-Hölder-smooth functions for some choice of p and C. For example, suppose that we could show that:

$$\mathcal{R}(f, \hat{f}_1) \leq \mathcal{R}(f, \hat{f}_2) \quad \text{for all } f \in \mathcal{F} \tag{26}$$

If we could prove this, then we could essentially discard $\mathcal{D}^n \mapsto \hat{f}_2$ as an estimation procedure since there is always a better alternative. In this situation, statisticians would call $\mathcal{D}^n \mapsto \hat{f}_2$ an *inadmissible* estimator.

While it is possible to prove that estimators are inadmissible, it is generally quite challenging since one needs to show that the inequality in eq. (26) holds over all possible target functions. Even a obviously bad estimator can have small risk for a carefully chosen distribution. For example, consider an estimator that ignores the training data entirely and estimates $\hat{f}(\boldsymbol{x}) = 0$. This estimator has zero risk when the target function is $f(\boldsymbol{x}) = 0$, but clearly has very large risk in almost any other circumstance (e.g. $f(\boldsymbol{x}) = c$ for some constant $|c| \gg 0$). In other words, when comparing two estimators, associated to $\hat{f}_1$ and $\hat{f}_2$, it is often the case that one can construct target functions $f_1 \in \mathcal{F}$ and $f_2 \in \mathcal{F}$ such that:

$$\mathcal{R}(f_1, \hat{f}_1) < \mathcal{R}(f_1, \hat{f}_2) \quad \text{while} \quad \mathcal{R}(f_2, \hat{f}_1) > \mathcal{R}(f_2, \hat{f}_2) \tag{27}$$

thus it is not possible to rank order the estimators simultaneously over all conceivable target functions.

To circumvent these difficulties, we can instead compare estimators by their *worst-case* performance. That is, if we wanted to show that $\mathcal{D}^n \mapsto \hat{f}_1$ is "better" than $\mathcal{D}^n \mapsto \hat{f}_2$ in terms of worst-case performance, we could try to prove that:

$$\sup_{f \in \mathcal{F}} \mathcal{R}(f, \hat{f}_1) \leq \sup_{f \in \mathcal{F}} \mathcal{R}(f, \hat{f}_2) \tag{28}$$

where $\sup_f$ denotes the supremum over f.[3][4]

By comparing estimators on the basis of their worst-case risk, then we avoid the situation presented in eq. (27). At least in principle, we can truly rank order the estimators from "best" to "worst." This leads us to the notion of the *minimax risk* which corresponds to the best possible performance we can hope to achieve.

Definition 1.6: Minimax risk

The *minimax* risk, $\mathcal{M}$, is the smallest worst-case risk achievable by the set of all estimation procedures. Specifically,

$$\mathcal{M} = \inf_{\hat{f}} \sup_{f} \mathcal{R}(f, \hat{f}) \tag{29}$$

where the infimum is taken over all estimation procedures mapping $\mathcal{D}^n \mapsto \hat{f}$ and the supremum is taken over a set of possible data generating distributions $P_{X,Y}$ with target functions $f \in \mathcal{F}$.

As in definition 1.2, we have introduced some abusive notation and it would be more precise for us to define the minimax risk as:

$$\mathcal{M} = \inf_{\mathcal{D}^n \mapsto \hat{f}} \sup_{P_{X,Y}} \mathcal{R}(P_{X,Y}, \mathcal{D}^n \mapsto \hat{f}) \tag{30}$$

But we prefer to the more compact expression in eq. (29).

At a first glance, the minimax risk seems like a very formidable calculation. An explicit calculation would require a minimization over all possible estimators, cascaded with a maximization over all possible data generating distributions and an expectation over instantiations of these data distributions (which is needed to compute each estimator's risk). Remarkably, statisticians have developed a number of techniques to lower bound and upper bound the minimax risk, giving us precise information into the worst-case difficulty of nonparametric regression. The following result summarizes the punchline.

[3] More precisely, in the setting of random design regression the supremum should be over the full data generating distribution $P_{X,Y}$, not just a supremum over the target function.

[4] If you are unfamiliar with the concept of a supremum, it is okay to pretend it is the same as a maximum. Similarly if you are unfamiliar with the concept of an infimum, it is okay to pretend it is a minimum. That is, you can mentally replace $\sup_f \leftrightarrow \max_f$ and $\inf_f \leftrightarrow \min_f$ everywhere in these notes.

Result 1.2: Minimax rate for nonparametric regression with quadratic loss

Let $\mathcal{F}$ be the set of (p, L)-Hölder-smooth functions mapping $\mathbb{R}^d \mapsto \mathbb{R}$. Let $\mathcal{M}(n, L)$ be the minimax risk associated with estimating an unknown target function $f \in \mathcal{F}$ with quadratic loss and n training samples. Then, there exist constants $C_2 \geq C_1 \geq 0$ such that:

$$C_1 \cdot L^{\frac{2p+d}{2d}} \cdot n^{\frac{-2p}{2p+d}} \leq \mathcal{M}(n, L) \leq C_2 \cdot L^{\frac{2p+d}{2d}} \cdot n^{\frac{-2p}{2p+d}} \tag{31}$$

holds. The value of the constants C_1 and C_2 may depend on p and d, but they do not depend on L or n.

Roughly speaking, this result tells us that the optimal nonparametric regression method (judged in terms of worst-case risk) incurs an expected loss proportional to $L^{(2p+d)/2d} n^{-2p/(2p+d)}$ when given n training samples on a d-dimensional regression problem where the target function is (p, L)-Hölder-smooth. For example, if if the target function is L-Lipschitz and univariate (i.e. $p = 1$ and $d = 1$), then the minimax risk is proportional to $L^{3/2} n^{-2/3}$.

Qualitatively, we see that the minimax risk decreases when L decreases—i.e. performance improves as the target function gets smoother. We also see that the minimax risk decreases when n increases—i.e. performance improves as we observe more data. Of course, these trends agree with our intuition, but result 1.2 strengthens this into a precise quantitative guarantee.

By definition, we can never develop a practical method that *beats* the minimax risk. The best we can do is hope to match it—and, again remarkably, it can be shown that we can. We summarize this second punchline as follows.

Result 1.3: Achieving the minimax rate

Under the same assumptions as result 1.2, we can construct estimators $\mathcal{D}^n \mapsto \hat{f}$ that are both useful in practice and which satisfy:

$$\mathcal{M}(n, L) \leq \sup_f \mathcal{R}(f, \hat{f}) \leq C_2 \cdot L^{\frac{2p+d}{2d}} \cdot n^{\frac{-2p}{2p+d}} \tag{32}$$

Estimators satisfying eq. (32) are called "minimax rate optimal" or are said to "achieve the optimal minimax rate" of convergence. Together Result 1.2 and 1.3 imply that these estimators are unimprovable, except potentially up to a multiplicative factor. That is, some estimators may suffer from a larger value of C_2, relative to others.

The proofs for Result 1.2 and 1.3 are rather involved and will be developed in subsequent chapters. We refer the advanced reader seeking formal proofs to theorems 3.2, 3.3, 19.4, and corollary 19.1 appearing in Györfi et al. [5]. The proofs found therein can be directly applied to the results we have cited above. Lower bounds on the minimax risk for Hölder-smooth target functions were first derived by Stone [10].

1.5 Bayesian perspectives

We have thus far spent a lot of ink covering the frequentist framing of nonparametric regression. We now give a brief description of the Bayesian perspective, which we will develop in much greater detail in subsequent chapters. The main punchline of this section is that Bayesian procedures for nonparametric regression have nice frequentist properties. Under certain assumptions, they converge to the true target function f at the minimax optimal rate of $n^{-2p/(2p+d)}$ (see Result 1.2 and 1.3).

A Bayesian approach begins by placing a prior distribution on the target function. A full generative model of the data under a random design is:

$$\begin{aligned} f &\sim P_f \\ X_i &\sim P_X && \text{independently for } i = 1, \dots, n \\ \epsilon_i \mid X_i &\sim P_\epsilon(X_i) && \text{independently for } i = 1, \dots, n \\ Y_i &= f(X_i) + \epsilon_i && \text{for } i = 1, \dots, n \end{aligned} \tag{33}$$

where P_f is the prior over the target function, P_X is a distribution over the independent variables (random design setting), and $P_\epsilon(X)$ is a noise distribution with zero mean which may vary as a function of X. As before, we assume that the X_i and Y_i random variables are observed and we use $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ as a shorthand. In contrast, f and the ϵ_i variables are unobserved (or "latent") random objects.

Our primary goal will be to infer the posterior distribution of the target function conditioned on the observed data. The usual way one goes about this is to use Bayes rule to relate the posterior density to the likelihood and prior density as follows:

$$p(f \mid \mathcal{D}_x) \propto p(\mathcal{D}_x \mid f) p(f) \tag{34}$$

Typically, we implement a function that evaluates the right hand side—which is an *unnormalized* density—and pass this off to a Markov Chain Monte Carlo (MCMC) sampler or variational inference routine (for background, see e.g. [8]). **Unfortunately, eq. (34) cannot be made rigorous under most circumstances**. The problem is that, while we can rigorously define a probability distribution over an infinite-dimensional function space, such as P_f, we usually cannot define a probability density, such as $p(f)$. For further technical details, see Eldredge [4].

We will return to these subtleties later, but for now I want to sidestep them as much as possible. It turns out that we can rigorously define a posterior distribution $P_{f \mid \mathcal{D}_n}$, even though we cannot explicitly write down a density. Intuitively, if we assume there is a "true" function f, we hope that $P_{f \mid \mathcal{D}_n}$ assigns large probability to the region of function space close to f. One way to quantify this intuition would be to extract a point estimate from the posterior, such as the mean of $P_{f \mid \mathcal{D}_n}$. Then, we could measure the risk $\mathcal{R}(f, \hat{f})$, as before under a frequentist framework. While this is possible, it feels against the Bayesian philosophy—why bother with Bayesian inference in the first place, if we just revert to a frequentist mindset at the final step?

An alternative is to use the *posterior risk*, which we define below. Whereas the frequentist risk quantified the performance of a point estimator $\mathcal{D}_n \mapsto \hat{f}$; the posterior risk quantifies the performance of a procedure that outputs a distribution over functions $\mathcal{D}_n \mapsto P_{\hat{f} \mid \mathcal{D}_n}$.

Definition 1.7: Posterior risk

Recall from definition 1.3 the expected out-of-sample loss

$$\mathcal{L}_{P_X}(g, h) = \mathbf{E}_{X \sim P_X}[\ell(g(X), h(X))]$$

which is a measure of distance between any two functions g and h mapping $\mathcal{X} \mapsto \mathbb{R}$. The **posterior risk** is the expected distance under the posterior distribution $P_{\hat{f} \mid \mathcal{D}_n}$ relative to the "true" target function f. Formally,

$$\mathcal{R}\left(f, P_{\hat{f} \mid \mathcal{D}_n}\right) = \mathbf{E}_{\mathcal{D}_n}\left[\int \mathcal{L}_{P_X}(f, \hat{f})\, dP_{\hat{f} \mid \mathcal{D}_n}\right] \tag{35}$$

Note that we use the same symbol $\mathcal{R}$ to denote the risk of a point estimate $\mathcal{R}(f, \hat{f})$ and the posterior risk $\mathcal{R}(f, P_{\hat{f} \mid \mathcal{D}_n})$. It will always be clear from context which of the two we are referencing. More importantly, one can interpret the frequentist risk $\mathcal{R}(f, \hat{f})$ as the posterior risk under a distribution with a point Dirac mass placed at $\hat{f}$. Thus, the frequentist risk can loosely be viewed as a special case of posterior risk, the only difference being that the former quantifies the performance of an estimator that outputs a singular posterior distribution at $\hat{f}$.

Now that we have defined posterior risk and shown that it is closely related to the frequentist risk, we are in a position to state the main result of this section (without proof, for now). It tells us that if we choose a good prior distribution, Bayesian inference will often produce a good estimate at the same rate we can expect of optimal frequentist estimators.

Result 1.4: Convergence of Bayesian inference at minimax optimal rates

Let f denote a ground truth target function that is (p, L)-Hölder-smooth. One can show under certain conditions[a] that there are constants C_1 and C_2 such that, for all n, we have:

$$C_1 n^{\frac{-2p}{2p+d}} \leq \mathcal{R}(f, P_{\hat{f} \mid \mathcal{D}_n}) \leq C_2 n^{\frac{-2p}{2p+d}} \tag{36}$$

where $\mathcal{R}$ denotes the posterior risk under a quadratic loss function.

[a] Among other things, these conditions include that we need to choose a "good" prior distribution P_f.

This result is stated loosely to provide intution; we will sharpen it to a more precise statement later. For immediate details, the reader can consult Castillo [2] for the first inequality (lower bound), and consult Van Der Vaart and Van Zanten [12] for the second inequality (upper bound).

1.6 Overview of these notes

We have now posed the problem of nonparametric regression (section 1.1), established necessary smoothness assumptions on the target function f (section 1.2), and discussed how to quantify the performance of regression methods from a frequentist perspective (section 1.3). We have also stated, without proof, that the worst case approximation error of *any estimation procedure* scales unfavorably with the number of independent variables. For example, the minimax risk is proportional to $n^{-2/(2+d)}$ for Lipschitz smooth functions (section 1.4). Finally, as discussed in section 1.5, these concepts can be connected to Bayesian approaches to nonparametric regression.

In particular, we have seen that optimal Bayesian and frequentist approaches converge, in some sense, to the solution at the same rate as a function of n (the amount of training data).

We have yet to discuss any practical procedures to solve the nonparametric regression problem on real data. There are in fact a variety of methods that work well both in practice and in theory. Having a flexible toolkit of methods is useful, but presents an organizational challenge. When reading through the literature, one finds estimators based on: partitioning/local averaging, k-nearest neighbors, Nadaraya-Watson kernels, local polynomial regression,[5] kernel ridge regression, regression splines, smoothing splines, et cetera. Many of these methods are closely related to each other, or even identical under special circumstances (see e.g. Silverman [9]).

Our strategy to navigate this complex ecosystem will be as follows. We first begin with a frequentist approach to nonparametric regression. In this setting, many (though not all) estimators can be expressed as optimization problem over candidate functions $\hat{f}$ from a function space $\hat{F}_n$. Specifically, we aim to minimize the empirical risk $E^*(\hat{f}) = \frac{1}{n}\sum_i \ell(Y_i, \hat{f}(X))$, as previously given in definition 1.5 plus a penalty function $\mathcal{H}_n : \hat{\mathcal{F}}_n \mapsto \mathbb{R}$ which intuitively penalizes more complex (often "wigglier") functions. Altogether, we aim to:

$$\begin{aligned} \underset{\hat{f}}{\text{minimize}} \quad & E^*(\hat{f}, \mathcal{D}_n) + \mathcal{H}_n(\hat{f}) \\ \text{subject to} \quad & \hat{f} \in \hat{\mathcal{F}}_n \end{aligned}$$

to estimate the target function. The subscript of n appearing in $\hat{F}_n$ and $\mathcal{H}_n$ denotes that we are allowed to choose the set of candidate models and the penalty function adaptively depending on n. Intutively, for larger n we may need less regularization to prevent overfitting, so we could choose a more lenient penalty function and/or a larger family of candidate functions.

After characterizing these frequentist methods, we turn to Bayesian approaches. Our move will be to interpret the optimization problem above as performing *maximum a posteriori* (MAP) inference. Roughly speaking, we can interpret $E^*(\hat{f}) = \frac{1}{n}\sum_i \ell(Y_i, \hat{f}(X))$ as a negative log-likelihood term and $\mathcal{H}(\hat{f})$ as the contribution of the prior distribution over $\hat{f}$. The set of candidate functions $\hat{\mathcal{F}}_n$ can be interpreted as the support of the prior. Thus, it will not be too much work for us to extend our understanding of frequentist nonparametric regression to the Bayesian setting.

Methods and models that do not fit into the above framework will be dealt with in the final chapters (if I ever get around to writing them).

[5] Which is sometimes called locally estimated scatterplot smoothing (LOESS) or locally weighted scatterplot smoothing (LOWESS) in the case of $d = 1$ independent variables.

2 Lower bounds on minimax risk

Our goal in this chapter is to prove the first inequality in result 1.2, which is a lower bound on minimax risk for nonparametric regression under a quadratic loss. Let us recall the generative model:

$$\begin{aligned} X_i &\sim P_X && \text{independently for } i \in \{1, \dots, n\} \\ \epsilon_i \mid X_i &\sim P_\epsilon(X_i) && \text{independently for } i \in \{1, \dots, n\} \\ Y_i &= f(X_i) + \epsilon_i && \text{for } i \in \{1, \dots, n\} \end{aligned}$$

where $f : \mathcal{X} \mapsto \mathbb{R}$ is a (p, L)-Hölder-smooth function that we'd like to estimate. The dimensionality of the problem is $d = \dim(\mathcal{X})$. In this chapter we will use σ^2 to denote the maximum noise variance over all independent variables, that is $\sigma^2 = \sup_{x \in \mathcal{X}} \mathbf{Var}[\epsilon \mid X = x]$.

The concrete result we seek to prove is that:

$$\inf_{\mathcal{D}_n \mapsto \hat{f}} \sup_{P_{X,Y}} \mathbf{E} \|f(X) - \hat{f}(X)\|^2 \geq C L^{\frac{2d}{2p+d}} n^{\frac{-2p}{2p+d}} \tag{37}$$

for some constant C. We will see that the constant C depends on σ^2 and p, but does not depend on d, L, or n. In the equation above, recall that the expectation is taken with respect to a dataset $\mathcal{D}_n$, the supremum is taken over all data generating distributions satisfying the Hölder-smoothness constraint on f, and the infimum is taken over all estimation procedures.

In section 2.1 we prove the lower bound in a special case where $d = 1$ and p is an integer. That is, we consider the case of estimating a univariate function that is $p - 1$ times continuously differentiable with Lipschitz parameter L. Although this is a simple case and we often care deeply about the high-dimensional case (where $d > 1$), the proof technique is instructive and can be extended to handle the fully general case.

In the future, I hope to fill out the proof for the case of general d and p in section 2.2.

2.1 Proof when $d = 1$ and p is an integer

Consider the case where $d = 1$ (univariate regression), and p is an integer. Our proof technique follows Theorem 3.2 in Györfi et al. [5], which proves the general case.

Since p is an integer, (p, L)-Hölder-smoothness condition can be expressed:

$$|f^{(p-1)}(x) - f^{(p-1)}(z)| \leq L|x - z| \tag{38}$$

for all x, z in the domain of f.

Step 1: *Choose a distribution over independent variables and noise distribution*

The overarching idea behind this proof will be to explicitly construct a data generating distribution $P_{X,Y}$ that is difficult to estimate. We start by defining a distribution over X and the noise.

Suppose that the input variable is uniformly distributed on the unit interval and the noise is standard Gaussian. That is, we choose $X_i \sim \text{Unif}([0,1])$ and $\epsilon_i \sim \mathcal{N}(0,1)$ independently for $i, \dots, n$. All that remains is to specify a space of target functions $\mathcal{F}$.

Note that, as long as $\mathcal{F}$ is a subset of all (p, L)-Hölder-smooth functions, we can lower bound the minimax risk:

$$\inf_{\mathcal{D}_n \mapsto \hat{f}} \sup_{P_{X,Y}} \mathbf{E}\|f(X) - \hat{f}(X)\|^2 \geq \inf_{\mathcal{D}_n \mapsto \hat{f}} \sup_{f \in \mathcal{F}} \mathbf{E}\|f(X) - \hat{f}(X)\| \tag{39}$$

where on the right hand side we are assuming that X_i and ϵ_i are distributed as we specified above. The inequality follows from the fact that we are taking the supremum over a restricted set of distributions $P_{X,Y}$, so the supremum can only be "worse" (i.e. achieve a smaller value).

In short, we can now write down an explicit and simple expression for the *risk*.

$$\mathcal{R}(f, \hat{f}) := \mathbf{E}(f(X) - \hat{f}(X))^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx \tag{40}$$

Our goal is to lower bound $\mathcal{M} = \inf_{\hat{f}} \sup_f \mathcal{R}(f, \hat{f})$.

Step 2: *Specify a space of target functions $\mathcal{F}$*

Choose a function $\tilde{\phi} : \mathbb{R} \mapsto \mathbb{R}$ that is (p, L)-Hölder-smooth and which satisfies $\tilde{\phi}(x) = 0$ for all $x \leq 0$ and all $x \geq 1$. Note that this implies that all $p - 1$ derivatives are zero at the borders:

$$\tilde{\phi}^{(1)}(0) = \cdots = \tilde{\phi}^{(p-1)}(0) = 0 \qquad \text{and} \qquad \tilde{\phi}^{(1)}(1) = \cdots = \tilde{\phi}^{(p-1)}(1) = 0 \tag{41}$$

otherwise the derivatives would be discontinuous and violate the smoothness condition.

Next, partition the unit interval into $J \geq 1$ equi-spaced bins. We will chose the number of bins, J, later. Within each bin, we place a translated, compressed, and re-scaled copy of $\tilde{\phi}$. Specifically,

$$\phi_1(x) = \frac{\tilde{\phi}(Jx)}{J^p} \quad , \quad \phi_2(x) = \frac{\tilde{\phi}(Jx + 1/J)}{J^p} \quad , \quad \dots \quad , \quad \phi_J(x) = \frac{\tilde{\phi}(Jx + (J-1)/J)}{J^p} \tag{42}$$

Our target function $f : [0,1] \mapsto \mathbb{R}$ will be a linear combination of these basis functions with coefficients equal to $+1$ or -1. That is our class of functions will be

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{J} s_j \phi_j(x) \;\middle|\; s_j = \pm 1, \text{for } j = 1, \dots, J \right\} \tag{43}$$

Note that this is a discrete set containing 2^J possible target functions.

It is easy to prove that every $f \in \mathcal{F}$ satisfies the Hölder smoothness condition.

First, we show that every basis function is (p, L)-Hölder smooth due the re-scaling by J^{-p}. Consider the first basis function ϕ_1. For any $x \in \mathbb{R}$ and $z \in \mathbb{R}$, we have:

$$\left| \phi_1^{(p-1)}(x) - \phi_1^{(p-1)}(z) \right| = \left| \frac{d^{p-1}}{dx^{p-1}} \frac{\tilde{\phi}(Jx)}{J^p} - \frac{d^{p-1}}{dx^{p-1}} \frac{\tilde{\phi}(Jz)}{J^p} \right| \tag{44}$$

$$= \left| J^{p-1} \frac{\tilde{\phi}^{(p-1)}(Jx)}{J^p} - J^{p-1} \frac{\tilde{\phi}^{(p-1)}(Jz)}{J^p} \right| \tag{45}$$

$$= J^{-1} \left| \tilde{\phi}^{(p-1)}(Jx) - \tilde{\phi}^{(p-1)}(Jz) \right| \tag{46}$$

$$\leq LJ^{-1} |Jx - Jz| \tag{47}$$

$$= L|x - z| \tag{48}$$

where the inequality follows from the requirement that $\tilde{\phi}$ is (p, L)-H"older smooth. The remaining basis functions are translated copies of ϕ_1, so clearly all basis functions are (p, L)-Hölder smooth.

Now we turn to prove that $f \in \mathcal{F}$ is also (p, L)-Hölder smooth:

$$f(x) = \phi_j(x) \qquad \text{and} \qquad f(z) = \phi_\ell(z) \tag{49}$$

where j and ℓ denote the bin indices for x and z, respectively. First, consider the case where $j = \ell$, meaning that x and z are in the same bin. Since every ϕ_j is (p, L)-Hölder-smooth we trivially have:

$$|f^{(p-1)}(x) - f^{(p-1)}(z)| = |\phi_j^{(p-1)}(x) - \phi_j^{(p-1)}(z)| \leq L|x - z| \tag{50}$$

Now consider the case where $j \neq \ell$. Without loss of generality we can assume that $z > x$ so $j < \ell$. Now we define $\overline{x}$ as the right edge of bin j, and define $\overline{z}$ as the left edge of bin ℓ. We then have the following sequence of inequalities:

$$\begin{aligned}
|f^{(p-1)}(x) - f^{(p-1)}(z)| &\leq |f^{(p-1)}(x)| + |f^{(p-1)}(z)| && \text{(triangle inequality)} \\
&= |\phi_j^{(p-1)}(x)| + |\phi_\ell^{(p-1)}(z)| && \text{(disjoint basis funcs)} \\
&= |\phi_j^{(p-1)}(x) - \phi_j^{(p-1)}(\overline{x})| + |\phi_\ell^{(p-1)}(z) - \phi_\ell^{(p-1)}(\overline{z})| && \text{(from eq. 41)} \\
&\leq L\left(|x - \overline{x}| + |z - \overline{z}|\right) && \text{(basis funcs are smooth)} \\
&\leq L|x - z|
\end{aligned}$$

which proves the claim.

Step 3: *Project our estimator onto the linear subspace that contains $\mathcal{F}$*

Let $\mathcal{D}_n \mapsto \hat{f}_0$ be any arbitrary estimator of f. We can always improve this estimator by taking advantage of the special structure of $\mathcal{F}$. Intuitively, we can bring $\hat{f}_0$ closer to the true target function by projecting it onto the set of functions:

$$\text{span}(\mathcal{F}) = \left\{ f(x) = \sum_{j=1}^{J} w_j \phi_j(x) \;\middle|\; w_1, \ldots, w_J \in \mathbb{R}^J \right\} \tag{51}$$

which is a J-dimensional linear subspace of functions that contains $\mathcal{F}$ as a subset.

To project $\hat{f}_0$ onto this set, we must solve:

$$\hat{f}_1 = \underset{g \in \text{span}(\mathcal{F})}{\text{argmin}} \; \mathbf{E}_{X \sim P_X} \left(g(X) - \hat{f}_0(X)\right)^2 \tag{52}$$

This produces a new estimator $\hat{f}_1$, which can only have lower risk.

Since P_X is the uniform distribution over $[0, 1]$ the minimized expression can be reformulated:

$$\mathbf{E}_{X\sim P_X}(g(X) - \hat{f}_0(X))^2 = \int_0^1 (g(x) - \hat{f}_0(x))^2 dx \tag{53}$$

$$= \int_0^1 (g(x) - \hat{f}_0(x))^2 dx \tag{54}$$

$$= \int_0^1 g^2(x)dx + \int_0^1 \hat{f}_0^2(x)dx - 2\int_0^1 g(x)\hat{f}_0(x)dx \tag{55}$$

The second term is constant with respect to g and can be dropped. To evaluate the remaining two terms we substitute $g(x) = \sum_j w_j\phi_j(x)$. The first term becomes:

$$\int_0^1 g^2(x)dx = \int_0^1 \left(\sum_{j=1}^J w_j\phi_j(x)\right)\left(\sum_{\ell=1}^J w_\ell\phi_\ell(x)\right)dx \tag{56}$$

$$= \sum_{j=1}^J \sum_{\ell=1}^J w_j w_\ell \int_0^1 \phi_\ell(x)\phi_j(x)dx \tag{57}$$

$$= \sum_{j=1}^J w_j^2 \int_0^1 \phi_j^2(x)dx \qquad (\text{since } \textstyle\int_0^1 \phi_j(x)\phi_\ell(x)dx = 0 \text{ if } j \neq \ell.) \tag{58}$$

Thus, the minimization problem eq. (52) is seen to be equivalent to:

$$\underset{w_1,\ldots w_J}{\operatorname{argmin}} \sum_{j=1}^J w_j^2 \int_0^1 \phi_j^2(x)dx - 2w_j \int_0^1 \phi_j(x)\hat{f}_0(x)dx \tag{59}$$

this problem is convex over $w_1, \ldots, w_J$ and we can solve it by differentiating with respect to these variables and setting the resulting system of equations equal to zero. This yields the solution:

$$\hat{f}_1(x) = \sum_{j=1}^J w_j\phi_j(x)\,, \qquad \text{where} \qquad w_j = \frac{\int_0^1 \phi_j(x)\hat{f}_0(x)dx}{\int_0^1 \phi_j^2(x)dx} \tag{60}$$

Let us pause to provide some intuition and interim summary. We have shown that any estimator can be transformed into a linear combination over the basis functions $\phi_1, \ldots, \phi_J$ and that such a transformation only has the potential improve the performance (decrease the risk) of the estimator.

This greatly simplifies our analysis since we've reduced the problem down to estimating J coefficients: $w_1, \ldots, w_J$. The value of w_j tells us how to scale basis function ϕ_j, which is nonzero only in bin j. Since all derivatives of ϕ_j decay to zero at the edge of the bins (see eq. (41)), datapoints where X_i falls outside of bin j provide us no information about the optimal value of w_j.

Step 4: *Choose a random target function from $\mathcal{F}$ and express the lower bound in terms of an expectation*

The supremum over $f \in \mathcal{F}$ of the risk is lower bounded by the expected risk we incur by sampling uniformly from $\mathcal{F}$ to produce our target function. That is,

$$\sup_{f \in \mathcal{F}} \left[\mathbf{E}_{\mathcal{D}_n} \int_0^1 (f(x) - \hat{f}(x))^2 dx \right] \geq \mathbf{E}_{f, \mathcal{D}_n} \left[\int_0^1 (f(x) - \hat{f}(x))^2 dx \right] \tag{61}$$

In the final expression, the outer iteration is taken jointly over target functions and training datasets. Concretely, we sample $f \sim \text{Unif}(\mathcal{F})$ and then (conditioned on this target function) sample a dataset from the induced distribution $\mathcal{D}_n \sim P_{X,Y}^n$ and use this dataset to determine $\hat{f}$. Thus, both f and $\hat{f}$ appearing inside the final expectation are random functions.

Following the construction in step 3, let $\hat{f}(x) = \sum_j w_j \phi_j(x)$ be the estimated function. Then,

$$\int_0^1 (f(x) - \hat{f}(x))^2 dx = \int_0^1 \sum_j (s_j - w_j)^2 \phi_j^2(x) dx = \sum_j (s_j - w_j)^2 \int_0^1 \phi_j^2(x) dx \tag{62}$$

The final integral can be written in terms of $\tilde{\phi}$:

$$\int_0^1 \phi_1^2(x) dx = J^{-2p} \int_0^1 \tilde{\phi}^2(Jx) dx = J^{-2p-1} \int_0^1 \tilde{\phi}^2(u) du \tag{63}$$

where we substituted $u = Jx$ (so $du/dx = J$ and $dx = du/J$). So, returning to eq. (61), we have:

$$\mathbf{E}_{f, \mathcal{D}_n} \left[\int_0^1 (f(x) - \hat{f}(x))^2 dx \right] = \left(J^{-2p-1} \int_0^1 \tilde{\phi}^2(x) dx \right) \mathbf{E}_{f, \mathcal{D}_n} \left[\sum_j (s_j - w_j)^2 \right] \tag{64}$$

Recall that $s_j \in \{-1, +1\}$. As a consequence, for every bin j we have:

$$(s_j - w_j)^2 \geq \boldsymbol{I}[\, s_j \neq \text{sign}(w_j)\,] \tag{65}$$

where $\boldsymbol{I}[q]$ evaluates to one when q is a true statement. Additionally, every bin is identical and statistically independent so the probability that $s_j \neq \text{sign}(w_j)$ is the same as the probability that $s_1 \neq \text{sign}(w_1)$. Thus,

$$\begin{aligned} \mathbf{E}_{f, \mathcal{D}_n} \left[\sum_j (s_j - w_j)^2 \right] &\geq \sum_j \mathbf{E}_{f, \mathcal{D}_n} \Big[\boldsymbol{I}\left[s_j \neq \text{sign}(w_j) \right] \Big] \\ &= J \cdot \mathbf{E}_{f, \mathcal{D}_n} \Big[\boldsymbol{I}\left[s_1 \neq \text{sign}(w_1) \right] \Big] \\ &= J \cdot \Pr\left[s_1 \neq \text{sign}(w_1) \right] \end{aligned}$$

Where we have substituted in the final line that the expectation of an indicator function of a set is simply the probability of that set occurring. Here we understand this probability as occuring over randomly sampled target functions f and datasets $\mathcal{D}_n$. Putting this all together, our lower bound on the maximal risk becomes:

$$\mathbf{E}_{f, \mathcal{D}_n} \left[\int_0^1 (f(x) - \hat{f}(x))^2 dx \right] \geq \left(J^{-2p} \int_0^1 \tilde{\phi}^2(x) dx \right) \Pr\left[s_1 \neq \text{sign}(w_1) \right] \tag{66}$$

where the first term is constant (does not depend on the estimation procedure).

Step 5:

Let us summarize our position. We have a lower bound that holds for any estimation procedure. Now we need to take the infimum over all estimation procedures to obtain a bound on the minimax risk. In step 4 we reduced the estimation procedure down to a binary decision. Specifically, we only need to consider an estimator of $s_1 \in \{+1, -1\}$. Taking the infimum over all estimators amounts to:

$$\inf_{\mathcal{D}_n \mapsto \hat{s}_1} \Pr[s_1 \neq \hat{s}_1] \tag{67}$$

and it is easy to see that the optimal estimation procedure is to set:

$$\hat{s}_1 = \begin{cases} +1 & \text{if } p(\mathcal{D}_n \mid s_1 = 1) \geq p(\mathcal{D}_n \mid s_1 = -1) \\ -1 & \text{if } p(\mathcal{D}_n \mid s_1 = 1) < p(\mathcal{D}_n \mid s_1 = -1) \end{cases} \tag{68}$$

To implement this rule, let us define:

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{u} = \begin{bmatrix} \phi_1(X_1) \\ \vdots \\ \phi_1(X_n) \end{bmatrix} \tag{69}$$

and consider the inner product $\boldsymbol{y}^\top \boldsymbol{u}$. Recall that $\phi_1(X_i) = 0$ if X_i is not in bin 1. Therefore, when computing $\boldsymbol{y}^\top \boldsymbol{u}$, we can effectively ignore all datapoints falling outside bin 1 (since $\boldsymbol{u}_i = 0$ for these terms). For the terms that do fall within bin 1, the distribution of Y_i conditioned on X_i is normal with mean $s_1 \boldsymbol{u}_i = s_1 \phi_1(X_i)$ and variance σ^2. Thus, conditioned on a realization of $\boldsymbol{u}$, the inner product $\boldsymbol{y}^\top \boldsymbol{u}$ follows a normal distribution since it is a weighted sum of independent normal variables. Further, $\mathbf{E}[\boldsymbol{y}^\top \boldsymbol{u}] = \mathbf{E}[\boldsymbol{y}^\top]\boldsymbol{u} = s_1 \boldsymbol{u}^\top \boldsymbol{u}$. Again, this holds because $\mathbf{E}[\boldsymbol{y}_1] \neq s_1 \boldsymbol{u}_i$ only for terms outside of bin 1, which do not contribute to the inner product. And finally, $\mathbf{Var}[\boldsymbol{y}^\top \boldsymbol{u}] = \sum_i \boldsymbol{u}_i^2 \sigma^2 = \sigma^2 \boldsymbol{u}^\top \boldsymbol{u}$.

We have just shown that the distribution of $\boldsymbol{y}^\top \boldsymbol{u}$ is $\mathcal{N}(s_1 \boldsymbol{u}^\top \boldsymbol{u}, \sigma^2 \boldsymbol{u}^\top \boldsymbol{u})$. Thus, the distribution of $\boldsymbol{y}^\top \boldsymbol{u} / \|\boldsymbol{u}\|$ is equal to $\mathcal{N}(s_1 \|\boldsymbol{u}\|, \sigma^2)$. We can use this to evaluate whether $p(\mathcal{D}_n \mid s_1 = 1)$ is greater than or less than $p(\mathcal{D}_n \mid s_1 = -1)$, which gives us the optimal estimation procedure. Specifically, our estimate is:

$$\hat{s}_1 = \begin{cases} +1 & \text{if } (\boldsymbol{u}^\top \boldsymbol{y})/\|\boldsymbol{u}\| \geq 0 \\ -1 & \text{if } (\boldsymbol{u}^\top \boldsymbol{y})/\|\boldsymbol{u}\| < 0 \end{cases} \tag{70}$$

The probability that this decision rule makes an error is given by integrating the tail of the normal density with σ^2 variance, yielding:

$$\inf_{\mathcal{D}_n \mapsto \hat{s}_1} \Pr[s_1 \neq \hat{s}_1] = \mathbf{E}_{\mathcal{D}_n}[\Phi(-\|\boldsymbol{u}\|/\sigma)] \tag{71}$$

where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution with unit variance. The expectation on the right hand side is required because $\boldsymbol{u}$ is still a random variable (it depends on the sampled values of $X_1, \ldots, X_n$).

Now comes a brilliant observation. Notice that $\Phi(\cdot)$ a convex non-decreasing function over the interval $(-\infty, 0]$ and further $-(1/\sigma)\sqrt{\cdot}$ is a convex function. The composition of these functions is convex (see sec. 3.2.4 in Boyd and Vandenberghe [1]), so $\Phi(-(1/\sigma)\sqrt{\boldsymbol{u}^\top \boldsymbol{u}})$ is a convex function

of $\boldsymbol{u}^\top \boldsymbol{u}$.[6] This permits us to use Jensen's inequality:

$$\mathbf{E}\left[\Phi(-\|\boldsymbol{u}\|/\sigma)\right] \geq \Phi\left(\frac{-1}{\sigma}\sqrt{\sum_{i=1}^{n}\mathbf{E}\phi_1^2(X_i)}\right) = \Phi\left(\frac{-1}{\sigma}\sqrt{n\mathbf{E}\phi_1^2(X_i)}\right) = \Phi\left(\frac{-1}{\sigma}\sqrt{nJ^{-2p-1}\int_0^1 \tilde{\phi}^2(x)dx}\right)$$

Where the final equality follows from eq. (63). Plugging this into eq. (66), we finally obtain a lower bound on the minimax risk:

$$\inf_{\mathcal{D}_n \mapsto \hat{f}} \sup_{P_{X,Y}} \mathbf{E}\|f(X) - \hat{f}(X)\|^2 \geq \left(J^{-2p}\int_0^1 \tilde{\phi}^2(x)dx\right)\Phi\left(\frac{-1}{\sigma}\sqrt{nJ^{-2p-1}\int_0^1 \tilde{\phi}^2(x)dx}\right) \tag{72}$$

which is valid for any choice of J. Let us modify the bound to make the dependence on L explicit. Recall that $\tilde{\phi}$ was defined as a (p, L)-Hölder-smooth function. Define $\varphi(x) = (1/L)\cdot\tilde{\phi}(x)$ and verify that φ is $(p, 1)$-Hölder-smooth. Define $C_1 = \int_0^1 \varphi^2(x)dx$ as an absolute constant. Plugging $\int_0^1 \tilde{\phi}^2(x)dx = L^2\int_0^1 \varphi^2(x)dx = C_1L^2$ into eq. (72), our lower bound becomes

$$\left(C_1J^{-2p}L^2\right)\Phi\left(-\sqrt{C_1J^{-2p-1}L^2n\sigma^{-2}}\right) \tag{73}$$

To finish the proof, we need to come up with a good choice of J that makes this lower bound as large as possible. We allow ourselves to set J adaptively and adversarially based on the values of n, L, and σ^2.[7]

Notice that $\Phi(-\sqrt{z})$ decays to zero very rapidly as z increases (*The reader should think about this and maybe check it out numerically!*). Thus, the second term in eq. (73) substantially weakens the lower bound as either n or L become large (or σ^2 becomes small). From this intuition, it turns out that a good choice is to set J so that the second term in eq. (73) becomes the constant $\Phi(-\sqrt{C_1})$. Thus, we choose:[8]

$$J = (L^2n\sigma^{-2})^{\frac{1}{2p+1}} \tag{74}$$

which means that $J^{-2p-1} = L^{-2}n^{-1}\sigma^2$ and also:

$$J^{-2p} = JL^{-2}n^{-1}\sigma^2 = (L^2n\sigma^{-2})^{\frac{1}{2p+1}}L^{-2}n^{-1}\sigma^2 = L^{\frac{2}{2p+1}}L^{-2}n^{\frac{-2p}{2p+1}}\sigma^{\frac{4p}{2p+1}} \tag{75}$$

Using the substitutions above, we find that this choice of J gives us the lower bound:

$$\left(C_1L^{\frac{2}{2p+1}}n^{\frac{-2p}{2p+1}}\sigma^{\frac{4p}{2p+1}}\right)\Phi\left(\sqrt{C_1}\right) \tag{76}$$

so we have proven the claimed result:

$$\inf_{\mathcal{D}_n \mapsto \hat{f}} \sup_{P_{X,Y}} \mathbf{E}\|f(X) - \hat{f}(X)\|^2 \geq C_2L^{\frac{2}{2p+1}}n^{\frac{-2p}{2p+1}}\sigma^{\frac{4p}{2p+1}} \tag{77}$$

[6] One can also verify this directly computing the second derivative of $\Phi(-\sqrt{z})$ and showing that it is nonnegative. *Hint: use the Leibniz's integral rule to differentiate the cumulative density function.*

[7] It is possible to relax this adversarial posture and still recover the same essential lower bound. However, this requires a more complicated construction and proof. See section 3.3 in Györfi et al. [5] for details.

[8] Note that we are cheating here by not restricting J to be an integer (which we require since J is the number of bins we are using to partition the unit interval). The reader can verify that rounding upwards, i.e. choosing $J = \left\lceil (L^2n)^{\frac{1}{2p+1}} \right\rceil$, works to establish the claim.

with constant:

$$C_2 = \left(\int_0^1 \varphi^2(x)dx\right) \cdot \Phi\left(\sqrt{\int_0^1 \varphi^2(x)dx}\right) \tag{78}$$

Question: The dependence on L and n is optimal in the above result. But is the dependence on σ^2 optimal?

2.2 Proof for any d and $p > 0$

[To be completed at a later date]

3 Upper bounds on minimax risk

In the last chapter we proved a lower bound on minimax risk for estimating a (p, L)-Hölder-smooth function with respect to a quadratic loss. In this chapter, our goal is to prove an upper bound that matches up to a multiplicative constant. We will restrict our attention to the special cases that are easy to prove. Our first proof assumes $d = 1$, but works for any p. Our second proof assumes that $p = 1$, but works for any d. Subsequent chapters will provide upper bounds in greater generality.

The two bounds we prove come with an explicit estimators that can be used in practice. These estimators are said to be "minimax optimal" since their worst-case risk is equal to the minimax risk, up to multiplicative factors.

3.1 Proof when $d = 1$ and any $p > 0$

Our proof follows section 1.6 of Tsybakov [11]. The high level idea is to locally estimate $f(x)$ with a p-degree polynomial. This is called a *local polynomial estimator*.

The target function $f : [0, 1] \mapsto \mathbb{R}$ is (p, L)-Hölder smooth by assumption. We will consider p to be an integer for simplicity, but the proof also works for fractional values of p with minimal changes.

The smoothness assumption means that we can differentiate the target function p times almost everywhere. Thus, we can form a Taylor approximation around a point $x \in [0, 1]$ as follows:

$$f(z) \approx f(x) + f^{(1)}(x)(z - x) + \cdots + f^{(p)}(x)\frac{(z - x)^p}{p!} = \sum_{k=0}^{p} f^{(k)}(x)\frac{(z - x)^k}{k!} \tag{79}$$

We proceed by introducing a *bandwidth parameter*, $h > 0$. We are free to choose h to be any value and its role will become clear later. We multiply and divide each term inside the sum by h^k to obtain:

$$f(z) \approx \sum_{k=0}^{p} h^k f^{(k)}(x)\frac{(z - x)^k}{h^k k!} = \sum_{k=0}^{p} \theta_k(x) u_k(z - x) \tag{80}$$

In the final step, we have defined and substituted:

$$\theta_k(x) = h^k f^{(k)}(x) \qquad u_k(z - x) = \frac{(z - x)^k}{h^k k!} \tag{81}$$

for $k = 0, \ldots, p$. Notice that $\theta_0(x) = f(x)$, so finding a good estimate of $\theta_0, \ldots \theta_p$ immediately gives us a good estimate of the target function.

We observe $(X_1, Y_1), \ldots (X_n, Y_n)$. Intuitively, if we Taylor expand f around a point $x \in [0, 1]$, we should be able to approximate $f(X_i)$ accurately so long as x is sufficiently close to X_i. In other words, we should be able to find coefficients $\hat{\theta}_1(x), \ldots, \hat{\theta}_n(x)$ to accurately predict Y_i, but only if $|X_i - x|$ is relatively small. This motivates us to solve the following weighted least-squares problem:

$$\underset{\hat{\theta}_1(x), \ldots, \hat{\theta}_p(x)}{\text{minimize}} \quad \sum_{i=1}^{n} \left(Y_i - \sum_{k=0}^{p} \hat{\theta}_k(x) u_k(X_i - x) \right)^2 K\left(\frac{X_i - x}{h}\right) \tag{82}$$

where $K : \mathbb{R} \mapsto \mathbb{R}$ is a weighting function or "kernel." In the present context,[9] the kernel is a nonnegative function that integrates to one and which satisfies $K(z) = 0$ for $|z| > 1$. There are many choices for this kernel, but two simple ones are the rectangular and triangular kernels, respectively defined:

$$K(z) = \frac{1}{2}\mathbf{1}[|z| < 1] \qquad \text{and} \qquad K(z) = (1 - |z|)\mathbf{1}[|z| < 1] \tag{83}$$

where $\mathbf{1}[\cdot]$ is the indicator function.

We can interpret eq. (80) as fitting a degree p polynomial using datapoints (X_i, Y_i) that satisfy $|x - X_i| \leq h$ and ignoring datapoints for which $|x - X_i| > h$. Thus, we see that the bandwidth parameter determines the size of the local averaging window. The rectangular kernel places equal weights on all datapoints falling within this window, while the triangular kernel (and many other kernels) place smaller weight on datapoints close to the boundary of the window.

Equation (82) can be reformulated in terms of matrix-vector multiplications as follows:

$$\underset{\hat{\boldsymbol{\theta}}}{\text{minimize}} \quad (\boldsymbol{y} - \boldsymbol{U}\hat{\boldsymbol{\theta}})^\top \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{U}\hat{\boldsymbol{\theta}}) \tag{84}$$

where $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$, and:

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} u_0(X_1 - x) & \dots & u_p(X_1 - x) \\ \vdots & & \vdots \\ u_0(X_n - x) & \dots & u_p(X_n - x) \end{bmatrix}, \quad \boldsymbol{K} = \begin{bmatrix} K((X_1 - x)/h) & & \\ & \ddots & \\ & & K((X_n - x)/h) \end{bmatrix}.$$

The objective function in eq. (84) is quadratic in $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{K}$ is positive semidefinite. Thus, the problem is a convex quadratic program. By differentiating with respect to $\hat{\boldsymbol{\theta}}$ and setting the result to zero we find that the optimal $\hat{\boldsymbol{\theta}}$ satisfies:

$$\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{U}\hat{\boldsymbol{\theta}} = \boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{y} \tag{85}$$

Assuming that $\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{U}$ is nonsingular, the solution is:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{y} \tag{86}$$

Recalling that $\theta_0(x) = f(x)$ we recover our estimate of $f(x)$ by taking the inner product with the vector $\boldsymbol{e}_1 = [\,1 \;\; 0 \;\; \dots \;\; 0\,]^\top$. Thus, the estimate

$$\hat{f}(x) = \boldsymbol{e}_1^\top (\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{y} \tag{87}$$

is a linear function of $\boldsymbol{y}$. This linear function mapping $\mathbb{R}^n \mapsto \mathbb{R}$ can be represented by a vector $\boldsymbol{w} \in \mathbb{R}^n$, given by:

$$\boldsymbol{w} = (\boldsymbol{e}_1^\top (\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{K})^\top = \boldsymbol{K}\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{K}\boldsymbol{U})^{-1}\boldsymbol{e}_1 \tag{88}$$

Note that $\boldsymbol{w}$ is not a function of $\boldsymbol{y}$. This turns out to have a very interesting consequence. Suppose for a moment that the target function were some polynomial $Q(x)$ with degree less than or equal

[9] The reader should be warned that a "kernel" function is an overloaded term. In other contexts, such as in Gaussian process regression and kernel ridge regression a "kernel" will take on a different meaning.

to p and there were no noise. Intuitively, our fitting procedure should work perfectly in this setting so we would have $\boldsymbol{w}^\top \boldsymbol{y} = Q(x)$. Suppose that $Q(x) = 1$ (this is indeed a polynomial less than degree p). Then, $\boldsymbol{y} = \mathbf{1}$ and we have $\boldsymbol{w}^\top \boldsymbol{y} = \boldsymbol{w}^\top \mathbf{1} = 1$. Thus, we have proven:

Lemma 3.1

Let $\boldsymbol{w}$ be defined as in eq. (88). The sum of elements $\sum_{i=1}^n \boldsymbol{w}_i$ equals 1.

Having defined our estimator now proceed to upper bounding the risk (expected squared error). We use the bias-variance decomposition of the risk at a fixed point $x \in [0, 1]$

$$\mathbf{E}[(f(x) - \hat{f}(x))^2] = (f(x) - \mathbf{E}[\hat{f}(x)])^2 + \mathbf{E}[(f(x) - \mathbf{E}[\hat{f}(x)])^2] \tag{89}$$

All expectations here are taken with respect to random observations of $Y_1, \ldots, Y_n$ which induce randomness in the estimator $\hat{f}$. It is important to note that x is fixed (not random).

First, we upper bound the bias. We begin by reformulating:

$$\begin{aligned}
f(x) - \mathbf{E}[\hat{f}(x)] &= f(x) - \boldsymbol{w}^\top \mathbf{E}[\boldsymbol{y}] && (90)\\
&= \boldsymbol{w}^\top (f(x)\mathbf{1} - \mathbf{E}[\boldsymbol{y}]) && \text{(since } \boldsymbol{w}^\top \mathbf{1} = 1 \text{ from lemma 3.1)} \quad (91)\\
&= \sum_{i=1}^n \boldsymbol{w}_i (f(x) - f(X_i)) && (\mathbf{E}[\boldsymbol{y}_i] = f(X_i) \text{ since noise is mean zero}) \quad (92)\\
&= \sum_{i=1}^n \boldsymbol{w}_i (f^{(\ell)}(z) - f(X_i)) && \text{(Taylor's theorem????} \quad (93)
\end{aligned}$$

3.2 Proof for $p = 1$ and $d \geq 2$

The k-nearest neighbor regression method gives a minimax optimal bound in this setting. The high level idea is to estimate $f(\boldsymbol{x})$ by taking a local average the K nearest points in the training set.

Formally, given paired observations $\mathcal{D}_n = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ sampled independently from $P_{X,Y}$, define for any $\boldsymbol{x} \in \mathbb{R}^d$ the sequence $(X_{(1,\boldsymbol{x})}, Y_{(1,\boldsymbol{x})}), (X_{(2,\boldsymbol{x})}, Y_{(2,\boldsymbol{x})}), \ldots (X_{(n,\boldsymbol{x})}, Y_{(n,\boldsymbol{x})})$ such that the observations are in ascending order of their distance to $\boldsymbol{x}$:

$$\|X_{(1,\boldsymbol{x})} - \boldsymbol{x}\| \leq \|X_{(2,\boldsymbol{x})} - \boldsymbol{x}\| \leq \cdots \leq \|X_{(n,\boldsymbol{x})} - \boldsymbol{x}\| \tag{94}$$

The mapping from the original data $\{(X_i, Y_i)\}$ to ordered data $\{(X_{(i,\boldsymbol{x})}, Y_{(i,\boldsymbol{x})})\}$ can be viewed as deterministic, since we can break ties arbitrarily. When the reference point $\boldsymbol{x}$ is clear from context we will simply use $X_{(i)}$ and $Y_{(i)}$ to denote the i^{th} nearest neighbor instead of the more explicit notation above.

Equipped with this notation, the k-nearest neighbor regression estimate can be written as:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{k} \sum_{i=1}^k Y_{(i,\boldsymbol{x})} \tag{95}$$

One can show that the risk of this estimator is minimax optimal for $p = 1$, but suboptimal for higher smoothness classes (see Györfi et al. [5], chapter 6).

Step 1: *Condition on independent variables, and use bias-variance decomposition*

The risk can be expressed using iterated expectations:

$$R(f, \hat{f}) = \mathbf{E}_{X_0, \mathcal{D}_n}[(f(X_0) - \hat{f}(X_0))^2] \tag{96}$$
$$= \mathbf{E}_{X_0, X_1, \ldots, X_n}[\mathbf{E}_{Y_1, \ldots, Y_n}[(f(X_0) - \hat{f}(X_0))^2 \mid X_0, X_1, \ldots, X_n]] \tag{97}$$

Our approach will begin by obtaining an upper bound on the inner conditional expectation. To proceed we use the bias-variance decomposition:

$$\mathbf{E}_{Y_1, \ldots, Y_n}[(f(X_0) - \hat{f}(X_0))^2 \mid X_0, \ldots, X_n] = B(X_0, \ldots, X_n) + V(X_0, \ldots, X_n) \tag{98}$$

where the bias B and variance V are respectively:

$$B(X_0, \ldots, X_n) = (f(X_0) - \mathbf{E}[\hat{f}](X_0))^2 \tag{99}$$
$$V(X_0, \ldots, X_n) = \mathbf{E}(\hat{f}(X_0) - \mathbf{E}[\hat{f}](X_0))^2 \tag{100}$$

All of the expectations above are over $Y_1, \ldots, Y_n$, conditioning on $X_0, \ldots, X_n$.

Step 2: *Upper bound the variance, conditioned on $X_0, \ldots X_n$*

Throughout the rest of the proof we will only require the nearest neighbors of X_0, so we will use $X_{(i)}$ to denote the i^{th} nearest neighbor to X_0. Note that:

$$\mathbf{E}[\hat{f}](X_0) = \frac{1}{k} \sum_{i=1}^{k} \mathbf{E}[Y_{(i)}] = \frac{1}{k} \sum_{i=1}^{k} f(X_{(i)}) \tag{101}$$

Recall that the residuals $\epsilon_{(i)} = Y_{(i)} - f(X_{(i,x)}$ are mean zero, independent random variables and by assumption $\mathbf{Var}[\epsilon_i] \leq \sigma^2$. Using this, we have:

$$V(X_0, \ldots, X_n) = \mathbf{E}_{Y_1, \ldots, Y_n}\left[\left(\hat{f}(X_0) - \mathbf{E}[\hat{f}](X_0)\right)^2\right] \tag{102}$$
$$= \mathbf{E}_{Y_1, \ldots, Y_n}\left[\left(Y_{(i)} - \frac{1}{k} \sum_{i=1}^{k} f(X_{(i)})\right)^2\right] \tag{103}$$
$$= \mathbf{E}_{Y_1, \ldots, Y_n}\left[\left(\frac{1}{k} \sum_{i=1}^{k} \left(Y_{(i)} - f(X_{(i)})\right)\right)^2\right] \tag{104}$$
$$= \mathbf{E}_{Y_1, \ldots, Y_n}\left[\frac{1}{k^2}\left(\sum_{i=1}^{k} \epsilon_{(i)}\right)^2\right] \tag{105}$$
$$\leq \frac{\sigma^2}{k} \tag{106}$$

Step 3: *Upper bound the bias, conditioned on $X_0, \ldots X_n$*

Using the Lipshitz assumption, we find:

$$B(X_0, \ldots, X_n) = \left(f(X_0) - \mathbf{E}[\hat{f}](X_0)\right)^2 \tag{107}$$

$$= \left(f(X_0) - \frac{1}{k} \sum_{i=1}^{k} f(X_{(i)}) \right)^2 \tag{108}$$

$$= \left(\frac{1}{k} \sum_{i=1}^{k} (f(X_0) - f(X_{(i)})) \right)^2 \tag{109}$$

$$\leq \left(\frac{1}{k} \sum_{i=1}^{k} \left| f(X_0) - f(X_{(i)}) \right| \right)^2 \tag{110}$$

$$\leq \left(\frac{L}{k} \sum_{i=1}^{k} \left\| X_0 - X_{(i)} \right\|_2 \right)^2 \tag{111}$$

Step 4: *Take the expectation over $X_0, X_1, \ldots, X_n$*

Returning to eq. (97) and using our bounds on the bias and variance, we have thus far shown:

$$R(f, \hat{f}) = \mathbf{E}_{X_0,\ldots,X_n}[B(X_0, \ldots, X_n) + V(X_0, \ldots, X_n) \mid X_0, \ldots, X_n] \tag{112}$$

$$\leq \mathbf{E}_{X_0,\ldots,X_n} \left[\left(\frac{L}{k} \sum_{i=1}^{k} \left\| X_0 - X_{(i)} \right\|_2 \right)^2 + \frac{\sigma^2}{k} \right] \tag{113}$$

$$\leq \mathbf{E}_{X_0,\ldots,X_n} \left[L^2 \| X_0 - X_{(k)} \|_2^2 + \frac{\sigma^2}{k} \right] \tag{114}$$

where the final inequality follows from:

$$\frac{1}{k} \sum_{i=1}^{k} \| X_0 - X_{(i)} \|_2 \leq \| X_0 - X_{(k)} \|_2 \tag{115}$$

which uniformly holds over all realizations of $X_1, \ldots, X_n$. Thus, we can upper bound the risk as:

$$R(f, \hat{f}) \leq \frac{\sigma^2}{k} + L^2 \left(\mathbf{E}_{X_0,\ldots,X_n} \| X_0 - X_{(k)} \|_2^2 \right) \tag{116}$$

where the expected squared distance to the k^{th} nearest neighbor is the only unknown term.

Step 5: *Upper bound the distance to the k^{th} nearest neighbor*

Theorem 2.4 in Devroye and Biau [3] proves that

$$\mathbf{E} \| X_0 - X_{(k)} \|_2^2 \leq C_1 \left(\frac{k}{n} \right)^{2/d} \tag{117}$$

where

$$C_1 = \left(\frac{2^{3+\frac{2}{d}} (1 + \sqrt{d})^2}{\pi} \right) \Gamma \left(\frac{d}{2} + 1 \right)^{2/d} \tag{118}$$

Step 6: *Choose the optimal number of neighbors to use*

We have proven that for any choice of k, the nearest neighbor regression model has risk:

$$R(f, \hat{f}) \leq \frac{\sigma^2}{k} + C_1 L^2 \left(\frac{k}{n} \right)^{2/d} \tag{119}$$

The right hand side is a convex function of k, since the second derivative is nonnegative:

$$\frac{2\sigma^2}{k^2} + C_1 L^2 \left(\frac{1}{n}\right)^{2/d} \left(\frac{2}{d}\right)\left(\frac{2}{d} - 1\right) k^{(2/d-2)} \geq 0 \tag{120}$$

because of our assumption that $d \geq 2$ and because $k > 0$. Thus we proceed to set the first derivative to zero and solve for k. We get that $k = C_2 n^{2/(2+d)}$ for a constant C_2 that is independent of n:

$$\frac{d}{dk}\left[\frac{\sigma^2}{k} + C_1 L^2 \left(\frac{k}{n}\right)^{2/d}\right] = 0 \tag{121}$$

$$\Rightarrow \frac{-\sigma^2}{k^2} + C_1 L^2 n^{-2/d} \left(\frac{2}{d}\right) k^{(2/d-1)} = 0 \tag{122}$$

$$\Rightarrow -\sigma^2 + C_1 L^2 n^{-2/d} \left(\frac{2}{d}\right) k^{(2/d-1+2)} = 0 \tag{123}$$

$$\Rightarrow k^{(2+d)/d} = C_1^{-1} L^{-2} \left(\frac{d}{2}\right) \sigma^2 n^{2/d} \tag{124}$$

$$\Rightarrow k = C_1^{-d/(2+d)} L^{-2d/(2+d)} \left(\frac{d}{2}\right)^{d/(2+d)} \sigma^{2d/(2+d)} n^{2/(2+d)} \tag{125}$$

$$\Rightarrow k = C_2 n^{2/(2+d)} \tag{126}$$

Plugging this into our bound we find:

$$R(f, \hat{f}) \leq C_2 \sigma^2 n^{-2/(2+d)} + C_1 L^2 C_2^{2/d} n^{-2/(2+d)} = C_3 n^{-2/(2+d)} \tag{127}$$

so we obtain the optimal minimax rate in n up to a multiplicative constant C_3.

4 Nonparametric regression via basis functions

In this chapter we'll explore estimators that can be expressed as the solution to an optimization problem. Specifically, we seek to:

$$\underset{\hat{f}\in\mathcal{F}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2 \tag{128}$$

where $\mathcal{F}$ is a set of candidate functions. We will see that this optimization framework is quite flexible and will provide us a foundation to develop Bayesian approaches to nonparametric regression.

Solving eq. (128) seems hard because we are optimizing over functions, which are infinite-dimensional objects. A simple and effective idea is to approximate f as a linear combination of a finite number of *basis functions*. Thus, our model takes the form:

$$f(x) \approx \hat{f}(x) = \sum_{i=1}^{J} w_j \phi_j(x) \tag{129}$$

where $\phi_1, \ldots, \phi_J$ are basis functions mapping $\mathcal{X} \mapsto \mathbb{R}$, and $w_1, \ldots, w_n$ are scalar coefficients that we adjust to minimize the prediction error on the training set. Thus, $\mathcal{F}$ is a J-dimensional linear subspace:

$$\mathcal{F}_J = \left\{ \boldsymbol{x} \mapsto \sum_{j=1}^{J} w_j \phi_j(\boldsymbol{x}) \;\middle|\; w_1, \ldots, w_J \in \mathbb{R}^J \right\}. \tag{130}$$

and eq. (128) is equivalent to:

$$\underset{\boldsymbol{w}}{\text{minimize}} \;\; \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}\|_2^2 \tag{131}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{n\times J}$ is a matrix of evaluated basis functions $\boldsymbol{\Phi}_{ij} = \phi_j(X_i)$, $\boldsymbol{w} \in \mathbb{R}^J$ is a vector of learned coefficents, and $[\boldsymbol{y}]_i = Y_i$. This is a *linear least-squares problem*. Assuming that $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ is full rank, the solution is given by:

$$\hat{f}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \hat{\boldsymbol{w}} \quad \text{where} \quad \hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \tag{132}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = \left[\phi_1(\boldsymbol{x}) \;,\; \ldots \;,\; \phi_J(\boldsymbol{x})\right]^\top$.

Voilà! We have turned a seemingly intractable optimization problem (optimization over infinite-dimensional function spaces) into a classical and well-understood ordinary least-squares problem. Doing so comes at a cost. We are fitting a parametric model (parametrized by $w_1, \ldots, w_J$) to approximate a general (not necessarily parametric) target function $f \notin \mathcal{F}$. Thus, we always incur some *approximation error*, given by:

$$\min_{g\in\mathcal{F}_J} \mathbf{E}[(g(X) - f(X))^2] \tag{133}$$

where the expectation is taken with respect to $X \sim P_X$. Note that if we were to assume that $f \in \mathcal{F}_J$, then the approximation error would be zero and we would then be studying a parametric (as opposed to nonparametric) regression problem.

4.1 Choosing basis functions

A natural idea is to specify an infinite sequence of basis functions $\{\phi_1, \phi_2, \dots\}$ and set J as a finite truncation of this series. For example, consider estimating a univariate function $f : [0,1] \mapsto \mathbb{R}$. One choice is to use Fourier basis functions:

$$\phi_1(x) = 1/\sqrt{2}, \quad \phi_{2k-1}(x) = \cos(2\pi kx), \quad \phi_{2k}(x) = \sin(2\pi kx), \quad \text{for } k \in \{1, 2, \dots\} \tag{134}$$

Any target function $f : [0,1] \mapsto \mathbb{R}$ that is Hölder smooth can be arbitrarily well approximated by this sequence of basis functions. In particular, if we define:

$$g_J(x) = \sum_{j=1}^{J} w_j^* \phi_j(X) \qquad \text{where } w_j^* = 2\int_0^1 f(x)\phi_j(x)dx. \tag{135}$$

Then it can be shown that as $J \to \infty$ we have that g_J converges uniformly to f. This implies pointwise convergence, so we have $g_J(x) = f(x)$ for any $x \in (0,1)$ in the limit of $J \to \infty$. More importantly, uniform convergence implies that the approximation error goes to zero as we increase the number of basis functions, i.e.

$$\lim_{J\to\infty} \mathbf{E}\left[\left(f(X) - g_J(X)\right)^2\right] = 0 \tag{136}$$

where the expectation is taken with respect to $X \sim P_X$.[10]

add discussion of Legendre / Chebyshev polynomials.

4.2 Analysis of risk with respect to empirical norm

Let's revisit the optimization problem in eq. (128). For any fixed candidate model $g \in \mathcal{F}_J$, including but not necessarily the optimal model $\hat{f}$, the expected squared error on a heldout sample is equal in expectation to the minimized objective:

$$\mathcal{L}_{P_X}(f, g) = \mathbf{E}_{X\sim P_X}[(f(X) - g(X))^2] = \mathbf{E}_{X_1,\dots,X_n}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2\right] \tag{138}$$

Intuitively, the law of large numbers tells us that the empirical mean squared error is not only equal to $\mathcal{L}_{P_X}(f, g)$ in expectation, but very close to it in reality. That is, as $n \to \infty$ we have:

$$\mathbf{Var}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2\right] \to 0 \tag{139}$$

Indeed, we can make this convergence precise without appealing to asymptotics. That is, for finite n one can prove that:

$$\mathcal{L}_{P_X}(f, g) \le \frac{2}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 \tag{140}$$

[10] We can prove that uniform convergence implies convergence in expectation easily in one line:

$$\mathbf{E}\left[\left(f(X) - g_J(X)\right)^2\right] = \int_0^1 \left(f(X) - g_J(X)\right)^2 P_X(dx) \le P_X([0,1])\left(\sup_{x\in[0,1]} |f(x) - g(x)|\right)^2 \to 0 \tag{137}$$

since uniform convergence tells us that this supremum converges to zero.

holds *simultaneously* for all $g \in \mathcal{F}$ with a probability that monotonically increases towards one as n increases. The fact that this bound holds *simultaneously* for all candidate functions is critical, since it means we can apply the bound to our optimal model $\hat{f} \in \mathcal{F}$. See Theorem 14.12 in Wainwright [13] for a precise statement.

In summary, we have articulated a probabilistic upper bound on the risk:

$$R(f, \hat{f}) = \mathbf{E}_{\mathcal{D}}\left[\mathcal{L}_{P_X}(f, \hat{f})\right] \leq \frac{2}{n}\sum_{i=1}^{n}(f(X_i) - \hat{f}(X_i))^2 \tag{141}$$

where the inequality holds with high probability assuming that n is sufficiently large. At some future date I may include more details on this probabilistic upper bound, for now I refer the reader to chapter 14 in Wainwright.

In the rest of the chapter we will focus on the right hand side of eq. (141). We will prove the following.

Result 4.1: Upper bound on empirical mean squared error

Let $\hat{f}$ be a solution to eq. (128). Assuming that $\mathbf{\Phi}^\top\mathbf{\Phi}$ is full rank (i.e. rank = J), we have:

$$\mathbf{E}_{Y_1,\ldots Y_n}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2 \,\middle|\, X_1, \ldots X_n\right] \leq \frac{J\sigma^2}{n} + \frac{1}{n}\left(\min_{g\in\mathcal{F}_J}\sum_{i=1}^{n}(g(X_i) - f(X_i))^2\right)$$

where $\sigma^2 = \sup_{x\in\mathcal{X}} \mathbf{Var}[\epsilon_i \mid X_i = x]$.

This result characterizes the empirical mean squared error loss, conditioned on values of $X_1, \ldots, X_n$. This can be plugged into the right hand side of eq. (141) because $\hat{f}$ is viewed as a fixed function for the purposes of that bound—in other words the only randomness that persists in eq. (141) is the randomness over $X_1, \ldots X_n$.

The result shows that the expected empirical means squared error is equal to the sum of two terms, both of which are proportional to $1/n$. This is the usual "parametric rate" of convergence (todo: explain this more).

Proof of result 4.1

To reduce notational burden we will use $\mathbf{E}^*[\cdot]$ to denote the conditional expectation $\mathbf{E}_{Y_1,\ldots Y_n}[\cdot \mid X_1, \ldots X_n]$.

Step 1: Use the bias-variance decomposition. Define $\bar{f}(x) = \mathbf{E}[\hat{f}](x)$ and verify that:

$$\mathbf{E}^*\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - f(X_i))^2\right] = \mathbf{E}^*\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X_i) - \bar{f}(X_i))^2\right] + \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \bar{f}(X_i))^2 \tag{142}$$

Step 2: The bias term. Define

$$\boldsymbol{f} = \begin{bmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{f}} = \begin{bmatrix} \hat{f}(X_1) \\ \vdots \\ \hat{f}(X_n) \end{bmatrix} \tag{143}$$

Using eq. (132) we have:

$$\hat{f} = \mathbf{\Phi}\hat{w} = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top y \tag{144}$$

since we are conditioning on $X_1, \ldots, X_n$, the matrix $\mathbf{\Phi}$ is not a random variable. Thus, we have:

$$\bar{f} = \mathbf{E}^*[\hat{f}] = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{E}^*[y] = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top f \tag{145}$$

To conclude this step we need to show that $\bar{f}$ coincides with the solution to the optimization problem:

$$\underset{g}{\text{minimize}} \quad \sum_{i=1}^{n}(f(X_i) - g(X_i))^2 \qquad \text{subject to } g \in \mathcal{F}_J \tag{146}$$

Define:

$$g = \begin{bmatrix} g(X_1) \\ \vdots \\ g(X_n) \end{bmatrix} \tag{147}$$

because $g \in \mathcal{F}_J$, there exists a set of coefficents $w_1, \ldots, w_n$ such that $g(X_i) = \sum_j w_j\phi_j(X_i)$. Thus, we can write $g = \mathbf{\Phi}w$ and reformulate the optimization problem:

$$\underset{w}{\text{minimize}} \quad \|f - \mathbf{\Phi}w\|_2^2 \tag{148}$$

As before, this is a least squares problem and plugging in the closed form solution for w gives

$$g = \mathbf{\Phi}w = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top f = \bar{f} \tag{149}$$

as claimed.

Step 3: The variance term. Using eqs. (144) and (145) we have:

$$\mathbf{E}^*\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_i) - \bar{f}(X_i)\right)^2\right] = \frac{1}{n}\mathbf{E}^*\left[\|\hat{f} - \bar{f}\|_2^2\right] = \frac{1}{n}\mathbf{E}^*\left[\|\mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top(y - f)\|_2^2\right] \tag{150}$$

To reduce notational burden let $H = \mathbf{\Phi}(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top$ and note that H is symmetric. Then, some basic algebraic manipulations give:

$$\begin{aligned}
\frac{1}{n}\mathbf{E}^*\left[\|H(y - f)\|_2^2\right] &= \frac{1}{n}\mathbf{E}^*\left[(y - f)^\top H(y - f)\right] && (151)\\
&= \frac{1}{n}\mathbf{E}^*\left[y^\top Hy + f^\top Hf - 2f^\top Hy\right] && (152)\\
&= \frac{1}{n}\left(\mathbf{E}^*\left[y^\top Hy\right] + f^\top Hf - 2f^\top H\mathbf{E}^*[y]\right) && (153)\\
&= \frac{1}{n}\left(\mathbf{E}^*\left[y^\top Hy\right] - f^\top Hf\right) && (154)
\end{aligned}$$

Now we use a classic "trace operator trick." For any two vectors u and v we have that $u^\top v = \text{Tr}[uv^\top]$. Using this and the fact that trace is a linear operator, we have:

$$\mathbf{E}^*\left[y^\top Hy\right] = \mathbf{E}^*\,\text{Tr}\left[Hyy^\top\right] = \text{Tr}\left[H\mathbf{E}^*\left[yy^\top\right]\right] \tag{155}$$

Element (i, j) in the matrix $\mathbf{E}^*[yy^\top]$ is given by:

$$\mathbf{E}^*[Y_iY_j] = \mathbf{E}^*[(f(X_i) + \epsilon_i)(f(X_j) + \epsilon_j)] \tag{156}$$

$$= \mathbf{E}^*[f(X_i)f(X_j) + \epsilon_i f(X_j) + f(X_i)\epsilon_j + \epsilon_i\epsilon_j] \tag{157}$$

$$= f(X_i)f(X_j) + \underbrace{\mathbf{E}^*[\epsilon_i]}_{=0} f(X_j) + f(X_i)\underbrace{\mathbf{E}^*[\epsilon_j]}_{=0} + \underbrace{\mathbf{E}^*[\epsilon_i\epsilon_j]}_{\leq\sigma^2\delta_{ij}} \tag{158}$$

Thus, we have

$$\mathbf{E}^*\left[\boldsymbol{y}^\top \boldsymbol{H}\boldsymbol{y}\right] \leq \mathrm{Tr}[\boldsymbol{H}(\boldsymbol{f}\boldsymbol{f}^\top + \sigma^2\boldsymbol{I})] = \mathrm{Tr}[\boldsymbol{H}\boldsymbol{f}\boldsymbol{f}^\top] + \mathrm{Tr}[\sigma^2\boldsymbol{H}] = \boldsymbol{f}^\top\boldsymbol{H}\boldsymbol{f} + \sigma^2 J \tag{159}$$

Appendix A Miscellaneous Proofs

Lemma A.1

Let f be a real-valued function with K continuous derivatives. Then for any x, there exists a number ξ that lies on the closed interval between zero and x and which satisfies:

$$f(x) = \sum_{k=0}^{K} \frac{f^{(k)}(0)}{k!} x^k + \frac{f^{(K)}(\xi) - f^{(K)}(0)}{K!} x^K \tag{160}$$

Proof. Define $r(x)$ as the remainder for the $K-1$ order Taylor approximation of f, expanded around zero:

$$r(x) = f(x) - \sum_{k=0}^{K-1} \frac{f^{(k)}(0)}{k!} x^k \tag{161}$$

Note that r is K times differentiable and $r^{(K)}(x) = f^{(K)}(x)$ since the Kth derivative of the second term above is zero. Also note that $r(0) = r^{(1)}(0) = \cdots = r^{(K-1)}(0) = 0$.

Now we use an elementary result in calculus, *Cauchy's mean value theorem*. Briefly, this states that for any two continuously differentiable functions g and h, and scalars $a < b$, there exists a scalar c satisfying $a \leq c \leq b$ and:

$$\frac{g(a) - g(b)}{h(a) - h(b)} = \frac{g'(c)}{h'(c)} \tag{162}$$

We can apply this result iteratively K times as follows:

$$\frac{r(x)}{x^K} = \frac{r(x) - r(0)}{x^K - 0^K} = \frac{r^{(1)}(x_1)}{K x_1^{K-1}} = \frac{r^{(1)}(x_1) - r^{(1)}(0)}{K x_1^{K-1} - K \cdot 0^{K-1}} = \frac{r^{(2)}(x_2)}{K(K-1) x_2^{K-2}} = \cdots = \frac{r^{(K)}(x_K)}{K!} \tag{163}$$

where $x_1, \ldots, x_K$ is a monotonic sequence of scalars satisfying $|x_1| \leq \cdots \leq |x_K| \leq |x|$. Recalling that $r^{(K)}(x) = f^{(K)}(x)$ we can conclude that:

$$r(x) = \frac{f^{(K)}(x_K)}{K!} x^K \tag{164}$$

which is called the *Lagrange form* of the remainder in Taylor's approximation. Now we choose $\xi = x_K$ to conclude:

$$f(x) = \sum_{k=0}^{K-1} \frac{f^{(k)}(0)}{k!} x^k + \frac{f^{(K)}(\xi)}{K!} x^K = \sum_{k=0}^{K} \frac{f^{(k)}(0)}{k!} x^k + \frac{f^{(K)}(\xi) - f^{(K)}(0)}{K!} x^K \tag{165}$$

where we simply add and subtract a term of $f^{(K)}(0) x^K / K!$ in the final step. □

References

[1] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Ismaël Castillo. "Lower bounds for posterior rates with Gaussian process priors". *Electronic Journal of Statistics* 2.none (2008), pp. 1281–1299.

[3] Luc Devroye and Gerard Biau. *Lectures on the nearest neighbor method*. 1st ed. Springer Series in the Data Sciences. Cham, Switzerland: Springer International Publishing, 2015.

[4] Nathaniel Eldredge. "Analysis and probability on infinite-dimensional spaces". *arXiv preprint arXiv:1607.03591* (2016).

[5] László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A distribution-free theory of nonparametric regression*. Vol. 1. Springer, 2002.

[6] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010.

[7] Ron Kohavi, David H Wolpert, et al. "Bias plus variance decomposition for zero-one loss functions". *ICML*. Vol. 96. Citeseer. 1996, pp. 275–283.

[8] Osvaldo A. Martin, Ravin Kumar, and Junpeng Lao. *Bayesian Modeling and Computation in Python*. Boca Raton, 2021.

[9] Bernard W Silverman. "Spline smoothing: the equivalent variable kernel method". *The annals of Statistics* (1984), pp. 898–916.

[10] Charles J. Stone. "Optimal Global Rates of Convergence for Nonparametric Regression". *The Annals of Statistics* 10.4 (1982), pp. 1040–1053.

[11] Alexandre B Tsybakov. "Introduction to Nonparametric Estimation". Springer Series in Statistics. New York, NY: Springer New York, 2009.

[12] Aad Van Der Vaart and Harry Van Zanten. "Information Rates of Nonparametric Gaussian Process Methods." *Journal of Machine Learning Research* 12.6 (2011).

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