Theory of nonparametric regression

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

Last Updated: January 7, 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 $\mathbf{Var}[\epsilon \mid X = x]$.

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:

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

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:

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

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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ as a sequence of random variables. The *estimation procedure* or *estimator* is a deterministic mapping[2] of these random variables to a function,

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

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 [4] 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 $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) = |y - \hat{y}|$ 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. [5]). 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 $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. 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:

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

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:

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

□

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 $\hat{f}(x) = 0$. This estimator has zero risk when the target function is $f(x) = 0$, but clearly has very large risk in almost any other circumstance (e.g. $f(x) = c$ 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. [3]. 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 [8].

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:

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

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 $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ 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. [6]). **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 [2].

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 [1] for the first inequality (lower bound), and consult Van Der Vaart and Van Zanten [9] 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 [7]).

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.

References

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

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

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

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

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

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

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

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

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