Lecture 15: Denoising II – Adapting to Unknown Smoothness

1 Maximum Penalized Likelihood Estimators

Let's recap the last two lectures. Suppose we have data $\{x_i, Y_i\}_{i=1}^n$ where $\{x_i\}$ are deterministic sampling points and $\{Y_i\}$ are independently distributed according to

$$Y_i \sim p_{f^*(x_i)}(y), \quad i = 1, \ldots, n$$

where $p_{f^*(x_i)}$ is a density function parameterized by $f^*(x_i)$, the evaluation of the target function at the point x_i. Perhaps the most familiar setting is the Gaussian case where

$$p_{f^*(x_i)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(y - f^*(x_i))^2/2\sigma^2\right]$$

Suppose that we have a countable collection of models $\mathcal{F}$, and a complexity $c(f)$ assigned to each $f \in \mathcal{F}$ such that $\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1$. The empirical risk of a model f is defined to be the average negative log-likelihood:

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n -\log p_{f(x_i)}(Y_i)$$

The risk is the expectation of the empirical risk:

$$R(f) = \frac{1}{n} \sum_{i=1}^n -E[\log p_{f(x_i)}(Y_i)]$$

And the excess riks (also called the *regret*) is

$$R(f) - R(f^*) = \frac{1}{n} \sum_{i=1}^n K(p_{f(x_i)}, p_{f^*(x_i)})$$

where for any two densities p and q the number $K(p, q)$ is the Kullback-Leibler divergence which is defined as $K(p, q) = \int \log(q/p)q$. Thus, the natural measure of the approximation error or bias for ML estimators is the KL divergence.

The Maximum Penalized Likelihood Estimator (MPLE) is defined as

$$\widehat{f}_n = \arg\max_{f \in \mathcal{F}} \left\{ \widehat{R}_n(f) + \frac{2\log 2\, c(f)}{n} \right\}$$

and satisfies the following bound

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^n E[H^2(p_{\widehat{f}_n(x_i)}, p_{f^*(x_i)})] &\leq -\frac{2}{n} \sum_{i=1}^n E[\log A(p_{\widehat{f}_n(x_i)}, p_{f^*(x_i)})] \\
&\leq \min_{f \in F} \left\{ \frac{1}{n} \sum_{i=1}^n K(p_{f(x_i)}, p_{f^*(x_i)}) + \frac{2\log 2\, c(f)}{n} \right\}
\end{aligned}
$$

where for any two densities p and q, $H^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2$ is the squared Hellinger distance and $A(p, q) = \int \sqrt{pq}$ is the affinity of p and q.

In the case of the Gaussian noise model

$$Y_i = f^*(x_i) + W_i, \quad i = 1, \ldots, n$$

the MPLE and bound are given by

$$\widehat{f}_n = \arg \max_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(f(x_i) - Y_i)^2}{2\sigma^2} + \frac{2 \log 2 \, c(f)}{n} \right\}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} E[(\widehat{f}_n(x_i) - f^*(x_i))^2] \leq \min_{f \in F} \left\{ \frac{2}{n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2 + \frac{8 \log 2\sigma^2 \, c(f)}{n} \right\}$$

2   Linear Models

A particular class of models that is very useful and widespread are linear models. These models are defined in terms of an $n \times k$ *basis* matrix H and a $k \times 1$ parameter vector θ:

$$f = H\theta = \sum_{j=1}^{k} \theta_j h_j$$

where h_j is the j-th column of H. Notice that we are considering the model f as an $n \times 1$ column vector. Let us re-state the Gaussian MPLE problem in this vector notation. Let $Y = [Y_1, \ldots, Y_n]^T$, $f^* = [f * (1/n), f^*(2/n), \ldots, f^*(1)]^T$, and $W = [W_1, \ldots, W_n]^T$ so that we can write our observations as

$$Y = f^* + W$$

Define the model collection $\mathcal{F}_H = \{f \in R^n : f = H\theta, \theta \in R^k\}$ This collection is the subspace of all n-vectors lying in the span of the columns of H. This is the model collection we would like to work with, but it is uncountable. To remedy this, let us assume that each coefficient θ_i lies within the interval $[-B, B]$ and furthermore let us quantize the coefficient to one of m levels (uniformly distributed) within this interval. In other words, we discretize the model space by requiring that $\theta_i \in \{-B, \ldots, -B/(m/2), 0, B/(m/2), \ldots, B\}$. Denote the quantized version of $\mathcal{F}_H$ as $\mathcal{F}_{H,m}$. Note that $\mathcal{F}_{H,m}$ contains m^k elements, and so we can encode them uniquely with codes of length $c(f) = k \log m$. Since the dependence is only logarithmic in m, we can quantize extremely finely and cover $\mathcal{F}_H$ almost perfectly.

With this model space, we can now consider the MPLE

$$\widehat{f}_n = \arg \min_{f \in \mathcal{F}_{H,m}} \left\{ \|f - Y\|^2 + 4 \log(2)\sigma^2 k \log m \right\}$$

which satisfies the MSE bound

$$\frac{1}{n} E \left[\|\widehat{f}_n - f\|_2^2 \right] \leq \min_{f \in \mathcal{F}_{H,m}} \left\{ \frac{2}{n} \|f - f^*\|_2^2 + \frac{8\sigma^2 \log(2)k \log m}{n} \right\}$$

$$= \min_{\theta \in \Theta_m} \left\{ \frac{2}{n} \|H\theta - f^*\|_2^2 + \frac{8\sigma^2 \log(2)k \log m}{n} \right\}$$

where Θ_m denotes the quantized set of coefficient vectors. Note that the penalty term is the same for every $f \in \mathcal{F}_{H,m}$ and so the MPLE minimization is equivalent to minimizing over the choice of θ in the model $f = H\theta$, and since $\mathcal{F}_{H,m}$ almost perfectly covers $\mathcal{F}_H$, we may conlude that

$$\widehat{f}_n \approx H(H^T H)^{-1} HY$$

the standard least squares estimator of f given Y. Note that $H(H^T H)^{-1}H$ is the orthogonal projection matrix onto the subspace spanned by the columns of H. The error bound above shows that the least squares estimator's MSE is proportional to the sum of the squared bias of approximating f^* in terms of functions in the subspace and the variance which is proportional to the number of parameters k, as we know from standard multivariate statistics. Indeed, we for the simple least squares estimator one can derive a similar (actually tighter) MSE bound using standard techniques.

However, the MPLE framework can handle situations that cannot be easily characterized using standard techniques from multivariate statistics. Suppose that instead of just one subspace, we considered a sequence of subspaces of increasing dimension, say $H_1, H_2, \ldots$ where the subscript denotes the dimension (k in the analysis above). Each matrix defines a collection of linear models of the form $f = H_k \theta$, and for each we can define the least squares estimator

$$\widehat{f}_n^{(k)} \;\approx\; H_k (H_k^T H_k)^{-1} H_k Y$$

Now we can select the best overall estimator using our MPLE procedure. Note that if we define $\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_{H_k, m}$, then the MPLE is given by

$$\widehat{f}_n \;=\; \arg\min_{f \in \mathcal{F}} \left\{ \|f - Y\|^2 \;+\; 4\log(2)\sigma^2 k \log m \right\}$$

It is easy to see that $\widehat{f}_n \equiv \widehat{f}_n^{(\widehat{k})}$, where

$$\widehat{k} \;=\; \arg\min_{k \geq 1} \left\{ \frac{2}{n} \|H_k (H_k^T H_k)^{-1} H_k Y - Y\|_2^2 \;+\; \frac{8\sigma^2 \log(2) k \log m}{n} \right\}$$

In other words, we select the best dimension k by minimizing the combination of the residual sum-of-squared errors and the variance bound (proportional to k). The choice of k, and hence H_k, for our final estimator is an instance of model selection. The error bound shows that the MSE of the final estimator is essentially as good as we could achieve had we known a priori which value of k to use in the least squares proceedure. Note that the final estimator is a highly nonlinear function of the data, and therefore we could not arrive at a similar MSE bound using standard techniques.

3   Review: Denoising in Smooth Function Spaces I

We will put the above techniques into action in a denoising application. Let us first recall one of the basic denoising problems we considered earlier. Suppose we make noisy measurements of a smooth function:

$$Y_i = f^*(x_i) + W_i, \; i = \{1, \ldots, n\},$$

where

$$W_i \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$$

and

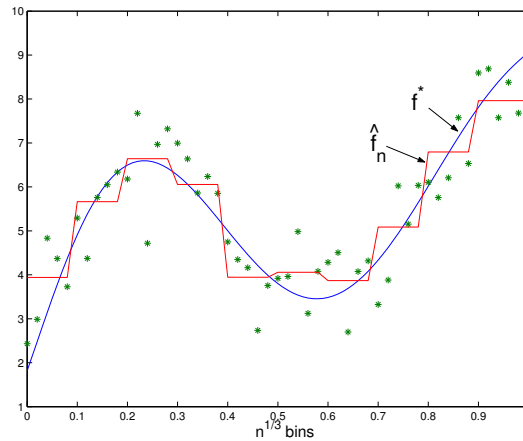$$x_i = \left(\frac{i}{n}\right).$$

The unknown function f^* is a map

$$f^* : [0, 1] \to \mathbf{R}$$

In Lecture 4, we consider this problem in the case where f^* was Lipschitz on $[0, 1]$. That is, f^* satisfied

$$|f^*(t) - f^*(s)| \leq L|t - s|, \quad \forall t, s \in [0, 1]$$

where $L > 0$ is a constant. In that case, we showed that by using a piecewise constant function on a partition of $n^{\frac{1}{3}}$ equal-size bins (Figure 3) we were able to obtain an estimator $\widehat{f}_n$ whose mean square error was

$$E\left[\|f^* - \widehat{f}_n\|^2\right] = O\left(n^{-\frac{2}{3}}\right)$$

Figure 1: Example of the piecewise constant approximation of f^*

In this lecture we will use the Maximum Complexity-Regularized Likelihood Estimation result we derived in Lecture 14 to extend our denoising scheme in several important ways.

To begin with let's consider a broader class of functions.

4   Hölder Spaces

For $0 < \alpha < 1$, define the space of functions

$$H^\alpha(C_\alpha) = \left\{ f : \sup_{x,h} \frac{|f(x+h) - f(x)|}{|h|^\alpha} \leq C_\alpha \right\}$$

for some constant $C_\alpha < \infty$ and where $f \in L_\infty$. H^α above contains functions that are bounded, but less smooth than Lipschitz functions. Indeed, the space of Lipschitz functions can be defined as H^1 ($\alpha = 1$)

$$H^1(C_1) = \left\{ f : \sup_{x,h} \frac{|f(x+h) - f(x)|}{|h|} \leq C_1 \right\}$$

for $C_1 < \infty$. Functions in H^1 are continuous, but those in $H^\alpha, \alpha < 1$, are not in general.

Let's also consider functions that are smoother than Lipschitz. If $\alpha = 1 + \beta$, where $0 < \beta < 1$, then define

$$H^\alpha(C_\alpha) = \left\{ f \in H^1(C_\alpha) : \frac{\partial f}{\partial x} \in H^\beta(C_\alpha) \right\}$$

In other words, $H^\alpha, 1 < \alpha < 2$, contains Lipschitz functions that are also differentiable **and** their derivatives are Hölder smooth with smoothness $\beta = \alpha - 1$.

And finally, let

$$H^2(C_2) = \left\{ f : \frac{\partial f}{\partial x} \in H^1(C_2) \right\}$$

contain functions that have continuous derivatives, but that are not necessarily twice-differentiable.

If $f \in H^\alpha(C_\alpha)$, $0 < \alpha \leq 2$, then we say that f is Hölder$-\alpha$ smooth with Hölder constant C_α. The notion of Hölder smoothness can also be extended to $\alpha > 2$ in a straightforward way.

Note: If $\alpha_1 < \alpha_2$ then

$$f \in H^{\alpha_2} \Rightarrow f \in H^{\alpha_1}$$

Summarizing, we can describe Hölder spaces as follows. If $f^* \in H^\alpha(C_\alpha)$ for some $0 < \alpha \leq 2$ and $C_\alpha < \infty$, then

(i) $0 < \alpha \leq 1$ $\hspace{3cm}$ $|f^*(t) - f^*(s)| \leq C_\alpha |t - s|^\alpha$

(ii) $1 < \alpha \leq 2$ $\hspace{2.7cm}$ $\left| \frac{\partial f^*}{\partial x}(t) - \frac{\partial f^*}{\partial x}(s) \right| \leq C_\alpha |t - s|^{\alpha - 1}$

Note that since Hölder smoothness essentially measures how differentiable functions are, the Taylor polynomial is the natural way to approximate Hölder smooth functions. We will focus on Hölder smooth function classes with $0 < \alpha \leq 2$. Thus, we will work with piecewise linear approximations, the Taylor polynomial of degree 1. If we were to consider smoother functions, $\alpha > 2$ we would need consider higher degree Taylor polynomial approximation functions, i.e. quadratic, cubic, etc.

5 Denoising Example for Signal-plus-Gaussian Noise Observation Model

Now let's assume $f^* \in H^\alpha(C_\alpha)$ for some ***unknown*** α $(0 < \alpha \leq 2)$; i.e. we don't know how smooth f^* is. We will use our observations

$$Y_i = f^*(x_i) + W_i, \ i = \{1, \ldots, n\},$$

to construct an estimator $\widehat{f}_n$. Intuitively, the smoother f^* is, the better we should be able to estimate it. Can we take advantage of extra smoothness in f^* if we don't know how smooth it is? The smoother f^* is, the more averaging we can perform to reduce noise. In other words for smoother f^* we should average over larger bins. Also, we will need to exploit the extra smoothness in our approximation of f^*. To that end, we will consider candidate functions that are piecewise ***linear*** functions on uniform partitions of $[0, 1]$. Let

$$\mathcal{F}_k = \left\{ \ |f| \leq C : \quad \begin{array}{l} f \text{ is piecewise linear on } \left[0, \frac{1}{k}\right), \left[\frac{1}{k}, \frac{2}{k}\right), \ldots \left[\frac{k-1}{k}, 1\right) \text{ and the} \\ \text{coefficients of each line segment are quantized to } \frac{1}{2} \log n \text{ bits.} \end{array} \right\}$$

The start and end points of each line segment are each one of $\sqrt{n}$ discrete values, as indicated in Figure 2. Since each line may start at any of the $\sqrt{n}$ levels and terminate at any of the $\sqrt{n}$ levels, there are a total of n possible lines for each segment.
Given that there are k intervals we have

$$|\mathcal{F}_k| = n^k \Rightarrow \log |\mathcal{F}_k| = k \log n$$

Therefore we can use $k \log n$ bits to describe a function $f \in \mathcal{F}_k$. Also observe that each function $f \in \mathcal{F}_k$ has
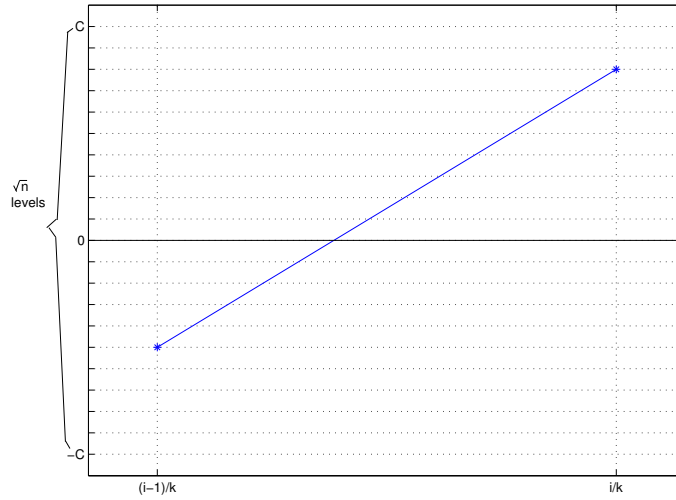
Figure 2: Example on the quantization of f on interval $\left[\frac{i-1}{k}, \frac{i}{k}\right)$

the form $f = H_k \theta$, where

$$
H_k =
\begin{bmatrix}
1 & 1 & 0 & \dots & \dots & \dots & 0 \\
1 & 2 & 0 & \dots & \dots & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \ell & 0 & \dots & \dots & \dots & 0 \\
0 & 0 & 1 & 1 & 0 & \dots & 0 \\
0 & 0 & 1 & 2 & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 1 & \ell & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \dots & \dots & \dots & 0 & 1 & 1 \\
0 & \dots & \dots & \dots & 0 & 1 & 2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \dots & \dots & \dots & 0 & 1 & \ell
\end{bmatrix}
$$

Note that each of the $2k$ columns corresponds to a constant $[1, 1, \dots, 1]$ or linear $[1, 2, \dots, \ell]$ basis function of length $\ell = n/k$ for one of the k subintervals.

Let

$$
\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k.
$$

Construct a prefix code for every $f \in \mathcal{F}$ by

    (i)   Use $\underbrace{000 \dots 1}_{k \ bits}$ to encode the smallest k such that $f \in \mathcal{F}_k$

    (ii)   Use $k \log n$ bits to encode which element of $\mathcal{F}_k$ we are considering.

Thus, if $f \in \mathcal{F}_k$, then the prefix code associated with f has codeword length

$$
c(f) = k + k \log n = k(1 + \log n)
$$

which satisfies the Kraft Inequality

$$
\sum_{f \in \mathcal{F}} 2^{-c(f)} \leq 1.
$$

Now we will apply our complexity regularization result to select a function $\widehat{f}_n$ from $\mathcal{F}$ and bound its risk. We are assuming Gaussian errors, so

$$-\log p_f(Y_i) = \frac{\left(Y_i - f\left(\frac{i}{n}\right)\right)^2}{2\sigma^2} + constant.$$

We can ignore the constant term and so our empirical selection is

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\left(Y_i - f\left(\frac{i}{n}\right)\right)^2}{2\sigma^2} + \frac{2c(f)\log 2}{n} \right\}$$

We can compute $\widehat{f}_n$ according to:

For $k = 1, \ldots, n$

$$\widehat{f}_n^{(k)} = \arg\min_{f \in \mathcal{F}_k} \hat{R}_n(f) = \arg\min_{f \in \mathcal{F}_k} \frac{1}{n} \sum_{i=1}^{n} \frac{\left(Y_i - f\left(\frac{i}{n}\right)\right)^2}{2\sigma^2}$$

then select

$$\hat{k} = \arg\min_{k=1,\ldots,n} \left\{ \hat{R}_n\left(\widehat{f}_n^{(k)}\right) + \frac{2k(1 + \log n)\log 2}{n} \right\}$$

and finally

$$\widehat{f}_n = \widehat{f}_n^{(\hat{k})}.$$

Because the KL divergence and $-2\log$ *affinity* simply reduce to squared error in the Gaussian case, the risk bound in Theorem 1, Lecture 14, produces a relatively simple bound on the mean square error of $\widehat{f}_n$

$$\frac{1}{n} \sum_{i=1}^{n} E\left[\left(\widehat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2\right] \leq \min_{f \in \mathcal{F}} \left\{ \frac{2}{n} \sum_{i=1}^{n} \left(f\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2 + \frac{8\sigma^2 c(f)\log 2}{n} \right\}$$

The first term on the lefthand side above is the error incurred by approximating f^* by an element of $\mathcal{F}$. The second term is an upper bound on the estimation error involved with the model selection process.

Let's focus on the approximation error. First, suppose $f^* \in H^\alpha\left(C_\alpha\right)$ for $1 < \alpha \leq 2$. Let f_k^* be the "best" piecewise linear approximation to f^*, with k pieces on intervals $\left[0, \frac{1}{k}\right), \left[\frac{1}{k}, \frac{2}{k}\right), \ldots \left[\frac{k-1}{k}, 1\right)$. Consider the difference between f^* and f_k^* on one such interval, say $\left[\frac{i-1}{k}, \frac{i}{k}\right)$. By applying Taylor's theorem with remainder we have

$$f^*(t) = f^*\left(\frac{i}{k}\right) + \frac{\partial f^*}{\partial x}(t')\left(t - \frac{i}{k}\right)$$

for $t \in \left[\frac{i-1}{k}, \frac{i}{k}\right)$ and some $t' \in \left[t, \frac{i}{k}\right]$. Define

$$f_k^*(t) \equiv f^*\left(\frac{i}{k}\right) + \frac{\partial f^*}{\partial x}\left(\frac{i}{k}\right)\left(t - \frac{i}{k}\right).$$

Note that $f_k^*(t)$ is not necessarily the best piecewise linear approximation to f^*, but it is good enough for our purposes. Then using the fact that $f^* \in H^\alpha\left(C_\alpha\right)$, for $t \in [i-1/k, i/k)$ we have

$$
\begin{aligned}
|f^*(t) - f_k^*(t)| &= \left| \frac{\partial f^*}{\partial x}(t')\left(t - \frac{i}{k}\right) - \frac{\partial f^*}{\partial x}\left(\frac{i}{k}\right)\left(t - \frac{i}{k}\right) \right| \\
&\leq \frac{1}{k}\left| \frac{\partial f^*}{\partial x}(t') - \frac{\partial f^*}{\partial x}\left(\frac{i}{k}\right) \right| \\
&\leq \frac{1}{k}C_\alpha\left| t' - \frac{i}{k} \right|^{\alpha-1} \\
&\leq \frac{1}{k}C_\alpha\left(\frac{1}{k}\right)^{\alpha-1} = C_\alpha k^{-\alpha}.
\end{aligned}
$$

So, for all $t \in [0, 1]$

$$|f^*(t) - f_k^*(t)| \leq C_\alpha k^{-\alpha}.$$

Now let f_k be the element of $\mathcal{F}_k$ closest to f_k^* (f_k is the quantized version of f_k^*)

$$
\begin{aligned}
|f^*(t) - f_k(t)| &= |f^*(t) - f_k^*(t) + f_k^*(t) - f_k(t)| \\
&\leq |f^*(t) - f_k^*(t)| + |f_k^*(t) - f_k(t)| \\
&\leq C_\alpha k^{-\alpha} + \frac{1}{\sqrt{n}}
\end{aligned}
$$

since we used $\frac{1}{2} \log n$ bits to quantize the endpoints of each line segment. Consequently,

$$
\begin{aligned}
|f^*(t) - f_k^*(t)|^2 &\leq |f^*(t) - f_k^*(t)|^2 + 2|f^*(t) - f_k^*(t)| |f_k^*(t) - f_k(t)| + |f_k^*(t) - f_k(t)|^2 \\
&\leq C_\alpha^2 k^{-2\alpha} + 2C_\alpha \frac{k^{-\alpha}}{\sqrt{n}} + \frac{1}{n}.
\end{aligned}
$$

Thus it follows that

$$\min_{f \in \mathcal{F}_k} \left\{ \frac{2}{n} \sum_{i=1}^{n} (f(i/n) - f^*(i/n))^2 + \frac{8\sigma^2 c(f) \log 2}{n} \right\} \leq 2C_\alpha^2 k^{-2\alpha} + \frac{4C_\alpha k^{-\alpha}}{\sqrt{n}} + \frac{2}{n} + \frac{8\sigma^2 k(\log n + 1)\log 2}{n}.$$

The first and last terms dominate the above expression. Therefore, the upper bound is minimized when $k^{-2\alpha}$ and $\frac{k}{n}$ are balanced. This is accomplished by choosing $k = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$. Then it follows that

$$\min_{f \in \mathcal{F}_k} \left\{ \frac{2}{n} \sum_{i=1}^{n} \left(f\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right) \right)^2 + \frac{8\sigma^2 c(f) \log 2}{n} \right\} = O\left(n^{-\frac{2\alpha}{2\alpha+1}} \log n \right).$$

If $\alpha = 2$ then we have

$$\frac{1}{n} \sum_{i=1}^{n} E\left[\left(\widehat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right) \right)^2 \right] = O\left(n^{-\frac{4}{5}} \log n \right)$$

If $f^* \in H^\alpha(C_\alpha)$ for $0 < \alpha \leq 1$, let f_k^* be the following piecewise constant approximation to f^*. Note that constant functions are simply special cases of linear functions, and thus they are contained in $\mathcal{F}$. Let

$$f_k^*(t) \equiv f^*\left(\frac{i}{n}\right) \quad \text{on interval} \quad \left[\frac{i-1}{k}, \frac{i}{k}\right).$$

Then

$$
\begin{aligned}
|f^*(t) - f_k^*(t)| &= \left| f^*(t) - f^*\left(\frac{i}{n}\right) \right| \\
&\leq C_\alpha \left| t - \frac{i}{n} \right|^\alpha \\
&\leq C_\alpha k^{-\alpha}.
\end{aligned}
$$

Repeating the same reasoning as in the $1 < \alpha \leq 2$ case, we arrive at

$$\frac{1}{n} \sum_{i=1}^{n} E\left[\left(\widehat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right) \right)^2 \right] = O\left(n^{-\frac{2\alpha}{2\alpha+1}} \log n \right)$$

for $0 < \alpha \leq 1$. In particular, for $\alpha = 1$ we get

$$\frac{1}{n} \sum_{i=1}^{n} E\left[\left(\widehat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right) \right)^2 \right] = O\left(n^{-\frac{2}{3}} \log n \right)$$

within a logarithmic factor of the rate we had before (in Lecture 4) for that case!

6 Summary

1. $\widehat{f}_n$ can be computed by finding least-square line fits to the data on partitions of the form $\left[\frac{i-1}{k}, \frac{1}{k}\right)$ for $i = 1, \ldots, n$, and then selecting the best fit by choosing $\hat{k}$ that minimizes the complexity regularization criterion.

2. If $f^* \in H^\alpha\left(C_\alpha\right)$ for some $0 < \alpha \leq 2$, then

$$MSE\left(\widehat{f}_n\right) = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(\widehat{f}_n\left(\frac{i}{n}\right) - f^*\left(\frac{i}{n}\right)\right)^2\right] = O\left(n^{-\frac{2\alpha}{2\alpha+1}} \log n\right).$$

3. $\widehat{f}_n$ **automatically** picks the optimal number of bins. Essentially $\widehat{f}_n$ adapts to the unknown smoothness of f^* and produces a rate which is near minimax optimal ! ($n^{-\frac{2\alpha}{2\alpha+1}}$ is the best possible).

4. The larger α is the faster the convergence and the better the denoising !