1 Review: Denoising in Smooth Function Spaces I - Method of Sieves

Suppose we make noisy measurements of a smooth function:

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

where

\[ W_i \sim_{i.i.d.} N(0, \sigma^2) \]

and

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

The unknown function \( f^* \) is a map

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

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 \( n^{1/3} \) equal-size bins (Figure 1) we were able to obtain an estimator \( \hat{f}_n \) whose mean square error was

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

![Figure 1: Example of the piecewise constant approximation of \( f^* \)](image-url)
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.

## 2 Hölder Spaces

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

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

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

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

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 $C_\alpha$. 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) $0 < \alpha \leq 1$

$$|f^*(t) - f^*(s)| \leq C_\alpha |t - s|^\alpha$$

(ii) $1 < \alpha \leq 2$

$$\left| \frac{\partial f^*}{\partial x}(t) - \frac{\partial f^*}{\partial x}(s) \right| \leq C_\alpha |t - s|^{\alpha - 1}$$

Note that in general there is a natural relationship between the Hölder space containing the function and the approximation class used to estimate the function. Here we will consider functions which are Hölder–$\alpha$ smooth where $0 < \alpha \leq 2$ and work with piecewise linear approximations. If we were to consider smoother functions, $\alpha > 2$ we would need consider higher order approximation functions, i.e. quadratic, cubic, etc.
3 Denoising Example for Signal-plus-Gaussian Noise Observation Model

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

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

to construct an estimator \( \hat{f}_n \). 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 \textit{linear} functions on uniform partitions of \([0, 1]\). Let

\[
F_k = \{ |f| \leq C : \text{f is piecewise linear on \([0, \frac{1}{k}\), \([\frac{1}{k}, \frac{2}{k}\), \ldots \([\frac{k-1}{k}, 1]\) and the coefficients of each line segment are quantized to \(\frac{1}{2} \log n\) bits.} \}
\]

Figure 2: Example on the quantization of \( f \) on interval \([\frac{i-1}{k}, \frac{i}{k})\)

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

\[
|F_k| = n^k \Rightarrow \log |F_k| = k \log n
\]

Therefore we can use \(k \log n\) bits to describe a function \( f \in F_k \).

Let

\[
\mathcal{F} = \bigcup_{k \geq 1} F_k.
\]

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

(i) Use \(\underbrace{000\ldots1}_{k \text{ bits}}\) to encode the smallest \( k \) such that \( f \in F_k \)

(ii) Use \(k \log n\) bits to encode which element of \( F_k \) we are considering.
Thus, if $f \in F_k$, 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

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

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

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

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

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

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

For $k = 1, \ldots, n$

$$\hat{f}_n^{(k)} = \arg \min_{f \in F_k} \hat{R}_n(f) = \arg \min_{f \in F_k} \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - f(\frac{i}{n}))^2}{2\sigma^2}$$

then select

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

and finally

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

Because the KL divergence and $-2 \log \text{affinity}$ simply reduce to squared error in the Gaussian case (Lecture 14), we arrive at a relatively simple bound on the mean square error of $\hat{f}_n$

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

The first term in the brackets above is related to the error incurred by approximating $f^*$ by an element of $F$. The second term is related to the estimation error involved with the model selection process.

Let’s focus on the approximation error. First, suppose $f^* \in H^\alpha(C_\alpha)$ for $1 < \alpha \leq 2$. Let $f_k^*$ be the “best” piecewise linear approximation to $f^*$, with $k$ pieces on intervals $[0, \frac{1}{k})$, $[\frac{1}{k}, \frac{2}{k})$, $\ldots$, $[\frac{k-1}{k}, 1)$. Consider the difference between $f^*$ and $f_k^*$ on one such interval, say $[\frac{i-1}{k}, \frac{i}{k})$. 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 $t \in [\frac{i-1}{k}, \frac{i}{k})$ and some $t' \in [t, \frac{i}{k}]$. 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 $f^*$, just good enough for our purposes. Then using the fact that $f^* \in H^\alpha (C_\alpha)$, for $t \in [i-1/k,i/k)$ we have

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

$$\leq \frac{1}{k} \left| \frac{\partial f^*}{\partial x} (t') - \frac{\partial f^*}{\partial x} \left( \frac{i}{k} \right) \right|$$

$$\leq \frac{1}{k} C_\alpha \left| t' - \frac{i}{k} \right|^{\alpha - 1}$$

$$\leq \frac{1}{k} C_\alpha \left( \frac{1}{k} \right)^{\alpha - 1} = C_\alpha k^{-\alpha}.$$

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 $F_k$ closest to $f_k^*$ ($f_k$ is the quantized version of $f_k^*$)

$$|f^*(t) - f_k(t)| = |f^*(t) - f_k^*(t) + f_k^*(t) - f_k(t)|$$

$$\leq |f^*(t) - f_k^*(t)| + |f_k^*(t) - f_k(t)|$$

$$\leq C_\alpha k^{-\alpha} + \frac{1}{\sqrt{n}}$$

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

$$|f^*(t) - f_k^*(t)|^2 \leq |f^*(t) - f_k^*(t)|^2 + 2 |f^*(t) - f_k^*(t)| |f_k^*(t) - f_k(t)| + |f_k^*(t) - f_k(t)|^2$$

$$\leq C_\alpha^2 k^{-2\alpha} + 2C_\alpha k^{-\alpha} \frac{1}{\sqrt{n}} + \frac{1}{n}.$$

Thus it follows that

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

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

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

If $\alpha = 2$ then we have

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

If $f^* \in H^\alpha (C_\alpha)$ for $0 < \alpha \leq 1$, let $f_k^*$ be the following piecewise constant approximation to $f^*$. Let

$$f_k^*(t) \equiv f^* \left( \frac{i}{n} \right) \text{ on interval } \left[ \frac{i-1}{k} , \frac{i}{k} \right).$$
Then
\[
|f^*(t) - f_k(t)| = \left| f^*(t) - f^* \left( \frac{i}{n} \right) \right| \\
\leq C_{\alpha} \left| t - \frac{i}{n} \right|^\alpha \\
\leq C_{\alpha} k^{-\alpha}.
\]
Repeating the same reasoning as in the $1 < \alpha \leq 2$ case, we arrive at
\[
\frac{1}{n} \sum_{i=1}^{n} E \left[ \left( \hat{f}_n \left( \frac{i}{n} \right) - f^* \left( \frac{i}{n} \right) \right)^2 \right] = O \left( n^{-\frac{2\alpha}{2\alpha+1}} \log n \right)
\]
for $0 < \alpha \leq 1$. In particular, for $\alpha = 1$ we get
\[
\frac{1}{n} \sum_{i=1}^{n} E \left[ \left( \hat{f}_n \left( \frac{i}{n} \right) - f^* \left( \frac{i}{n} \right) \right)^2 \right] = O \left( n^{-\frac{3}{2}} \log n \right)
\]
within a logarithmic factor of the rate we had before (in Lecture 4) for that case!

4 Summary

1. $\hat{f}_n$ can be computed by finding least-square line fits to the data on partitions of the form $[0, \frac{1}{k}), [\frac{1}{k}, \frac{2}{k}) \ldots \left[ \frac{k-1}{k}, 1 \right)$ for $k = 1, \ldots, n$, and then selecting the best fit by the $k$ that gives the minimum of the complexity regularization criterion.

2. If $f^* \in \mathcal{H}^\alpha (C_{\alpha})$ for some $0 < \alpha \leq 2$, then
\[
MSE \left( \hat{f}_n \right) = \frac{1}{n} \sum_{i=1}^{n} E \left[ \left( \hat{f}_n \left( \frac{i}{n} \right) - f^* \left( \frac{i}{n} \right) \right)^2 \right] = O \left( n^{-\frac{2\alpha}{2\alpha+1}} \log n \right).
\]

3. $\hat{f}_n$ automatically picks the optimal number of bins. Essentially $\hat{f}_n$ (indirectly) estimates the 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 $\alpha$ is the faster the convergence and the better the denoising!