ECE 901
Lecture 15: Denoising Smooth Functions with Unknown Smoothness

R. Nowak

5/17/2009

1 Review: Denoising in Smooth Function Spaces I

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

$$x_i = \frac{i}{n}.$$

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 1) we were able to obtain an estimator $\widehat{f}_n$ whose mean square error was

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

Lipschitz functions are interesting, but can be very rough (these can have many kinks). In many situations the functions can be much smoother. This is how you would model the temperature inside a museum room for example. Often we don't know how smooth the function might be, so an interesting question is if we can adapt to the unknown smoothness. 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 $0 < \alpha \leq 1$, define the space of functions

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

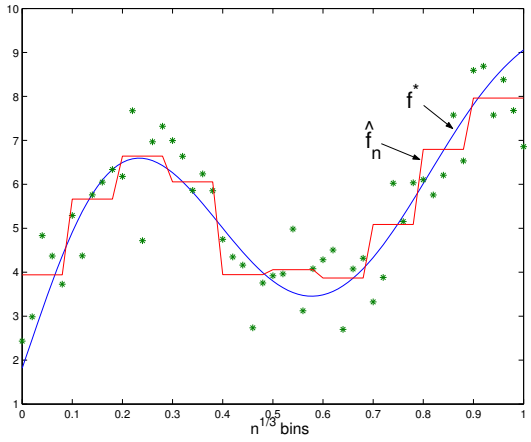

Figure 1: Example of the piecewise constant approximation of f^*

for some constant $C < \infty$ 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 ($\alpha = 1$)

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

for $C < \infty$. Functions in H^1 are uniformly continuous, but functions in $H^\alpha, \alpha < 1$, are generally not uniformly continuous (although are still continuous). Therefore a larger α corresponds to smoother functions.

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

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

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

If $f \in H^\alpha(C)$, $0 < \alpha \leq 2$, then we say that f is Hölder$-\alpha$ smooth with Hölder constant C. The notion of Hölder smoothness can also be extended to $\alpha > 2$ in a straightforward way. There are other equivalent ways of defining Hölder smooth functions, in particular characterizing their local approximation by polynomials. In particular a function f is Hölder-α smooth if it has $\lfloor \alpha \rfloor$ derivatives and

$$|f(x) - T_y^{\lfloor \alpha \rfloor}(x)| \leq C|x - y|^\alpha \quad \forall x, y \ ,$$

where $\lfloor \alpha \rfloor$ is the largest integer such that $\lfloor \alpha \rfloor < \alpha$, and $T_y^{\lfloor \alpha \rfloor}$ is the Taylor polynomial of degree $\lfloor \alpha \rfloor$ around the point y. In words, a Hölder-α smooth function is locally well approximated by a polynomial of degree $\lfloor \alpha \rfloor$. In this lecture we will work with the previous definition (this will also give you an indication why the two definitions are equivalent).

Note: If a function is Hölder-α_2 smooth and $\alpha_1 < \alpha_2$ then the function is also Hölder-α_1 smooth.

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

(i) $0 < \alpha \leq 1$ $\qquad |f^*(t) - f^*(s)| \leq C|t - s|^\alpha$

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

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...

3 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

$$\tilde{\mathcal{F}}_k = \left\{\ |f| \leq C : \quad f \textit{ is piecewise linear on } \left[0, \tfrac{1}{k}\right), \left[\tfrac{1}{k}, \tfrac{2}{k}\right), \ldots \left[\tfrac{k-1}{k}, 1\right)\ \right\}$$

To be able to apply our maximum penalized likelihood estimator (MPLE) error analysis we need a class of models that is countable. Unfortunately the class $\tilde{\mathcal{F}}_k$ above is not countable. An easy but effective way of addressing this problem is to discretize the possibilities for the various end-points of each segment. We will consider $\sqrt{n}$ possibilities for each end-point, as illustrated in Figure 2. Denote this class of models by $\mathcal{F}_k$.

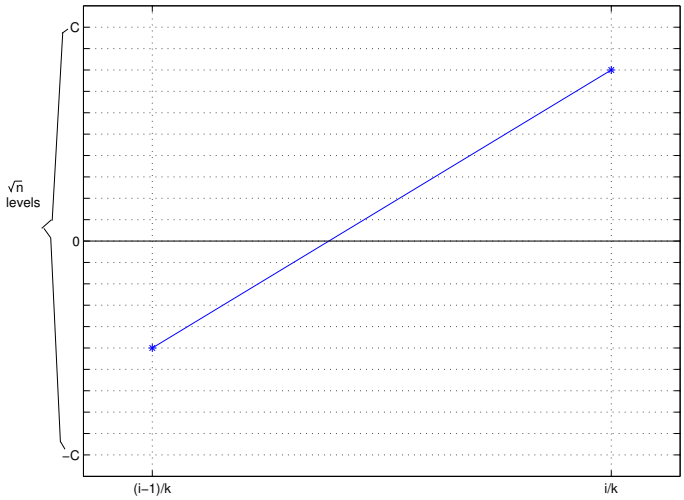

Figure 2: Example on the quantization of f on interval $\left[\frac{i-1}{k}, \frac{i}{k}\right)$

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

$$|\mathcal{F}_k| = n^k \Rightarrow \log_2 |\mathcal{F}_k| = k \log_2 n$$

Therefore we can use $k \log_2 n$ bits to describe a function $f \in \mathcal{F}_k$.

Since we don't know the smoothness of the underlying function f^* we will need to search for a good model with k bins, where k depends on the unknown smoothness. In essence, we want to consider all possible models with an arbitrary number of bins.

Let

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

We have encountered this kind of construction before, and seen that we can devise a simple prefix code to encode all the elements in $\mathcal{F}$:

(i) Use $\underbrace{000\ldots1}_{k(f)\,bits}$ to encode the smallest k such that $f \in \mathcal{F}_k$

(ii) Use $k(f)\log_2 n$ 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

$$c(f) = k(f) + k(f)\log_2 n = k(f)(1 + \log n)$$

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 $f^* \in H^\alpha(C)$ for $1 < \alpha \leq 2$. Let f_k^* be the "good" 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 $t' \in \left[t, \frac{i}{k}\right]$. This suggests the definition

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

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 $f^* \in H^\alpha(C)$, for $t \in [i-1/k, i/k)$ we have

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

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

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

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

$$\begin{aligned}
|f^*(t) - f_k(t)| &= |f^*(t) - f_k^*(t) + f_k^*(t) - f_k(t)| \\
&\leq |f^*(t) - f_k^*(t)| + |f_k^*(t) - f_k(t)| \\
&\leq Ck^{-\alpha} + \frac{2C}{\sqrt{n}}
\end{aligned}$$

since each segment endpoint is quantized with a step $2C/\sqrt{n}$. Consequently,

$$|f^*(t) - f_k^*(t)|^2 \leq C^2k^{-2\alpha} + 4C^2\frac{k^{-\alpha}}{\sqrt{n}} + \frac{4C}{n}\ .$$

Thus it follows that

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

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 $k = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$.

Finally using this observation in the oracle bound derived before we get that

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

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 $f^* \in H^\alpha(C)$ 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

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

Then

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

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![1]

4 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 $f^* \in H^\alpha(C)$ 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). The extra log factor is the price we pay for the adaptivity to the unknown smoothness.

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

[1] Note that in Lecture 4 we measured the error as $\|\hat{f}_n - f^*\|_{L2}^2 = \int_0^1 (\hat{f}_n(t) - f^*(t))^2 dt$, which is different than the way we are measuring the error here. It is possible to relate the two, but this involves more work, specially for the case when $\alpha > 1$.