

Lecture 4: Denoising in Smooth Function Spaces

An example of the use of Sieves for complexity regularization in denoising

Consider the following setting. Let

$$Y = f^*(X) + W,$$

where X is a random variable (r.v.) on $\mathcal{X} = [0, 1]$, W is a r.v. on $\mathcal{Y} = \mathbf{R}$, independent of X and satisfying

$$E[W] = 0 \quad \text{and} \quad E[W^2] = \sigma^2 < \infty.$$

Finally let $f^* : [0, 1] \rightarrow \mathbf{R}$ be a function satisfying

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

where $L > 0$ is a constant. A function satisfying condition (1) is said to be Lipschitz on $[0, 1]$. Notice that such a function must be continuous, but it is not necessarily differentiable. An example of such a function is depicted in Figure (a).

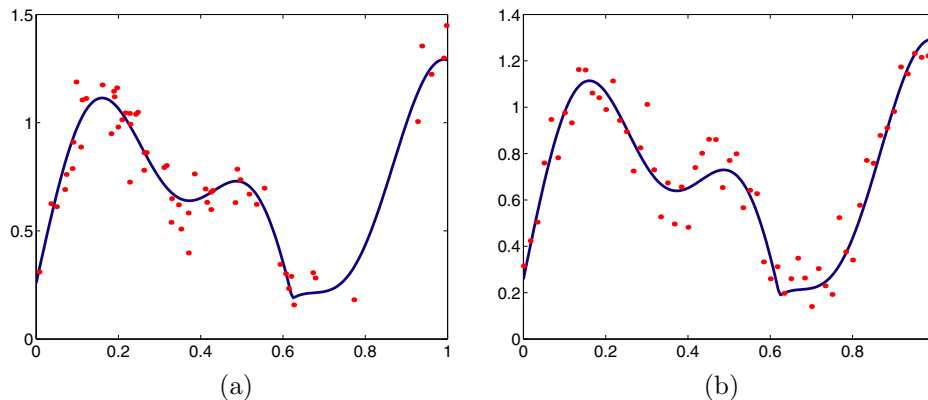


Figure 1: Example of a Lipschitz function, and our observations setting. (a) random sampling of f^* , the points correspond to (X_i, Y_i) , $i = 1, \dots, n$; (b) deterministic sampling of f^* , the points correspond to $(i/n, Y_i)$, $i = 1, \dots, n$.

Note that

$$\begin{aligned} E[Y|X = x] &= E[f^*(X) + W|X = x] \\ &= E[f^*(x) + W|X = x] \\ &= f^*(x) + E[W] = f^*(x). \end{aligned}$$

Consider our usual setup: Estimate f^* using n training examples

$$\begin{aligned} \{X_i, Y_i\}_{i=1}^n &\stackrel{i.i.d.}{\sim} P_{XY}, \\ Y_i &= f^*(X_i) + W_i, \quad i = \{1, \dots, n\}, \end{aligned}$$

where $\stackrel{i.i.d.}{\sim}$ means *independently and identically distributed*. Figure (a) illustrates this setup.

In many applications we can sample $\mathcal{X} = [0, 1]$ as we like, and not necessarily at random. For example we can take n samples uniformly on $[0, 1]$

$$\begin{aligned} x_i &= \frac{i}{n}, \quad i = 1, \dots, n, \\ Y_i &= f(x_i) + W_i \\ &= f\left(\frac{i}{n}\right) + W_i. \end{aligned}$$

We will proceed with this setup (as in Figure (b)) in the rest of the lecture.

Our goal is to find \hat{f}_n such that $E[\|f^* - \hat{f}_n\|^2] \rightarrow 0$, as $n \rightarrow \infty$ (here $\|\cdot\|$ is the usual L_2 -norm; i.e., $\|f^* - \hat{f}_n\|^2 = \int_0^1 |f^*(t) - \hat{f}_n(t)|^2 dt$).

Let

$$\mathcal{F} = \{f : f \text{ is Lipschitz with constant } L\}.$$

The **Risk** is defined as

$$R(f) = \|f^* - f\|^2 = \int_0^1 |f^*(t) - f(t)|^2 dt.$$

The **Expected Risk** (recall that our estimator \hat{f}_n is based on $\{x_i, Y_i\}$ and hence is a r.v.) is defined as

$$E[R(\hat{f}_n)] = E[\|f^* - \hat{f}_n\|^2].$$

Finally the **Empirical Risk** is defined as

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \left(f\left(\frac{i}{n}\right) - Y_i \right)^2.$$

Let $0 < m_1 \leq m_2 \leq m_3 \leq \dots$ be a sequence of integers satisfying $m_n \rightarrow \infty$ as $n \rightarrow \infty$, and $k_n m_n = n$ for some integer $k_n > 0$. That is, for each value of n there is an associated integer value m_n . Define the **Sieve** $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots$,

$$\mathcal{F}_n = \left\{ f : f(t) = \sum_{j=1}^{m_n} c_j \mathbf{1}_{\left\{ \frac{j-1}{m_n} \leq t < \frac{j}{m_n} \right\}}, \quad c_j \in \mathbf{R} \right\}.$$

\mathcal{F}_n is the space of functions that are constant on intervals

$$I_{j, m_n} \equiv \left[\frac{j-1}{m_n}, \frac{j}{m_n} \right), \quad j = 1, \dots, m_n.$$

From here on we will use m and k instead of m_n and k_n (dropping the subscript n) for notational ease. Define

$$f_n(t) = \sum_{j=1}^m c_j^* \mathbf{1}_{\{t \in I_{j, m}\}}, \quad \text{where} \quad c_j^* = \frac{1}{k} \sum_{i: \frac{i}{n} \in I_{j, m}} f^*\left(\frac{i}{n}\right)$$

Note that $f_n \in \mathcal{F}_n$.

Exercise 1 Upper bound $\|f^* - f_n\|^2$.

$$\begin{aligned}
\|f^* - f\|^2 &= \int_0^1 |f^*(t) - f_n(t)|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} |f^*(t) - f_n(t)|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} |f^*(t) - c_j^*|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} \left| f^*(t) - \frac{1}{k} \sum_{i: \frac{i}{n} \in I_{j,m}} f^*\left(\frac{i}{n}\right) \right|^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} \left(\frac{1}{k} \left| \sum_{i: \frac{i}{n} \in I_{j,m}} \left(f^*(t) - f^*\left(\frac{i}{n}\right) \right) \right| \right)^2 dt \\
&\leq \sum_{j=1}^m \int_{I_{j,m}} \left(\frac{1}{k} \sum_{i: \frac{i}{n} \in I_{j,m}} \left| f^*(t) - f^*\left(\frac{i}{n}\right) \right| \right)^2 dt \\
&\leq \sum_{j=1}^m \int_{I_{j,m}} \left(\frac{1}{k} \sum_{i: \frac{i}{n} \in I_{j,m}} \frac{L}{m} \right)^2 dt \\
&= \sum_{j=1}^m \int_{I_{j,m}} \left(\frac{L}{m} \right)^2 dt \\
&= \sum_{j=1}^m \frac{1}{m} \left(\frac{L}{m} \right)^2 = \left(\frac{L}{m} \right)^2.
\end{aligned}$$

The above implies that $\|f^* - f_n\|^2 \rightarrow 0$ as $n \rightarrow \infty$, since $m = m_n \rightarrow \infty$ as $n \rightarrow \infty$. In words, with n sufficiently large we can approximate f^* to arbitrary accuracy using models in \mathcal{F}_n (even if the functions we are using to approximate f^* are not Lipschitz!).

For any $f \in \mathcal{F}_n$, $f = \sum_{j=1}^m c_j \mathbf{1}_{\{t \in I_{j,m}\}}$, we have

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{j=1}^m \left(\sum_{i: \frac{i}{n} \in I_{j,m}} (c_j - Y_i)^2 \right).$$

Let $\widehat{f}_n = \arg \min_{f \in \mathcal{F}_n} \widehat{R}_n(f)$. Then

$$\widehat{f}_n(t) = \sum_{j=1}^m \widehat{c}_j \mathbf{1}_{\{t \in I_{j,m}\}}, \quad \text{where } \widehat{c}_j = \frac{1}{k} \sum_{i: \frac{i}{n} \in I_{j,m}} Y_i \quad (2)$$

Exercise 2 Show (2).

Note that $E[\widehat{c}_j] = c_j^*$ and therefore $E[\widehat{f}_n(t)] = f_n(t)$. Lets analyze now the expected risk of \widehat{f}_n :

$$\begin{aligned}
 E[\|f^* - \widehat{f}_n\|^2] &= E[\|f^* - f_n + f_n - \widehat{f}_n\|^2] \\
 &= \|f^* - f_n\|^2 + E[\|f_n - \widehat{f}_n\|^2] + 2E[\langle f^* - f_n, f_n - \widehat{f}_n \rangle] \\
 &= \|f^* - f_n\|^2 + E[\|f_n - \widehat{f}_n\|^2] + 2\langle f^* - f_n, E[f_n - \widehat{f}_n] \rangle \\
 &= \|f^* - f_n\|^2 + E[\|f_n - \widehat{f}_n\|^2],
 \end{aligned} \tag{3}$$

where the final step follows from the fact that $E[\widehat{f}_n(t)] = f_n(t)$. A couple of important remarks pertaining the right-hand-side of equation (3): The first term, $\|f^* - f_n\|^2$, corresponds to the approximation error, and indicates how well can we approximate the function f^* with a function from \mathcal{F}_n . Clearly, the larger the class \mathcal{F}_n is, the smallest we can make this term. This term is precisely the squared bias of the estimator \widehat{f}_n . The second term, $E[\|f_n - \widehat{f}_n\|^2]$, is the estimation error, the variance of our estimator. We will see that the estimation error is small if the class of possible estimators \mathcal{F}_n is also small.

The behavior of the first term in (3) was already studied. Consider the other term:

$$\begin{aligned}
 E[\|f_n - \widehat{f}_n\|^2] &= E\left[\int_0^1 |f_n(t) - \widehat{f}_n(t)|^2 dt\right] \\
 &= E\left[\sum_{j=1}^m \int_{I_{j,m}} |c_j^* - \widehat{c}_j|^2 dt\right] \\
 &= \sum_{j=1}^m \int_{I_{j,m}} E[|c_j^* - \widehat{c}_j|^2] dt \\
 &= \sum_{j=1}^m \int_{I_{j,m}} \frac{E[W^2]}{k} dt \\
 &\leq \sum_{j=1}^m \int_{I_{j,m}} \frac{\sigma^2}{k} dt \\
 &= \sum_{j=1}^m \frac{1}{m} \frac{\sigma^2}{k} = \frac{\sigma^2}{k} = \frac{m}{n} \sigma^2
 \end{aligned}$$

Combining all the facts derived we have

$$E[\|f^* - \widehat{f}_n\|^2] \leq \frac{L^2}{m^2} + \frac{m}{n} \sigma^2 = O\left(\max\left\{\frac{1}{m^2}, \frac{m}{n}\right\}\right).^1 \tag{4}$$

What is the best choice of m ? If m is small then the approximation error (*i.e.*, $O(1/m^2)$) is going to be large, but the estimation error (*i.e.*, $O(m/n)$) is going to be small, and vice-versa. This two conflicting goals provide a tradeoff that directs our choice of m (as a function of n). In Figure 2 we depict this tradeoff. In Figure 2(a) we considered a large m_n value, and we see that the approximation of f^* by a function in the class \mathcal{F}_n can be very accurate (that is, our estimate will have a small bias), but when we use the measured data our estimate looks very bad (high variance). On the other hand, as illustrated in Figure 2(b), using a very small m_n allows our estimator to get very close to the best approximating function in the class \mathcal{F}_n , so we have a low variance estimator, but the bias of our estimator (*i.e.*, the difference between f_n and f^*) is quite considerable.

¹The notation $x_n = O(y_n)$ (that reads “ x_n is big- O y_n ”, or “ x_n is of the order of y_n as n goes to infinity”) means that $x_n \leq C y_n$, where C is a positive constant and y_n is a non-negative sequence.

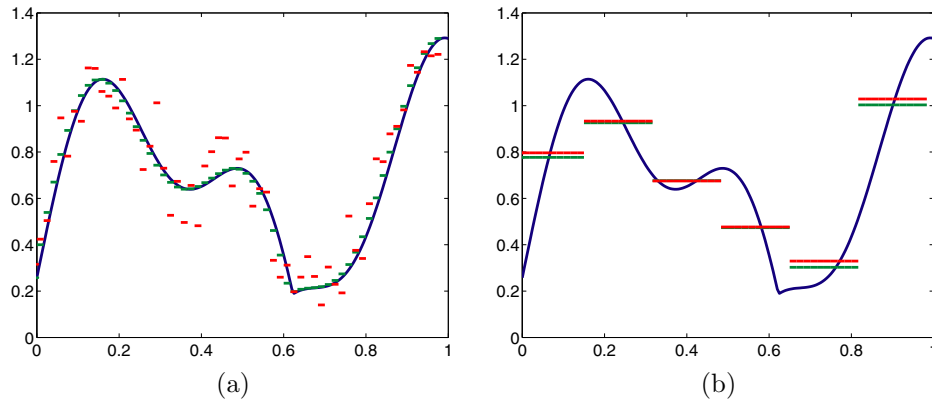


Figure 2: Approximation and estimation of f^* (in blue) for $n = 60$. The function f_n is depicted in green and the function \hat{f}_n is depicted in red. In (a) we have $m = 60$ and in (b) we have $m = 6$.

We need to balance the two terms in the right-hand-side of (4) in order to maximize the rate of decay (with n) of the expected risk. This implies that $\frac{1}{m^2} = \frac{m}{n}$ therefore $m_n = n^{1/3}$ and the Mean Squared Error (MSE) is

$$E[\|f_n - \hat{f}_n\|^2] = O(n^{-2/3}).$$

So the sieve $\mathcal{F}_1, \mathcal{F}_2, \dots$ with

$$\mathcal{F}_n = \left\{ f : f(t) = \sum_{j=1}^{m_n} c_j \mathbf{1}_{\{\frac{j-1}{m_n} \leq t < \frac{j}{m_n}\}}, c_j \in \mathbf{R} \right\},$$

produces a \mathcal{F} -consistent estimator for $f^* = E[Y|X + x] \in \mathcal{F}$.

It is interesting to note that the rate of decay of the MSE we obtain with this strategy cannot be further improved by using more sophisticated estimation techniques (that is, $n^{-2/3}$ is the *minimax* MSE rate for this problem). Also, rather surprisingly, we are considering classes of models \mathcal{F}_n that are actually not Lipschitz, therefore our estimator of f^* is not a Lipschitz function, unlike f^* itself.