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 \( \mathcal{Y} = \mathbb{R} \), independent of \( X \) and satisfying

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

Finally let \( f^* : [0, 1] \to \mathbb{R} \) be a function satisfying

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

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

Note that

\[ E[Y|X = x] = E[f^*(X) + W|X = x] \]
\[ = E[f^*(x) + W|X = x] \]
\[ = f^*(x) + E[W] = f^*(x). \]

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

\[ \{X_i, Y_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} P_{XY}, \]
\[ Y_i = f^*(X_i) + W_i, \ i = \{1, \ldots, n\}, \]
where \(i.i.d.\) means *independently and identically distributed*. Figure 1(a) illustrates this setup.

For simplicity we will consider a slightly different setting. In many applications we can sample \(X = [0, 1]\) as we like, and not necessarily at random. For example we can take \(n\) samples uniformly spaced on \([0, 1]\)

\[
x_i = \frac{i}{n}, \quad i = 1, \ldots, n,
\]

\[
Y_i = f^*(x_i) + W_i = f^* \left( \frac{i}{n} \right) + W_i.
\]

We will proceed with this setup (as in Figure 1(b)) in the rest of the lecture.

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

Let

\[
F = \{ f : f \text{ is Lipschitz with constant } L \}.
\]

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

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

For the estimation task we will use stair functions. Let \(m \in \mathbb{N}\) and define the class of piecewise constant functions

\[
F_m = \left\{ f : f(t) = \sum_{j=1}^m c_j 1_{\left( \frac{j-1}{m}, \frac{j}{m} \right]}, \quad c_j \in \mathbb{R} \right\}.
\]

\(F_n\) is the space of functions that are constant on intervals

\[
I_{j,m} = \left[ \frac{j-1}{m}, \frac{j}{m} \right], \quad j = 1, \ldots, m.
\]

Clearly if \(m\) is rather large we can approximate almost any bounded function arbitrarily well. So it make some sense to use these classes to construct a set of sieves.

Let \(0 < m_1 \leq m_2 \leq m_3 \leq \cdots\) be a sequence of integers satisfying \(m_n \to \infty\) as \(n \to \infty\). That is, for each value of \(n\) there is an associated integer value \(m_n\). Define the **Sieve** \(F_1, F_2, F_3, \ldots, F_{m_n} = \left\{ f : f(t) = \sum_{j=1}^{m_n} c_j 1_{\{t \in I_{j,m_n}\}}, \quad c_j \in \mathbb{R} \right\}.
\)

From here on we will use \(m\) instead of \(m_n\) and \(I_j\) instead of \(I_{j,m}\) for notational ease.

Define \(\hat{f}(t) \in F_m\) to be an approximation of \(f^*\), in particular

\[
\hat{f}(t) = \sum_{j=1}^m \bar{c}_j 1_{\{t \in I_j\}}, \quad \text{where} \quad \bar{c}_j = \frac{1}{|N_j|} \sum_{i : \frac{i}{n} \in I_j} f^* \left( \frac{i}{n} \right),
\]

where \(N_j = \{ i \in \{1, \ldots, n\} : \frac{i}{n} \in I_j \}\). Let \(|N_j|\) be the number of elements of \(N_j\), and assume \(m\) is not too large relative to \(n\) so that \(|N_j| > 0\). In fact \([\frac{m}{n}] \leq |N_j| \leq \frac{n}{m}\) so as long as \(m = m_n\) grows slightly slower than \(n\) we are okay.
Exercise 1  Upper bound the error of approximation of $\|f^* - \bar{f}\|^2$.

$$\|f^* - \bar{f}\|^2 = \int_0^1 |f^*(t) - \bar{f}(t)|^2 dt$$
$$= \sum_{j=1}^m \int_{I_j} |f^*(t) - \bar{f}(t)|^2 dt$$
$$= \sum_{j=1}^m \int_{I_j} |f^*(t) - \bar{c}_j|^2 dt$$
$$= \sum_{j=1}^m \int_{I_j} \left| f^*(t) - \frac{1}{|N_j|} \sum_{i \in I_j} f^* \left( \frac{i}{n} \right) \right|^2 dt$$
$$= \sum_{j=1}^m \int_{I_j} \left( \frac{1}{|N_j|} \sum_{i \in I_j} \left| f^*(t) - f^* \left( \frac{i}{n} \right) \right| \right)^2 dt$$
$$\leq \sum_{j=1}^m \int_{I_j} \left( \frac{1}{|N_j|} \sum_{i \in I_j} \left| f^*(t) - f^* \left( \frac{i}{n} \right) \right| \right)^2 dt$$
$$\leq \sum_{j=1}^m \int_{I_j} \left( \frac{1}{|N_j|} \sum_{i \in I_j} L_m \right)^2 dt$$
$$= \sum_{j=1}^m \int_{I_j} \left( \frac{L}{m} \right)^2 dt$$
$$= \sum_{j=1}^m \frac{1}{m} \left( \frac{L}{m} \right)^2 = \left( \frac{L}{m} \right)^2.$$

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

Of course we cannot compute $\bar{f}$ without knowing $f^*$, so let’s use the data to find a good model in $F_m$. For any $f \in F_m$, $f = \sum_{j=1}^m c_j 1_{\{t \in I_j\}}$, we have

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^m c_j 1_{\{t \in I_j\}} - Y_i \right)^2$$
$$= \frac{1}{n} \sum_{j=1}^m \left( \sum_{i \in I_j} (c_j - Y_i)^2 \right).$$

Let $\hat{f}_n = \arg \min_{f \in F_m} \hat{R}_n(f)$. Then

$$\hat{f}_n(t) = \sum_{j=1}^m \hat{c}_j 1_{\{t \in I_j\}}, \quad \text{where} \quad \hat{c}_j = \frac{1}{|N_j|} \sum_{i \in I_j} Y_i \quad (2)$$
Exercise 2  Show \([2]\).

Note that \(E[\hat{c}_j] = \bar{c}_j\) and therefore \(E[\hat{f}_n(t)] = \bar{f}(t)\). Let’s analyze now the expected risk of \(\hat{f}_n\):

\[
E[\|f^* - \hat{f}_n\|^2] = E[\|f^* - \bar{f} + \bar{f} - \hat{f}_n\|^2] \\
= \|f^* - \bar{f}\|^2 + E[\|\bar{f} - \hat{f}_n\|^2] + 2E[(f^* - \bar{f}, \bar{f} - \hat{f}_n)] \\
= \|f^* - \bar{f}\|^2 + E[\|\bar{f} - \hat{f}_n\|^2] + 2(f^* - \bar{f}, E[\bar{f} - \hat{f}_n]) \\
= \|f^* - \bar{f}\|^2 + E[\|\bar{f} - \hat{f}_n\|^2],
\]

where the final step follows from the fact that \(E[\hat{f}_n(t)] = \bar{f}(t)\). A couple of important remarks pertaining the right-hand-side of equation \([3]\): The first term, \(\|f^* - \bar{f}\|^2\), corresponds to the approximation error, and indicates how well can we approximate the function \(f^*\) with a function from \(\mathcal{F}_m\). Clearly, the larger the class \(\mathcal{F}_m\) is, the smaller we can make this term. This term is precisely the squared bias of the estimator \(\hat{f}_n\).

The second term, \(E[\|\bar{f} - \hat{f}_n\|^2]\), is the estimation error, the variance of our estimator. We will see that the estimation error is small if the class of possible estimators \(\mathcal{F}_m\) is also small.

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

\[
E[\|\bar{f} - \hat{f}_n\|^2] = E \left[ \int_0^1 |\bar{f}(t) - \hat{f}_n(t)|^2 dt \right] \\
= E \left[ \int_0^1 \sum_{j=1}^m (\hat{c}_j - \bar{c}_j)^2 1_{(t \in T_j)} dt \right] \\
= E \left[ \sum_{j=1}^m \int_{T_j} (\hat{c}_j - \bar{c}_j)^2 dt \right] \\
= \frac{1}{m} \sum_{j=1}^m E \left[ (\hat{c}_j - \bar{c}_j)^2 \right] \\
= \frac{1}{m} \sum_{j=1}^m E \left[ \left( \frac{1}{|N_j|} \sum_{i \in N_j} (f^*(i/n) - Y_i) \right)^2 \right] dt \\
= \frac{1}{m} \sum_{j=1}^m E \left[ \left( \frac{1}{|N_j|} \sum_{i \in N_j} W_i \right)^2 \right] dt \\
= \frac{1}{m} \sum_{j=1}^m \frac{\sigma^2}{|N_j|} \\
\leq \frac{1}{m} \sum_{j=1}^m \frac{\sigma^2}{|n/m|} \\
= \sigma^2 \frac{1}{|n/m|} \approx \sigma^2 \frac{m}{n} \leq (1 + \epsilon) \sigma^2 \frac{m}{n},
\]

for any \(\epsilon > 0\) provided \(|n/m|\) is large enough.

Combining all the facts derived we have

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

\(^1\)The 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.
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 \( F_{m_n} \) 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 \( F_n \), 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.

![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 \( m = 6 \).](image)

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 \( \frac{1}{m^2} = \frac{m_n}{n} \) therefore \( m_n = n^{1/3} \) and the Mean Squared Error (MSE) is

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

So the sieve \( F_{m_1}, F_{m_2}, \cdots \) with \( m_n \approx n^{1/3} \) produces a \( F \)-consistent estimator for \( f^* \in F \).

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 \( F_n \) that are actually not Lipschitz, therefore our estimator of \( f^* \) is not a Lipschitz function, unlike \( f^* \) itself.