# A Simple Regression Problem

R. M. Castro

March 23, 2011

In this brief note a simple regression problem will be introduced, illustrating clearly the bias-variance tradeoff.

Let

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

where $x_i = i/n$, $f : [0,1] \to \mathbb{R}$ is a function, and $W_i$'s are independent random variables such that

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

The object of interest is the function $f$. Using the data $\{Y_i\}$ we want to construct an estimate $\hat{f}_n$ that is good, in the sense that the squared $L_2$ distance

$$\|\hat{f}_n - f\|^2 = \int_0^1 (\hat{f}_n(t) - f(t))^2 dt \ ,$$

is small (note that the above is a random quantity). In particular we want to minimize the *expected risk*

$$\mathbb{E}[\|\hat{f}_n - f\|^2] \ .$$

In order to characterize the expected risk we need further assumptions on the function $f$, namely we assume it is Lipschitz smooth. Formally we assume

$$f \in \mathcal{F}_L \equiv \mathcal{F} = \{f : [0,1] \to \mathbb{R} \ : \ |f(s) - f(t)| \leq L|t-s|, \ \forall t, s \in [0,1]\} \ ,$$

where $L > 0$ is a constant. Notice that such functions are continuous, but not necessarily differentiable. An example of such a function is depicted in Figure 1(a).

Our approach will use piecewise constant functions, in what is usually referred to as a *regressogram* (this is the regression analogue of the histogram). Let $m \in \mathbb{N}$ and define the class of piecewise constant functions

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

The set $\mathcal{F}_m$ is a linear space consisting of functions that are constant on the intervals

$$\mathrm{I}_{j,m} \equiv \left[ \frac{j-1}{m}, \frac{j}{m} \right), \ j = 1, \ldots, m \ .$$

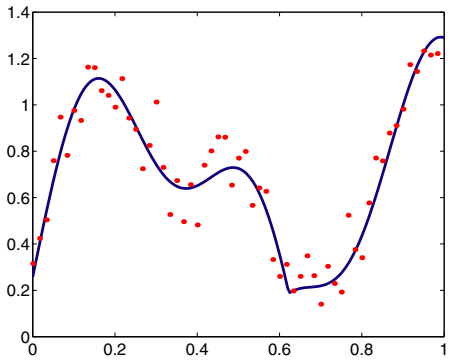


Figure 1: Example of a Lipschitz function (blue), and corresponding observations (red): The red dots correspond to $(i/n, Y_i),\ i = 1, \ldots, n$.

Clearly if $m$ is large we can approximate almost any bounded function arbitrarily well. For notational ease we will drop the subscript $m$ in $I_{j,m}$ and use simply $I_j$.

We are going to use a bias-variance decomposition. First let's define our estimator. It is going to be simply the average of the data in each one of the intervals $Y_i$. A way to motivate this estimator is as follows. Our goal is to minimize $\mathbb{E}[\|\hat{f}_n - f\|^2]$, but obviously we cannot compute this expectation. Let's consider instead an empirical surrogate for it, namely

$$\hat{R}_n(f') = \frac{1}{n}\sum_{i=1}^{n}(f'(x_i) - Y_i)^2 \ ,$$

where $f'$ is an arbitrary function. Now let $f' \in \mathcal{F}_m$, so that we can write it as

$$f'(t) = \sum_{j=1}^{m} c_j \mathbf{1}\{t \in I_j\} \ ,$$

where $c_j \in \mathbb{R}$. Define

$$N_j = \{i : x_i \in I_j\} \ .$$

We can rewrite the $\hat{R}_n(f')$ as

$$\begin{aligned} \hat{R}_n(f') &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} c_j \mathbf{1}\{x_i \in I_j\} - Y_i\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{m}\sum_{i \in N_j}(c_j - Y_i)^2 \ . \end{aligned}$$

Define the estimator

$$\hat{f}_n = \arg\min_{f' \in \mathcal{F}_m} \hat{R}_n(f') \ . \tag{1}$$

Then

$$\hat{f}_n(t) = \sum_{j=1}^{m} \hat{c}_j 1\{t \in I_j\} \ ,$$

where

$$\hat{c}_j = \frac{1}{|N_j|} \sum_{i \in N_j} Y_i \ , \tag{2}$$

where $|N_j|$ denotes the number of elements in $N_j$. Notice that $|N_j|$ is always greater than zero provided $m < n$. We will assume this throughout the entire document.

**Exercise 1** *Prove that the solution of* (1) *is given by* $\hat{f}_n(t) = \sum_{j=1}^{m} \hat{c}_j 1\{t \in I_j\}$, *where the* $\hat{c}_j$*'s are given by* (2).

Define also $\bar{f} \in \mathcal{F}_m$, the expected value of $\hat{f}_n$:

$$\bar{f}(t) = \mathbb{E}[\hat{f}_n(t)] = \sum_{j=1}^{m} \bar{c}_j \mathbf{1}\{t \in I_j\} \ , \qquad \bar{c}_j = \frac{1}{|N_j|} \sum_{i \in I_j} f(x_i) \ .$$

We are ready to do our bias-variance decomposition.

$$\begin{aligned}
\mathbb{E}[\|\hat{f}_n - f\|^2] &= \mathbb{E}[\|\hat{f}_n - \bar{f} + \bar{f} - f\|^2] \\
&= \mathbb{E}[\|\hat{f}_n - \bar{f}\|^2] + \mathbb{E}[\|\bar{f} - f\|^2] + 2\mathbb{E}[\langle \hat{f}_n - \bar{f}, \bar{f} - f \rangle] \\
&= \mathbb{E}[\|\hat{f}_n - \bar{f}\|^2] + \|\bar{f} - \hat{f}_n\|^2 + 2\langle \mathbb{E}[\hat{f}_n] - \bar{f}, \bar{f} - f \rangle \\
&= \mathbb{E}[\|\hat{f}_n - \bar{f}\|^2] + \|\bar{f} - \hat{f}_n\|^2 \ ,
\end{aligned}$$

where the final step follows from the fact that $\mathbb{E}[\hat{f}_n(t)] = \bar{f}(t)$. So the expected risk is decomposed in two terms, the first is the variance (or estimation error), and the second is the squared bias (or approximation error). Now we just need

to evaluate each one of these terms. Let's start with the bias term.

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

So we see that if $m$ is large (provided it is smaller than $n$) the bias term goes to zero. In other words we can approximate a Lipschitz smooth function arbitrarily

well with a piecewise constant function. Now for the variance.

$$\begin{aligned}
\mathbb{E}[\|\hat{f}_n - \bar{f}\|^2] &= \mathbb{E}\left[\int_0^1 (\hat{f}_n(t) - \bar{f}(t))^2 dt\right] \\
&= \mathbb{E}\left[\sum_{j=1}^m \int_{I_j} (\hat{f}_n(t) - \bar{f}(t))^2 dt\right] \\
&= \mathbb{E}\left[\sum_{j=1}^m \int_{I_j} (\hat{c}_j - \bar{c}_j)^2 dt\right] \\
&= \mathbb{E}\left[\sum_{j=1}^m (\hat{c}_j - \bar{c}_j)^2 \int_{I_j} dt\right] \\
&= \mathbb{E}\left[\sum_{j=1}^m (\hat{c}_j - \bar{c}_j)^2 \frac{1}{m}\right] \\
&= \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[(\hat{c}_j - \bar{c}_j)^2\right] \\
&= \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[\left(\frac{1}{|N_j|}\sum_{i\in N_j} Y_i - \frac{1}{|N_j|}\sum_{i\in N_j} f(x_i)\right)^2\right] dt \\
&= \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[\left(\frac{1}{|N_j|}\sum_{i\in N_j} (Y_i - f(x_i))\right)^2\right] dt \\
&= \frac{1}{m}\sum_{j=1}^m \mathbb{E}\left[\left(\frac{1}{|N_j|}\sum_{i\in N_j} W_i\right)^2\right] dt \\
&= \frac{1}{m}\sum_{j=1}^m \frac{\sigma^2}{|N_j|} \ .
\end{aligned}$$

Now notice that $|N_j| \approx n/m$. In fact, if we want to be precise we can say that $|N_j| \geq \lfloor n/m \rfloor$, where $\lfloor x \rfloor$ is the largest integer $k$ such that $k < x$. Therefore $|N_j| \geq n/m - 1$, and so

$$\begin{aligned}
\mathbb{E}[\|\hat{f}_n - \bar{f}\|^2] &\leq \frac{1}{m}\sum_{j=1}^m \frac{\sigma^2}{\lfloor n/m \rfloor} \\
&= \frac{\sigma^2}{\lfloor n/m \rfloor} \leq \frac{\sigma^2}{n/m - 1} = \sigma^2 \frac{m}{n}\left(\frac{n}{n-m}\right) \ .
\end{aligned}$$

So, as long as $m < cn$, with $0 < c < 1$ then

$$\mathbb{E}[\|\hat{f}_n - \bar{f}\|^2] \leq \sigma^2 \frac{m}{n} \frac{1}{1-c} \ ,$$

so the variance term is essentially proportional to $m/n$. In words this means the variance term is proportional to the number of model parameters $m$ divided by the amount of data $n$.

Combining everything we have

$$\mathbb{E}[\|\hat{f}_n\|^2 - f\|^2] \leq \sigma^2 \frac{m}{n} + \frac{L^2}{m^2} = O\left(\max\left\{\frac{1}{m^2}, \frac{m}{n}\right\}\right) \ , \tag{3}$$

where we make use of the Big-O notation[1]. At this point it becomes clear that there is an optimal choice for $m$, namely if $m$ is small then the squared bias term $O(1/m^2)$ is going to be large, but the variance term $O(m/n)$ 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 $m$ value, and we see that the approximation of $f$ by a function in the class $\mathcal{F}_m$ 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 $m$ allows our estimator to get very close to the best approximating function in the class $\mathcal{F}_m$, so we have a low variance estimator, but the bias of our estimator (the difference between $\bar{f}$ and $f$) is quite considerable.

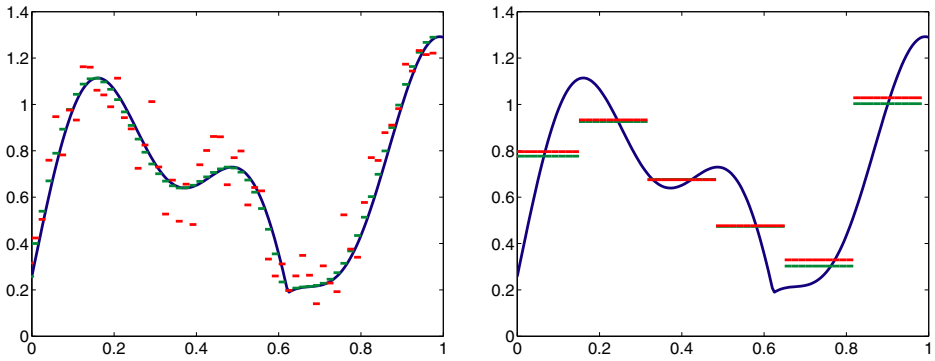


Figure 2: Approximation and estimation of $f$ (in blue) for $n = 60$. The function $\bar{f}$ is depicted in green and the function $\hat{f}_n$ is depicted in red. In (a)we have $m = 60$ and in (b) we have $m = 6$.

[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 $x_n \leq Cy_n$, where $C$ is a positive constant and $y_n$ is a non-negative sequence.

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

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

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. In fact we have the following *minimax lower bound*:

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}_m} \mathbb{E}[\|\hat{f}_n - f\|^2] \geq c(L, \sigma^2) n^{-2/3} \ ,$$

where $c(L, \sigma^2) > 0$, and the infimum is taken over all possible estimators (i.e., all measurable functions of the data).

Also, rather surprisingly, we are considering classes of models $\mathcal{F}_m$ that are actually not Lipschitz, therefore our estimator of $f$ is not a Lipschitz function, unlike $f$ itself.

**Exercise 2** *Suppose that the true regression function $f$ was not really a Lipschitz smooth function, but instead a piecewise Lipshitz functions. These are functions that are composed by a finite number of pieces that are Lipschitz. An example of such a function is $g(t) = f_1(t)\mathbf{1}\{t \in [0, 1/3]\} + f_2(t)\mathbf{1}\{t \in (1/3, 1/2)\} + f_3(t)\mathbf{1}\{t \in (1/2, 1]\}$, where $f_1, f_2, f_3 \in \mathcal{F}_L$.*

*Let $\mathcal{G}(M, L, R)$ denote the class of bounded piecewise Lipschitz functions. Each piece belongs to class $\mathcal{F}_L$, there are at most $M$ pieces, and any function $f \in \mathcal{G}(M, L, R)$ is bounded in the sense that $|f(x)| \leq R$ for all $x \in [0, 1]$. Study the performance of the above estimator when $f \in \mathcal{G}(M, L, R)$. Identify the best rate of error decay of the estimator risk.*

**Exercise 3** *Suppose you want to remove noise from an image (e.g., a medical image). An image can be thought of as a function $f : [0, 1]^2 \to \mathbb{R}$. Let's suppose it satisfies a 2-dimensional Lipschitz condition*

$$|f(x_1, y_1) - f(x_2, y_2)| \leq L \max(|x_1 - x_2|, |y_1 - y_2|) \quad , \forall x_1, y_1, x_2, y_2 \in [0, 1] \ .$$

1. *Do you think this is a good model for images? Why and why not.*
2. *Assume $n$, the number of samples you get from the function, is the square of an integer, therefore $\sqrt{n} \in \mathbb{N}$. Let $f$ be a function satisfying the above condition let the observation model be*

$$Y_{ij} = f(i/\sqrt{n}, j/\sqrt{n}) + W_{ij}, \quad i, j \in \{1, \ldots, \sqrt{n}\} \ ,$$

*where as before the noise variables are mutually independent and again $E[W_{ij}] = 0$ and $E[W_{ij}^2] \leq \sigma^2 < \infty$.*

*Using a similar approach to the one in class construct an estimator $\hat{f}_n$ for $f$. Using this procedure what is the best rate of convergence attainable when $f$ is a 2-dimensional Lipschitz function?*