Lectures 12 and 13 - Complexity Penalized Maximum Likelihood Estimation

Rui Castro

May 5, 2013

1 Introduction

As you learned in previous courses, if we have a statistical model we can often estimate unknown "parameters" by the maximum likelihood principle. Suppose we have independent, but not necessarily identically distributed, data. Namely, we model the data $\{Y_i\}_{i=1}^n$ as independent random variables with densities (with respect to a common dominating measure) given by $p_i(\cdot;\theta)$, where θ is an unknown "parameter"[1]. The Maximum Likelihood Estimator (MLE) of θ is simply given by

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \prod_{i=1}^n p_i(Y_i|\theta) \ ,$$

where Θ is the set of possible parameters. If the support of $p_i(\cdot;\theta)$ is not a function of θ then we re-write the above estimator in terms of the log-likelihood

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n \log p_i(Y_i|\theta) \ . \tag{1}$$

We will consider this setting in what follows.

Why is maximum likelihood a good idea? To better understand this let's introduce the Kullback-Leibler divergence.

Definition 1 (Kullback-Leibler Divergence)**.** *Convention* $0\cdot\log 0 = \lim_{x\to 0^+} x\log x = 0$*. Let* p *and* q *be two densities (for simplicity say these are with respect to the Lebesgue measure). Then the Kullback-Leibler divergence is defined as*

$$KL(p\|q) = \begin{cases} \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx & \text{if } p \ll q \\ \infty & \text{otherwise} \end{cases} \ ,$$

where $p \ll q$ *means* q dominates p*, that is, for almost all* x *if* $q(x) = 0$ *then* $p(x) = 0$*.*

Here are some important facts about the KL divergence.

[1] The word parameter is in between quotes as θ can be infinite dimensional, in what is generally referred to as non-parametric estimation. For instance, this is the case in non-parametric regression.

- $\mathrm{KL}(p\|q) \geq 0$, with equality if and only if $p(x) = q(x)$ almost everywhere, which means p and q represent the same probability measure. The proof of this follows easily from Jensen's inequality.

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\
&= \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] \ , \qquad \text{where } X \text{ is a r.v. with density } p \\
&= \mathbb{E}_p\left[-\log \frac{q(X)}{p(X)}\right] \\
&\geq -\log \mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] \\
&= -\log \int p(x) \frac{q(x)}{p(x)} dx \\
&= -\log \int q(x) dx = -\log 1 = 0 \ .
\end{aligned}$$

  where the inequality follows from Jensen's inequality and it is strict unless $p(x) = q(x)$ almost everywhere (note: there are minute technicalities omitted in the above proof).

- $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$. In other words, the KL divergence is not symmetric. It is therefore not a distance (in addition, it doesn't satisfy the triangle inequality). Let's see an example: let $p(x) = \mathbf{1}\{x \in [0,1]\}$ and $q(x) = \mathbf{1}\{x \in [0,2]\}/2$. Then $\mathrm{KL}(p\|q) = \log 2$ and $\mathrm{KL}(q\|p) = \infty$.

- The KL divergence is related to testing between two simple hypothesis.If the null hypothesis corresponds to density p and the alternative corresponds to density q the probabilities of type I and II error are intimately related to $\mathrm{KL}(p\|q)$ and $\mathrm{KL}(q\|p)$ respectively. In particular larger KL divergences correspond to smaller the error probability.

- **The product-density property:** Let p_i and q_i, $i \in 1, \ldots, n$ be univariate densities and define the following multivariate densities

$$\mathbf{p}(x_1, \ldots, x_n) = p_1(x_1) \cdot p_2(x_2) \cdots p_n(x_n) \ ,$$

$$\mathbf{q}(x_1, \ldots, x_n) = q_1(x_1) \cdot q_2(x_2) \cdots q_n(x_n) \ .$$

  Then

$$\mathrm{KL}(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^n \mathrm{KL}(p_i\|q_i) \ .$$

  The proof is quite simple and left as an exercise.

Now let's look again at the MLE formulation in (1). Begin by assuming Y_i are independent random variables with density q_i. We can re-write the MLE as

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{i=1}^n \log \frac{q_i}{p_i(Y_i; \theta)} \ .$$

The expected value of the r.h.s of the above is simply

$$\mathrm{KL}(\mathbf{q}\|\mathbf{p}(\theta)) \ ,$$

where $\mathbf{q} \equiv q_1(\cdot)\cdots q_n(\cdot)$ and $\mathbf{p}(\theta) \equiv p_1(\cdot;\theta)\cdots p_n(\cdot;\theta)$ are the product densities under the true data distribution and a model parameterized by θ, respectively. Therefore we see that the MLE is attempting to find the parameter $\theta \in \Theta$ that is "closer" to the true data distribution, where distance is measured as the KL divergence between the distributions of the data. Actually, with some small additional assumptions one can argue, by the strong law of large numbers, that

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{q_i}{p_i(Y_i;\theta)} - \frac{1}{n}\mathrm{KL}(\mathbf{q}\|\mathbf{p}(\theta)) \overset{\text{a.s.}}{\to} 0 \ ,$$

as $n \to \infty$, and that $\frac{1}{n}\mathrm{KL}(\mathbf{q}\|\mathbf{p}(\theta))$ converges to some point $c > 0$. So, in a sense, the MLE is asymptotically identifying the model in Θ that is closer to the true distribution of the data.

Example 1. The regression setting with deterministic design: *Let*

$$Y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where ϵ_i are i.i.d. normal random variables with variance σ^2. Then each Y_i has density

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - r(x_i)^2}{2\sigma^2}\right) \ .$$

The joint likelihood of $Y_1, \ldots, Y_n$ is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - r(x_i)^2}{2\sigma^2}\right) \ ,$$

and the MLE is simple given by

$$\hat{r}_n = \arg\min_{r\in\mathcal{R}} \sum_{i=1}^{n} (Y_i - r(x_i))^2 \ .$$

So we see that we recover our familiar sum of squared residuals in this case.

In what follows we'll make use of the KL divergence to characterize the behavior of the MLE for a variety of settings. For this we'll need two other measures between densities.

1.1 The Hellinger Distance and the Affinity

As we argued above, the KL divergence is not a proper distance (in particular it is not symmetric and does not satisfy the triangle inequality). However, it is not hard to define other distances between densities. Notably useful is the Hellinger distance

Definition 2 (Hellinger Distance)**.** *Let p and q be two densities with respect to the Lebesgue measure. The Hellinger distance is defined as*

$$H(p,q) = \left(\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx\right)^{1/2} \ .$$

This is a proper distance, as it is symmetric, since $H(p,q) = H(q,p)$, non-negative, and it satisfies the triangle inequality. Remarkably the squared Hellinger distance provides a lower bound for the KL divergence, so convergence in KL divergence implies convergence of the Hellinger distance.

Proposition 1.

$$H^2(p,q) \leq KL(p\|q) \ .$$

Proof.

$$\begin{aligned}
H^2(p,q) &= \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx \\
&= \int p(x)dx + \int q(x)dx - 2\int \sqrt{p(x)q(x)}dx \\
&= 2\left(1 - \int \sqrt{p(x)q(x)}dx\right) \\
&\leq -2\log \int \sqrt{p(x)q(x)}dx \ , \quad \text{since } 1 - x \leq -\log x \\
&= -2\log \int \sqrt{\frac{q(x)}{p(x)}} p(x)dx \\
&= -2\log \mathbb{E}\left[\sqrt{\frac{q(X)}{p(X)}}\right] \ , \quad \text{where } X \text{ has density } p \\
&\leq -2\mathbb{E}\left[\log \sqrt{q(X)/p(X)}\right] , \quad \text{by Jensen's inequality} \\
&= E\left[\log(p(X)/q(X))\right] \ \equiv \ \mathrm{KL}(p\|q) \ .
\end{aligned}$$

□

Note that we have also showed that

$$H^2(p,q) = 2\left(1 - \int \sqrt{p(x)q(x)}dx\right) \leq -2\log \int \sqrt{p(x)q(x)}dx \ .$$

The quantity inside the log is called the Affinity, and is a measure of the similarity between two densities (valued 1 is these are identical, and zero if they have disjoint support).

Definition 3 (Affinity)**.** *Let p and q be two densities with respect to the Lebesgue measure. The Affinity is defined as*

$$A(p,q) = \int \sqrt{p(x)q(x)}dx \ .$$

In summary, we have shown that

$$H^2(p,q) \leq -2\log A(p,q) \leq \mathrm{KL}(p\|q) \ .$$

1.2 The important case of Gaussian distributions

An important case to consider is when p and q are Gaussian distributions with the same variance but different means. Let

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} .$$

Then

$$\begin{aligned}
\mathrm{KL}(p_{\theta_1} \| p_{\theta_2}) &= \int \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} p_{\theta_1}(x) dx \\
&= \int \frac{(x-\theta_2)^2 - (x-\theta_1)^2}{2\sigma^2} p_{\theta_1}(x) dx \\
&= \mathbb{E}\left[\frac{(X-\theta_2)^2 - (X-\theta_1)^2}{2\sigma^2}\right] \quad \text{where } X \text{ has density } p_{\theta_1} \\
&= \frac{1}{2\sigma^2}\mathbb{E}[(X-\theta_2)^2] - \frac{1}{2\sigma^2}\mathbb{E}[(X-\theta_1)^2] \\
&= \frac{1}{2\sigma^2}\mathbb{E}[(X-\theta_1+\theta_1-\theta_2)^2] - \frac{1}{2\sigma^2}\mathbb{E}[(X-\theta_1)^2] \\
&= \frac{1}{2\sigma^2}(\sigma^2 + (\theta_1-\theta_2)^2) - \frac{1}{2\sigma^2}\sigma^2 \\
&= \frac{(\theta_1-\theta_2)^2}{2\sigma^2} .
\end{aligned}$$

So, in this case, the KL divergence is symmetric.

Now, for the affinity and Hellinger distance we proceed in a similar fashion.

$$\begin{aligned}
A(p_{\theta_1}, p_{\theta_2}) &= \int \sqrt{p_{\theta_1}(x) p_{\theta_2}(x)} dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta_1)^2}{4\sigma^2}} e^{-\frac{(x-\theta_2)^2}{4\sigma^2}} dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2-2\theta_1 x+\theta_1^2+x^2-2\theta_2 x+\theta_2^2}{4\sigma^2}} dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2-2\theta_1 x+\theta_1^2+x^2-2\theta_2 x+\theta_2^2}{4\sigma^2}} dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2-2\frac{\theta_1+\theta_2}{2}x+\frac{\theta_1^2+\theta_2^2}{2}}{2\sigma^2}} dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x-\frac{\theta_1+\theta_2}{2}\right)^2-\left(\frac{\theta_1+\theta_2}{2}\right)^2+\frac{\theta_1^2+\theta_2^2}{2}}{2\sigma^2}} dx \\
&= e^{\frac{\left(\frac{\theta_1+\theta_2}{2}\right)^2-\frac{\theta_1^2+\theta_2^2}{2}}{2\sigma^2}} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(x-\frac{\theta_1+\theta_2}{2}\right)^2}{2\sigma^2}} dx \\
&= e^{\frac{\left(\frac{\theta_1+\theta_2}{2}\right)^2-\frac{\theta_1^2+\theta_2^2}{2}}{2\sigma^2}} \\
&= e^{-\frac{(\theta_1-\theta_2)^2}{8\sigma^2}} .
\end{aligned}$$

Therefore

$$H^2(p_{\theta_1}, p_{\theta_2}) = 2\left(1 - e^{-\frac{(\theta_1-\theta_2)^2}{8\sigma^2}}\right) \ ,$$

and

$$-2\log A(p_{\theta_1}, p_{\theta_2}) = \frac{(\theta_1-\theta_2)^2}{4\sigma^2} \ .$$

2 Maximum Penalized Likelihood Estimation (MPLE)

Often time, in non-parametric settings, it is necessary to either restrict the choice of possible models under consideration, or penalize the MLE criterion. The second approach is more general and we will consider it here. Namely, we are going to study estimators of the form

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left\{-\sum_{i=1}^{n}\log p_i(Y_i;\theta) + \mathrm{pen}(\theta)\right\} \ ,$$

where pen(θ) is intended to penalize "complex" models. Next we prove the following important result.

Theorem 1 (Li-Barron 2000, Kolaczyk-Nowak 2002)**.** *Let $\mathbf{Y}$ be a random vector (e.g. $\mathbf{Y} = (Y_1, \ldots, Y_n)$). Assume the distribution of $\mathbf{Y}$ has unknown density p^*. Now suppose we have a class of probability density functions p_θ parameterized by $\theta \in \Theta$. Assume Θ is a countable set, and that for each θ we have an associated number $c(\theta)$ so that the following inequality holds*

$$\sum_{\theta\in\Theta} 2^{-c(\theta)} \leq 1 \quad \textit{(Kraft Inequality)} \ . \tag{2}$$

Define the Maximum Penalized Likelihood Estimator (MPLE) as

$$\hat{\theta}_n \equiv \arg\min_{\theta\in\Theta}\left\{-\log p_\theta(\mathbf{Y}) + 2c(\theta)\log 2\right\} \ .$$

Then

$$\begin{aligned} \mathbb{E}\left[H^2(p^*, p_{\hat{\theta}_n})\right] &\leq -2\mathbb{E}\left[\log A(p^*, p_{\hat{\theta}_n})\right] \\ &\leq \min_{\theta\in\Theta}\left\{KL(p^*\|p_\theta) + 2c(\theta)\log 2\right\} \ . \end{aligned}$$

This type of result, known as an *oracle bound*, tells us that the performance of the proposed estimator is essentially as good as the performance of a clairvoyant estimator where we replace the likelihood function by the theoretical counterpart (the KL divergence). Note also the result is extremely general - there are no assumptions made on the distribution of $\mathbf{Y}$ other than having a density. The major weakness of this result is that it only applies to countable (or finite) classes of candidate models. As we will see this creates some inconvenience, but we can still analyze many interesting settings by discretizing/quantizing the classes of models under consideration. There are possible extension of results of this type to uncountable classes of models, but these require several technical assumptions and heavy empirical processes machinery to prove. Nevertheless the main ideas, intuition and results are essentially preserved. Finally, note that this result gives you performance bounds for maximum likelihood estimation for finite classes of models, as in that case one can take $c(f) = \log_2|\Theta|$ and the estimator reduces to the usual MLE (as the penalty term does not depend on θ).

2.1 The Gaussian Regression Case

Before proving the theorem, let's look at a very special and important case. We will use these results quite a lot in the what follows. Recall the Gaussian regression model, used in many of the previous lectures. Suppose

$$Y_i = r^*(x_i) + \epsilon_i \quad , \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \ ,$$

where x_i are deterministic covariates (e.g., $x_i = i/n$), and r^* is the true, unknown, regression function. The density of Y_i is parameterized by r and given by

$$p_i(y_i; r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - r(x_i))^2}{2\sigma^2}} \ , \tag{3}$$

with $r = r^*$. We have seen before for this case we have

$$\text{KL}(p_i(r^*) \| p_i(r)) = \frac{1}{2\sigma^2}(r^*(x_i) - r(x_i))^2 \ ,$$

and

$$-2 \log A(p_i(r^*), p_i(r)) = \frac{1}{4\sigma^2}(r^*(x_i) - r(x_i))^2 \ .$$

This means that, for the joint density of the data, we have

$$\text{KL}(\mathbf{p}(r^*) \| \mathbf{p}(r)) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (r^*(x_i) - r(x_i))^2$$

and

$$-2 \log A(\mathbf{p}(r^*), \mathbf{p}(r)) = \frac{1}{4\sigma^2} \sum_{i=1}^{n} (r^*(x_i) - r(x_i))^2 \ .$$

Finally, as we have seen $-\log \mathbf{p}(\mathbf{Y}; r) = \sum_{i=1}^{n} \frac{(Y_i - r(x_i))^2}{2\sigma^2} + \text{const}$, where the constant term does not depend on r. So, using all this together with the above theorem we have the following, important corollary.

Corollary 1. *Let*

$$Y_i = r^*(x_i) + \epsilon_i \quad , \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \ ,$$

where x_i are deterministic. Let $\mathcal{R}$ be a class of models such that there is a map $c : \mathcal{R} \to [0, \infty]$ satisfying

$$\sum_{r \in \mathcal{R}} 2^{-c(r)} \leq 1 \ .$$

Define

$$\hat{r}_n = \arg\min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - r(x_i))^2 + \frac{4\sigma^2 c(r) \log 2}{n} \right\} \ .$$

Then

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\hat{r}_n(x_i) - r^*(x_i))^2 \right] \leq 2 \min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r(x_i) - r^*(x_i))^2 + \frac{4\sigma^2 c(r) \log 2}{n} \right\} .$$

This corollary shows that the maximum penalized likelihood estimator can perform almost as well (we have a factor 2 in the bound) has if we had observed $r^*(x_i)$ instead of Y_i (i.e., if we had removed the noise from the observations and used the same estimator).

Proof of Theorem 1. We have already shown that

$$H^2(p^*, p_\theta) \leq -2 \log A(p^*, p_\theta) \ ,$$

So, clearly

$$\mathbb{E}\left[H^2(p^*, p_{\hat{\theta}_n})\right] \leq \mathbb{E}\left[-2 \log A(p^*, p_{\hat{\theta}_n})\right] \ .$$

We are going to bound the right-hand-side of the above expression. Begin by defining the theorectical analog of $\hat{\theta}_n$

$$\tilde{\theta}_n = \arg\min_{\theta \in \Theta} \left\{\mathrm{KL}(p^* \| p_\theta) + 2c(f) \log 2\right\} \ .$$

Let's re-write the definition of the MPLE

$$\begin{aligned}
\hat{\theta}_n &= \arg\min_{\theta \in \Theta} \left\{-\log p_\theta(\mathbf{Y}) + 2c(\theta) \log 2\right\} \\
&= \arg\max_{\theta \in \Theta} \left\{\log p_\theta(\mathbf{Y}) - 2c(\theta) \log 2\right\} \\
&= \arg\max_{\theta \in \Theta} \left\{\frac{1}{2} \log p_\theta(\mathbf{Y}) - c(\theta) \log 2\right\} \\
&= \arg\max_{\theta \in \Theta} \left\{\log \sqrt{p_\theta(\mathbf{Y})} - c(\theta) \log 2\right\} \\
&= \arg\max_{\theta \in \Theta} \left\{\sqrt{p_\theta(\mathbf{Y})} \exp\left(-c(\theta) \log 2\right)\right\} \\
&= \arg\max_{\theta \in \Theta} \left\{\sqrt{p_\theta(\mathbf{Y})} 2^{-c(\theta)}\right\}
\end{aligned}$$

This means that, for ANY $\theta \in \Theta$

$$\sqrt{p_{\hat{\theta}_n}(\mathbf{Y})} 2^{-c(\hat{\theta}_n)} \geq \sqrt{p_\theta(\mathbf{Y})} 2^{-c(\theta)} \ .$$

In particular we have

$$\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p_{\tilde{\theta}_n}(\mathbf{Y})}} \frac{2^{-c(\hat{\theta}_n)}}{2^{-c(\tilde{\theta}_n)}} \geq 1 \ .$$

With this fact in hand let's proceed with the bound

$$\begin{aligned}
\mathbb{E}\left[-2\log A(p^*,p_{\hat{\theta}_n})\right] &= 2\mathbb{E}\left[\log\frac{1}{A(p^*,p_{\hat{\theta}_n})}\right] \\
&\le 2\mathbb{E}\left[\log\left(\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p_{\tilde{\theta}_n}(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{2^{-c(\tilde{\theta}_n)}}\frac{1}{A(p^*,p_{\hat{\theta}_n})}\right)\right] \\
&= 2\mathbb{E}\left[\log\left(\sqrt{\frac{p^*(\mathbf{Y})}{p_{\tilde{\theta}_n}(\mathbf{Y})}}\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{2^{-c(\tilde{\theta}_n)}}\frac{1}{A(p^*,p_{\hat{\theta}_n})}\right)\right] \\
&= \mathbb{E}\left[\log\frac{p^*(\mathbf{Y})}{p_{\tilde{\theta}_n}(\mathbf{Y})}\right] + 2c(\tilde{\theta}_n)\log 2 + \mathbb{E}\left[\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{A(p^*,p_{\hat{\theta}_n})}\right] \\
&= \mathrm{KL}(p^*\|p_{\tilde{\theta}_n}) + 2c(\tilde{\theta}_n)\log 2 \\
&\qquad +\mathbb{E}\left[\log\left(\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{A(p^*,p_{\hat{\theta}_n})}\right)\right] .
\end{aligned}$$

So, the first two terms of the last line are exactly what we have in the statement of the theorem. To conclude the proof we just need to show the last term is always smaller than zero. Note that there are two random quantities in that term, namely $\mathbf{Y}$ and $\hat{\theta}_n$ (which is a function of $\mathbf{Y}$). Next, we use first Jensen's inequality and then a very crude bound to essentially get rid of $\hat{\theta}_n$, taking advantage of the fact that $\hat{\theta}_n \in \Theta$.

First note that, by Jensen's inequality

$$\mathbb{E}\left[\log\left(\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{A(p^*,p_{\hat{\theta}_n})}\right)\right] \le \log\mathbb{E}\left[\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{A(p^*,p_{\hat{\theta}_n})}\right]$$

Now we can make use of the Kraft inequality through a union bound. In this case the union bound is simply saying that any individual term in a summation of positive terms is smaller than the summation.[2] So, we bound the inside of the expectation by a sum over Θ.

$$\begin{aligned}
\log\mathbb{E}\left[\sqrt{\frac{p_{\hat{\theta}_n}(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\hat{\theta}_n)}}{A(p^*,p_{\hat{\theta}_n})}\right] &\le \log\mathbb{E}\left[\sum_{\theta\in\Theta}\sqrt{\frac{p_\theta(\mathbf{Y})}{p^*(\mathbf{Y})}}\frac{2^{-c(\theta)}}{A(p^*,p_\theta)}\right] \\
&= \log\left(\sum_{\theta\in\Theta}2^{-c(\theta)}\frac{\mathbb{E}\left[\sqrt{\frac{p_\theta(\mathbf{Y})}{p^*(\mathbf{Y})}}\right]}{A(p^*,p_\theta)}\right) \\
&= \log\left(\sum_{\theta\in\Theta}2^{-c(\theta)}\right) \\
&\le \log 1 = 0
\end{aligned}$$

[2] Let $z_1, z_2, \ldots$ be non-negative. Then for all $i \in \mathbb{N}$ $z_i \le \sum_{j=1}^{\infty} z_j$

where the last step follows from Kraft's inequality, and the previous step follows simply by noting that

$$\mathbb{E}\left[\sqrt{\frac{p_\theta(\mathbf{Y})}{p^*(\mathbf{Y})}}\right] = \int \sqrt{\frac{p_\theta(\mathbf{y})}{p^*(\mathbf{y})}} p^*(\mathbf{y}) d\mathbf{y} = \int \sqrt{p_\theta(\mathbf{y}) p^*(\mathbf{y})} d\mathbf{y} = A(p^*, p_\theta) \ .$$

□

2.2 Choosing the values $c(\theta)$

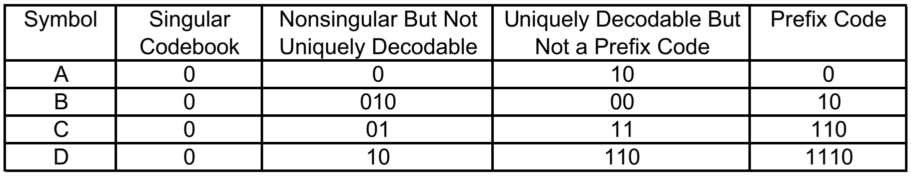
In order to apply the theorem or the corollary we need to construct a map $c(\cdot)$ from Θ to $[0,\infty)$. These values can be though of as a measure of the *complexity* of each model θ. If one has a proper probability mass distribution over Θ, say $P_{\text{prior}} : \Theta \to [0,1]$, then we can just use $c(\theta) = -\log_2 P_{\text{prior}}(\theta)$. Clearly this satisfies the Kraft inequality. Moreover this estimator as a strong Bayesian interpretation - it corresponds to the *Maximum a Posteriori* estimator for the prior P_{prior}. Another way to satisfy the Kraft inequality is to use coding arguments. This has the advantage that we can very easily devise maps that automatically satisfy the Kraft inequality.

Assume we have assign a binary codeword (sequence of 0s and 1s) to each element of Θ. Let $c(\theta)$ denote the size of the codeword (size of the sequence). The set Θ is called the alphabet, and the idea is to encode sequences of symbols from the alphabet by concatenating the corresponding codewords. A very useful class of codes are called **prefix codes**.

Definition 4. *A code is called a* prefix *or* instantaneous *code if no codeword is a prefix of any other codeword.*

The reason for such name and definition is clarified in the following example.

Example 2. *(From Cover & Thomas '91)*
Consider an alphabet of symbols, say A, B, C, and D and the codebooks in Figure 2. Suppose we want to convey the sentence ADBBAC. This will be encoded in each system respectively as 000000, 010010010001, 1011000001011, and 0111010100110. It is clear that with the singular codebook we assign the same codeword to each symbol - a system that is obviously flawed! In the second case the codes are not singular but the codeword 010 could represent B or CA or AD, so the above binary sequence is ambiguous, as it can be decoded both as ADBBAC or BBBAC for instance. Hence it is not a uniquely decodable codebook. The third and fourth cases are both examples of uniquely decodable codebooks, but the fourth has the added feature that no codeword is a prefix of another. Prefix codes can be decoded from left to right instantaneously *since each codeword is "self-punctuating" - that is, you know immediately when a codeword ends. Note that, until the tenth bit in the third case, you don't know if the sequence is ADBB... or ACBBB... and only then can you decide. In the fourth case you know immediately when a codeword ends and another starts.*



Symbol	Singular Codebook	Nonsingular But Not Uniquely Decodable	Uniquely Decodable But Not a Prefix Code	Prefix Code
A	0	0	10	0
B	0	010	00	10
C	0	01	11	110
D	0	10	110	1110

Figure 1: Four possible codes for an alphabet of four symbols.

The good thing about prefix codes is that these are easy to construct and automatically satisfy the Kraft inequality. We will use this approach often.

2.2.1 The Kraft Inequality

Theorem 2. *For any binary prefix code, the codeword lengths c_1, c_2, ... satisfy*

$$\sum_{k=1}^{\infty} 2^{-c_k} \leq 1 \ .$$

Conversely, given any c_1, c_2, ... satisfying the inequality above we can construct a prefix code with these codeword lengths.

Proof. We will prove that for any binary prefix code, the codeword lengths $c_1, c_2, \ldots$, satisfy $\sum_k 2^{-c_k} \leq 1$. The converse is easy to prove also, but it not central to our purposes here (for a proof, see Cover & Thomas '91). Consider a binary tree like the one shown in Figure 2.2.1

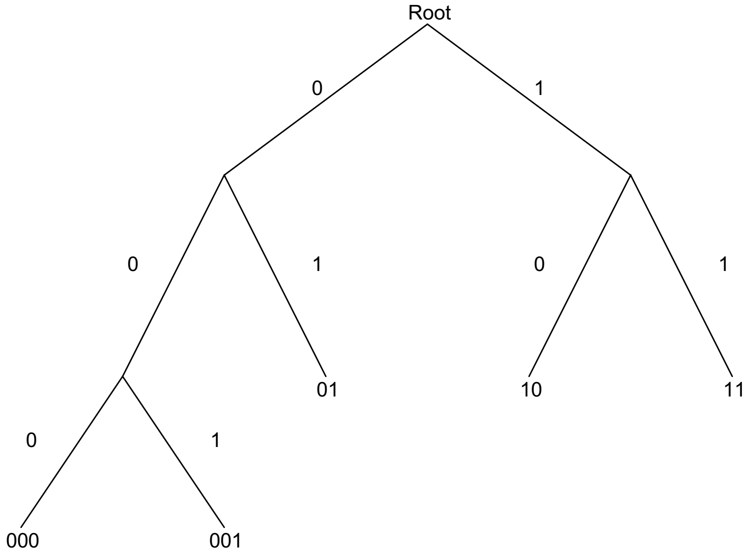


Figure 2: A binary tree.

The sequence of bit values leading from the root to a leaf of the tree represents a codeword. The prefix condition implies that no codeword is a descendant of any other codeword in the tree. Therefore we can each leaf of the tree represents a codeword in our code. Let $c_{\max}$ be the length of the longest codeword (also the number of branches to the deepest leaf) in the tree.

Consider a node in the tree at level c_i. This node in the tree can have at most $2^{c_{\max}-c_i}$ descendants at level $c_{\max}$. Furthermore, for each leaf of the tree the set of possible descendants at level $c_{\max}$ is disjoint (since no codeword can be a prefix of another). Therefore, since the total number of possible leafs at level $c_{\max}$ is $2^{c_{\max}}$, we have

$$\sum_{i \in \text{leafs}} 2^{c_{\max}-c_i} \leq 2^{c_{\max}} \quad \Rightarrow \quad \sum_{i \in \text{leafs}} 2^{-c_i} \leq 1 \ ,$$

which proves the case when the number of codewords is finite.

Suppose now that we have a countably infinite number of codewords. Let $b_1, b_2, \ldots, b_{c_i}$ be the i^{th} codeword and let

$$r_i = \sum_{j=i}^{c_i} b_j 2^{-j}$$

be the real number corresponding to the binary expansion of the codeword (in binary $r_i = 0.b_1\ b_2\ b_3 \ldots$). We can associate the interval $[r_i, r_i + 2^{-c_i})$ with the i^{th} codeword. This is the set of all real numbers whose binary expansion begins with $b_1, b_2, \ldots, b_{c_i}$. Since this is a subinterval of $[0, 1]$, and all such subintervals corresponding to prefix codewords are disjoint, the sum of their lengths must be less than or equal to 1. This proves the case where the number of codewords is infinite. □

3 Adapting to Unknown Parameters

Earlier in the course we saw that one can estimate Lipschitz smooth functions well. In this section we generalize those results to include smoother functions. Furthermore, we will be able to automatically adjust to the unknown level of smoothness of functions.

3.1 Lipschitz Functions

Suppose we have a function $r^* : [0, 1] \to [-R, R]$ satisfying the Lipschitz smoothness assumption

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

We have seen these functions can be well approximated by piecewise constant functions of the form

$$g(x) = \sum_{j=1}^{m} c_j \mathbf{1}\{x \in I_j\} \ , \quad \text{where } I_j = \left[\frac{j-1}{m}, \frac{j}{m}\right] \ .$$

Let's use maximum-likelihood to pick the "best" such model. Suppose we have following regression model

$$Y_i = r^*(x_i) + \epsilon_i \ , \quad i = 1, \ldots, n \ ,$$

where $x_i = i/n$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. To be able to use the corollary we derived we need a countable or finite class of models. The easiest way to do so is to discretize/quantize the possible values of each constant piece in the candidate models. Define

$$\mathcal{R}_m = \left\{ \sum_{j=1}^{m} c_j \mathbf{1}\{x \in I_j\} \ : c_j \in \mathcal{Q} \right\} \ ,$$

where

$$\mathcal{Q} = \{-R, -R\frac{n-1}{n}, \ldots, R\} = \left\{ R\frac{k-n}{n} \ , k = 0, \ldots, 2n \right\} \ .$$

Therefore $\mathcal{R}_m$ has exactly $(2n+1)^m$ elements in total. This means that, by taking $c(r) = \log_2\left((2n+1)^m\right) = m \log_2(2n+1)$ for all $r \in \mathcal{R}_m$ we satisfy the Kraft inequality

$$\sum_{r \in \mathcal{R}_m} 2^{-c(r)} = \sum_{r \in \mathcal{R}_m} \frac{1}{|\mathcal{R}_m|} = 1 \ .$$

So we are ready to apply our oracle bound. Since $c(r)$ is just a constant (not really a function of r) the estimator is simply the MLE

$$\begin{aligned}\hat{r}_n &= \arg\min_{r\in\mathcal{R}_m}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - r(x_i))^2 + \frac{4\sigma^2 c(r)\log 2}{n}\right\} \\ &= \arg\min_{r\in\mathcal{R}_m}\left\{\frac{1}{n}\sum_{i=1}^n (Y_i - r(x_i))^2\right\} .\end{aligned}$$

The corollary then says that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{r}_n(x_i) - r^*(x_i))^2\right] \quad\le\quad 2\min_{r\in\mathcal{R}_m}\left\{\frac{1}{n}\sum_{i=1}^n (r^*(x_i) - r(x_i))^2 + \frac{4\sigma^2 c(r)\log 2}{n}\right\} .$$

So far the result is extremely general, as we have not made use of the Lipschitz assumption. We have seen earlier that there is a piecewise constant function $\bar{r}_m(x) = \sum_{j=1}^m c_j \mathbf{1}\{x \in I_j\}$ such that for all $x \in [0,1]$ we have $|r^*(x) - \bar{r}_m(x)| \le L/m$. The problem is that, generally $\bar{r}_m \notin \mathcal{R}_m$ since $c_j \notin \mathcal{Q}$. Take instead the element of $\mathcal{R}_m$ that is closest to $\bar{r}_m$, namely

$$\tilde{r}_m = \arg\min_{f\in\mathcal{R}_m}\sup_{x\in[0,1]} |r(x) - \bar{r}_m(x)| \ .$$

It is clear that $|f(x) - \bar{f}_m(x)| \le R/n$ for all $x \in [0,1]$ therefore, by the triangle inequality we have

$$|f(x) - \tilde{f}_m(x)| \le |f(x) - \bar{f}_m(x)| + |\bar{f}_m(x) - \tilde{f}_m(x)| \le \frac{L}{m} + \frac{R}{n} \ .$$

Now, we can just use this in our bound

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{r}_n(x_i) - r^*(x_i))^2\right] &\le 2\min_{r\in\mathcal{R}_m}\left\{\frac{1}{n}\sum_{i=1}^n (r^*(x_i) - r(x_i))^2 + \frac{4\sigma^2 c(r)\log 2}{n}\right\} \\ &\le \frac{2}{n}\sum_{i=1}^n \left(\frac{L}{m} + \frac{R}{n}\right)^2 + \frac{8\sigma^2 m \log_2(2n+1)\log 2}{n} \\ &= 2\left(\frac{L}{m} + \frac{R}{n}\right)^2 + \frac{8\sigma^2 m \log(2n+1)}{n} \ .\end{aligned}$$

So, to ensure the best bound possible we should choose m minimizing the right-hand-side. This yields $m \sim (n/\log n)^{1/3}$ and

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{f}_n(x_i) - f^*(x_i))^2\right] = O\left((n/\log n)^{-2/3}\right)$$

which, apart from the logarithmic factor is the best we can ever hope for (this logarithmic factor is due to the discretization of the model classes, and is an artifact of this approach). If we want the truly best possible bound we need to minimize the above expression with respect to m, and for that we need to know L. Can we do better? Can we "automagically" choose m using the data? The answer is yes, and for this we will start taking full advantage of our oracle bound.

Since we want to choose the best possible m we must consider the following class of models

$$\mathcal{R} = \bigcup_{m=1}^{\infty} \mathcal{R}_m \ .$$

This is clearly a countable class of models (but not finite). So we need to be a bit more careful in constructing the map $c(\cdot)$. Let's use a coding argument: begin by defining

$$m(r) = \min_{m \in \mathbb{N}} \{ m : r \in \mathcal{R}_m \} \ .$$

Encode $r \in \mathcal{R}$ using first the bits $00 \ldots 01$ (total $m(r)$ bits) to encode $m(r)$ and them $\log_2 |\mathcal{R}_m|$ bits to encode which model inside $\mathcal{R}_m$ is r. This is clearly a prefix code and therefore satisfies the Kraft inequality. More formally

$$c(r) = m(r) + \log_2 |\mathcal{R}_{m(r)}| = m(r) + \log_2 \left((2n+1)^{m(r)} \right) = m(r)(1 + \log_2((2n+1)) \ .$$

Although we know, from the coding argument, that the map $c(\cdot)$ satisfies the Kraft inequality for sure, we can do a little sanity check, and ensure this is indeed true:

$$\begin{aligned}
\sum_{r \in \mathcal{R}} 2^{-c(r)} &\leq \sum_{m=1}^{\infty} \sum_{r \in \mathcal{R}_m} 2^{-c(r)} \\
&= \sum_{m=1}^{\infty} \sum_{r \in \mathcal{R}_m} 2^{-m(r) - \log_2 |\mathcal{R}_{m(r)}|} \\
&\leq \sum_{m=1}^{\infty} \sum_{r \in \mathcal{R}_m} 2^{-m - \log_2 |\mathcal{R}_m|} \\
&= \sum_{m=1}^{\infty} 2^{-m} \sum_{r \in \mathcal{R}_m} \frac{1}{|\mathcal{R}_m|} \\
&= \sum_{m=1}^{\infty} 2^{-m} = 1 \ .
\end{aligned}$$

Now, similarly to what we had before

$$\begin{aligned}
\hat{r}_n &= \arg\min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - r(x_i))^2 + \frac{4\sigma^2 c(r) \log 2}{n} \right\} \\
&= \arg\min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - r(x_i))^2 + \frac{4\sigma^2 m(r)(1 + \log_2(2n+1)) \log 2}{n} \right\} \ ,
\end{aligned}$$

which is no longer the MLE, but rather a maximum penalized likelihood estimator. Then

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(x_i)-r^*(x_i))^2\right] \\
\leq\;& 2\min_{r\in\mathcal{R}}\left\{\frac{1}{n}\sum_{i=1}^{n}(r^*(x_i)-r(x_i))^2+\frac{4\sigma^2 m(r)(1+\log_2(2n+1))\log 2}{n}\right\} \\
\leq\;& 2\min_{m\in\mathbb{N}}\left\{\min_{r\in\mathcal{R}_m}\left\{\frac{1}{n}\sum_{i=1}^{n}(r^*(x_i)-r(x_i))^2\right\}+\frac{4\sigma^2 m(1+\log_2(2n+1))\log 2}{n}\right\} \\
\leq\;& \min_{m\in\mathbb{N}}\left\{2\left(\frac{L}{m}+\frac{R}{n}\right)^2+\frac{8\sigma^2 m(\log 2+\log(2n+1))}{n}\right\} .
\end{aligned}$$

Therefore this estimator automatically chooses the best possible number of parameters m. Note that the price we pay is very very modest - the only change from what we had before was that the term $\log(2n+1)$ is replaced by $\log 2+\log(2n+1)$, which is a very minute change as $\log(2n+1)\gg\log 2$ for any interesting sample size. Although this is remarkable, we can probably do much better, and adjust to unknown smoothness. We'll see how to do this next.

4 Regression of Hölder smooth functions

Let's begin by defining a class of Hölder smooth functions. Let $\alpha>0$ and define let $H^\alpha(C,R)$ be the set of all functions $r:[0,1]\to[-R,R]$ which have $\lfloor\alpha\rfloor$ derivatives and

$$|f(x)-T_y^{\lfloor\alpha\rfloor}(x)|\leq C|x-y|^\alpha \quad \forall x,y\in[0,1] \ ,$$

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 $\lfloor\alpha\rfloor$.

Note that the above definition coincides with our Lipschitz function class when $\alpha=1$. Hölder smoothness essentially measures how differentiable functions are, and therefor Taylor polynomials are the natural way to approximate Hölder smooth functions. We will focus on Hölder smooth function classes with $0<\alpha\leq 2$ but the results presented can be easily generalized for larger values of α. Since $\alpha\leq 2$ 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...

4.1 Regression of Hölder smooth functions

Consider the usual regression model

$$Y_i=r^*(x_i)+\epsilon_i \ , \quad i=1,\ldots,n \ ,$$

where $x_i=i/n$ and $\epsilon_i\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,\sigma^2)$. Let's assume $r^*\in H^\alpha(C,R)$, with unknown smoothness $0<\alpha\leq 2$. Intuitively, the smoother r^* is the better we should be able to estimate it, as we can "average" more observations locally. In other words for smoother r^* we should average over larger bins. Also, we will need to exploit the extra smoothness in our models for r^*. To that end, we will consider candidate functions that are piecewise linear, i.e., functions of the form

$$\sum_{j=1}^{m}(a_j + b_j x)\mathbf{1}\{x \in I_j\} \ , \quad \text{where } I_j = \left[\frac{j-1}{m}, \frac{j}{m}\right] \ .$$

As before, we want to consider countable/finite classes of models to be able to apply our corollary, so we will consider a slight modification of the above. Each linear piece can be described by their beginning and end points respectively. So we are going to restrict those to lie on a grid. Namely refer to Figure 3

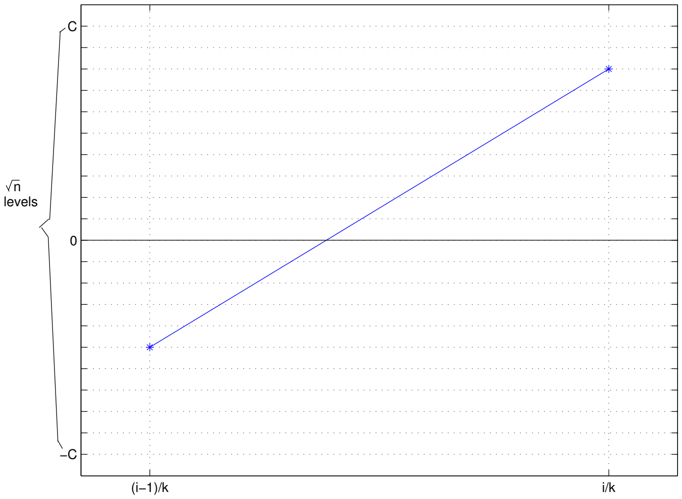


Figure 3: Example on the discretization of f on interval $\left[\frac{j-1}{m}, \frac{j}{m}\right)$

Define the class

$$\mathcal{R}_m = \left\{ r(x) = \sum_{j=1}^{m} \ell_j(x)\mathbf{1}\{x \in I_j\} \right\} \ ,$$

where

$$\ell_j(x) = \frac{x-(j-1)/m}{1/m} b_j + \frac{j/m - x}{1/m} a_j = (mx - j + 1)b_j + (j - mx)a_j \ ,$$

and $a_j, b_j \in \left\{ \frac{k-\sqrt{n}}{\sqrt{n}} R : k \in \{0, \ldots, 2\sqrt{n}\} \right\}$. Clearly $|\mathcal{R}_m| = (2\sqrt{n}+1)^{2m}$.

Since we don't know the smoothness a priori, we must choose m using the data. Therefore, in the same fashion as before we take the class

$$\mathcal{R} = \bigcup_{m=1}^{\infty} \mathcal{R}_m \ ,$$

with $m(r) = \min_{m \in \mathbb{N}}\{m : r \in \mathcal{R}_m\}$, and

$$c(r) = m(r) + \log_2 |\mathcal{R}_{m(r)}| = m(r)(1 + 2\log_2\left((2\sqrt{n}+1)\right) \ .$$

Exactly as before, define the estimator

$$\hat{r}_n \quad = \quad \arg\min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - r(x_i))^2 + \frac{4\sigma^2 m(r)(1 + 2\log_2(2\sqrt{n} + 1))\log 2}{n} \right\} \ ,$$

Then

$$\begin{aligned} & \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (\hat{r}_n(x_i) - r^*(x_i))^2 \right] \\ & \leq \ 2 \min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r^*(x_i) - r(x_i))^2 + \frac{4\sigma^2 m(r)(1 + 2\log_2(2\sqrt{n} + 1))\log 2}{n} \right\} \\ & \leq \ 2 \min_{m \in \mathbb{N}} \left\{ \min_{r \in \mathcal{R}_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r^*(x_i) - r(x_i))^2 \right\} + \frac{4\sigma^2 m(1 + 2\log_2(2\sqrt{n} + 1))\log 2}{n} \right\} \ . \end{aligned}$$

In the above, the first term is essentially our familiar approximation error, and the second term is in a sense bounding the estimation error. Therefore this estimator automatically seeks the best balance between the two. To say something more concrete about the performance of the estimator we need to bring in the assumptions we have on r^* (so far we haven't used any).

First, suppose $r^* \in H^\alpha(C, R)$ for $1 < \alpha \leq 2$. We need to find a"good" model in the class $\mathcal{R}_m$ that makes the approximation error small. Take $x \in I_j$ where j is arbitrary. From the definition of $H^\alpha(C, R)$ we have

$$|r^*(x) - r^*(j/m) - \frac{\partial}{\partial x} r^*(j/m)(x - j/m)| \leq C|x - j/m|^\alpha \leq C|1/m|^\alpha = Cm^{-\alpha} \ .$$

Therefore we can approximate $r^*(x)$ with $x \in I_j$ by a linear function, such that the error is bounded by $Cm^{-\alpha}$. In particular, the best piecewise linear approximation, defined as

$$r_m^* = \arg \min_{\text{piecewise linear functions}} \sup_{x \in [0,1]} |r(x) - r^*(x)| \ ,$$

satisfies $|r^*(x) - r_m^*(x)| \leq Cm^{-\alpha}$ for all $x \in [0, 1]$. Clearly, this function in not in $\mathcal{R}_m$, so we need to do some discretization. Take the function in $\mathcal{R}_m$ that is closest to that function r_m^*.

$$\tilde{r}_m^* = \arg \min_{r \in \mathcal{R}_m} \sup_{x \in [0,1]} |r(x) - r_m^*(x)| \ ,$$

because of the way we discretized we know that $\sup_{x \in [0,1]} |\tilde{r}_m^*(x) - r_m^*(x)| \leq R/\sqrt{n}$. Using this, together with the triangle inequality yields

$$\forall x \in [0, 1] \quad |\tilde{r}_m^*(x) - r^*(x)| \leq Cm^{-\alpha} + \frac{R}{\sqrt{n}} \ . \tag{4}$$

If $r^* \in H^\alpha(C, R)$ for $0 < \alpha \leq 1$ we can proceed in a similar fashion, but simply have to note that such functions are well approximated by piecewise constant functions. Furthermore these are a subset of $\mathcal{R}_m$. So, the reasoning we used for Lipschitz functions applies directly here, and we obtain expression (4) for that case as well.

Finally, we can just plug-in these results into the bound of the corollary.

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(x_i)-r^*(x_i))^2\right] \\
&\leq 2\min_{m\in\mathbb{N}}\left\{\min_{r\in\mathcal{R}_m}\left\{\frac{1}{n}\sum_{i=1}^{n}(r^*(x_i)-r(x_i))^2\right\}+\frac{4\sigma^2 m(1+2\log_2(2\sqrt{n}+1))\log 2}{n}\right\} \\
&\leq 2\min_{m\in\mathbb{N}}\left\{\left(Cm^{-\alpha}+\frac{R}{\sqrt{n}}\right)^2+\frac{4\sigma^2 m(1+2\log_2(2\sqrt{n}+1))\log 2}{n}\right\} \\
&\leq 2\min_{m\in\mathbb{N}} O\left(\max\left\{m^{-2\alpha},\frac{m^{-\alpha}}{n},\frac{1}{n},\frac{m\log n}{n}\right\}\right) .
\end{aligned}
$$

It is not hard to see that the first and last terms dominate the bound, and so we attain the minimum by taking (in the bound)

$$
m\sim\left(\frac{n}{\log n}\right)^{1/(2\alpha+1)} ,
$$

which yields

$$
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_n(x_i)-r^*(x_i))^2\right]=O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{2\alpha+1}}\right) .
$$

Note that the estimator does not know α!!! So we are indeed adapting to unknown smoothness. If the regression function r^* is Lipschitz this estimator has error rate $O\left((n/\log n)^{-2/3}\right)$. However, if the function is smoother (say $\alpha=2$) the estimator has error rate $O\left((n/\log n)^{-4/5}\right)$, which decays much quicker to zero. More remarkably, apart from the logarithmic factor it can be shown this is the best one can hope for! So the logarithmic factor is the very small price we need to pay for adaptivity.