

ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Statistical Learning Theory

So far in the course we have considered signal detection and estimation problems with parametric distributions.

Example 1. *Hypothesis testing:*

$$\begin{aligned} H_0 : x_i &\stackrel{iid}{\sim} p_0 \\ H_1 : x_i &\stackrel{iid}{\sim} p_1 \end{aligned}$$

Example 2. *Parametric estimation:*

$$x_i \stackrel{iid}{\sim} \text{Poisson}(\theta), \quad \hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

Example 3. *Signal estimator:*

$$x \sim \mathcal{N}(H\theta, \sigma^2 I), \quad \hat{\theta} = (H^T H)^{-1} H^T x$$

However in many problems we are faced with **unknown** distributional characteristics. The distribution generating the data may be non-parametric or even completely unknown.

Example 4. *Pattern Classification:* Suppose that $x \in \mathbb{R}^d$ are ‘features’ that can be used to predict the class ‘label’ $y = 0$ or 1 . This is similar to binary hypothesis testing if we have $p_0(x) = p(x|y = 0)$ and $p_1(x) = p(x|y = 1)$. Then given a feature x we can decide if it is more like a realization from p_0 or p_1 . However, suppose we don’t know p_0 and p_1 , but we have a set of labeled examples $\{(x_i, y_i)\}_{i=1}^n \stackrel{iid}{\sim} p(x, y)$. Given these data we can try to design a function to predict the proper label for other x .

Example 5. *Non-parametric Estimation:* Suppose that we make noisy observations of an unknown function f according to

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where the ϵ_i are iid noises with possibly unknown distribution. If the x_i are also iid, then

$$(x_i, y_i)_{i=1}^n \stackrel{iid}{\sim} p_{xy}(f).$$

How well can we estimate f from these data? If f were a parametric function of a single parameter, then we expect the MSE to be on the order of $\frac{1}{n}$. But what if f is a smooth (e.g. differentiable) but otherwise unknown function?

Density Estimation

Perhaps the most basic problem in statistical learning theory is density estimation. Suppose $x_i \stackrel{iid}{\sim} p$, $i = 1, \dots, n$ where the density p is unknown and doesn’t necessarily have a parametric form. For the moment let’s assume p could be any probability density function.

The most intuitive approach to density estimation is to estimate the density at a point x as

$$\hat{p}(x) \propto \text{the number of } x_i \text{ within a small neighborhood about } x$$

If there are more/less x_i near x , then the probability density is probably higher/lower at that point.

Histogram Density Estimators

Let's assume that the unknown density p is supported on the unit hypercube in d -dimensions, i.e. $\text{supp}(p) = [0, 1]^d$. We can always rescale any bounded region of support to the cube. Now divide $[0, 1]^d$ into m subcubes of sidelength $m^{-\frac{1}{d}}$. For example, if $d = 2$, then we have this partition of the unit square:

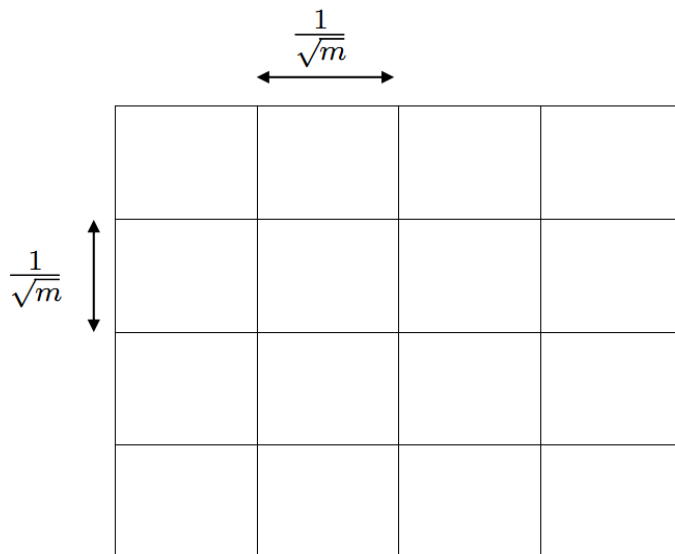


Figure 1: Partition of the hypercube

Let's call the subcubes 'bins' and enumerate them as B_j , $j = 1, \dots, m$. Let n_j be the number of $\{x_i\}$ in bin B_j , i.e.

$$n_j := \sum_{i=1}^n \mathbf{1}_{\{x_i \in B_j\}} .$$

The quantity

$$\hat{q}_j := \frac{n_j}{n}$$

is an estimator of the probability mass p places on bin B_j , i.e.

$$q_j := \int_{B_j} p(x) dx .$$

Now since \hat{q}_j is an unbiased estimator of the probability mass of B_j , we can estimate approximately the probability density on B_j as

$$\frac{\text{prob. mass}}{\text{volume}} = \frac{\hat{q}_j}{\frac{1}{m}} = m \hat{q}_j ,$$

which yields the *histogram density estimator* of the density function p

$$\begin{aligned} \hat{p}_m(x) &= \sum_{j=1}^m m \hat{q}_j \mathbf{1}_{\{x \in B_j\}} , \\ &= m \times \frac{1}{n} \times \{\#x_i \text{ in bin containing } x\} . \end{aligned}$$

How good of an estimator is \hat{p}_m ? Let's consider the squared error:

$$\|p - \hat{p}_m\|_2^2 = \int_{[0,1]^d} |p(x) - \hat{p}_m(x)|^2 dx .$$

The squared error is a random variable since it depends on the random sample $x_i \stackrel{iid}{\sim} p, i = 1, \dots, n$. So let's consider the MSE $\mathbb{E}[\|p - \hat{p}_m\|_2^2]$, where the expectation is with respect to the random sample used to design \hat{p}_m . The estimator \hat{p}_m is reasonable if it is consistent, i.e. if

$$\mathbb{E}[\|p - \hat{p}_m\|_2^2] \xrightarrow{n \rightarrow \infty} 0$$

Specifically this is L_2 or MSE consistency.

To analyze the MSE, first note that n_j is the number of samples in B_j out of a total of n , and q_j is the probability of a sample falling in the bin, therefore

$$n_j \sim \text{Binomial}(n, q_j) ,$$

and

$$\mathbb{P}(n_j = k) = \binom{n}{k} q_j^k (1 - q_j)^{n-k} .$$

Based on the distribution of n_j , we know $\mathbb{E}[n_j] = nq_j$, thus we have:

$$\mathbb{E}[\hat{q}_j] = \mathbb{E}\left[\frac{n_j}{n}\right] = q_j .$$

It also follows that $\text{var}(n_j) = nq_j(1 - q_j)$, thus we have:

$$\text{var}(\hat{q}_j) = \mathbb{E}[(q_j - \hat{q}_j)^2] = \frac{q_j(1 - q_j)}{n} .$$

Now consider the following decomposition into bias and variance. First define

$$p_m(x) = \mathbb{E}[\hat{p}_m] = \sum_{j=1}^m m q_j 1_{x \in B_j} .$$

This is the expected value of \hat{p}_m . Now write

$$\begin{aligned} \mathbb{E}[\|p - \hat{p}_m\|_2^2] &= \mathbb{E}[\|p - p_m + p_m - \hat{p}_m\|_2^2] , \\ &= \|p - p_m\|_2^2 + \mathbb{E}[\|p_m - \hat{p}_m\|_2^2] , \end{aligned}$$

since

$$\mathbb{E} \left[\int (p(x) - p_m(x))(p_m(x) - \hat{p}_m(x)) dx \right] = \int (p(x) - p_m(x)) \mathbb{E}[p_m(x) - \hat{p}_m(x)] dx = 0$$

The term $\|p - p_m\|_2^2$ measures the bias or 'approximation error' that is incurred by approximating the density as piecewise constant on the histogram partition. For example, if $d = 1$, then the picture looks like this:

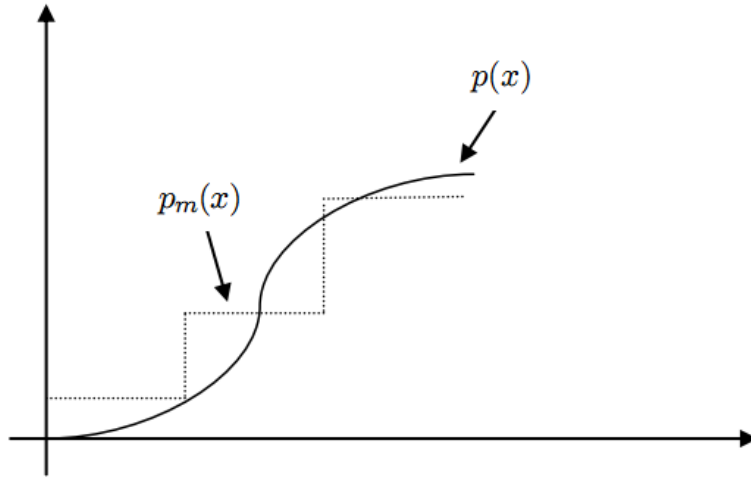


Figure 2: Picture of $p(x)$ and $p_m(x)$

It is clear that as $m \rightarrow \infty$, the approximation error $\|p - p_m\|_2^2 \rightarrow 0$. So more bins is better in this sense.

The term $\mathbb{E}[\|p_m - \hat{p}_m\|_2^2]$ is the variance or ‘stochastic error’ incurred because we must estimate the probability mass q_j on each bin using the training data $\{x_i\}_{i=1}^n$. Clearly, if $m \gg n$, then we will have no samples in most bins, which makes estimation impossible. So m should not be too large.

To get a better sense of this effect let’s look at the variance more closely.

$$\begin{aligned}
 \|p_m - \hat{p}_m\|_2^2 &= \int_{[0,1]^d} |p_m(x) - \hat{p}_m(x)|^2 dx \\
 &= \sum_{j=1}^m \int_{B_j} |mq_j - m\hat{q}_j|^2 dx \\
 &= \sum_{j=1}^m \frac{1}{m} |mq_j - m\hat{q}_j|^2 \\
 &= m \sum_{j=1}^m |q_j - \hat{q}_j|^2 .
 \end{aligned}$$

Therefore,

$$\mathbb{E}[\|p_m - \hat{p}_m\|_2^2] = m \sum_{j=1}^m \mathbb{E}[(q_j - \hat{q}_j)^2] = m \sum_{j=1}^m \frac{q_j(1 - q_j)}{n}$$

Since $q_j(1 - q_j) \leq q_j$, so we have the upper bound

$$\mathbb{E}[\|p_m - \hat{p}_m\|_2^2] \leq m \sum_{j=1}^m \frac{q_j}{n} = \frac{m}{n} .$$

If we want the variance to go to zero then a sufficient condition is

$$\frac{m}{n} \rightarrow 0 , \text{ as } n \rightarrow \infty .$$

To conclude so far:

$$\begin{aligned} \text{squared bias} &= \|p - p_m\|_2^2 \rightarrow 0 \text{ if } m \rightarrow \infty \\ \text{variance} &= \mathbb{E}[\|p_m - \hat{p}_m\|_2^2] \rightarrow 0 \text{ if } \frac{m}{n} \rightarrow 0 \end{aligned}$$

So we can take $m = m(n)$ any diverging function of n such that

$$\lim_{n \rightarrow \infty} \frac{m(n)}{n} = 0$$

For example, $m = \sqrt{n}$, $m = \log n$, etc. will suffice for consistency. What choice of m is the best? This depends on the underlying density p . If it is very smooth, like

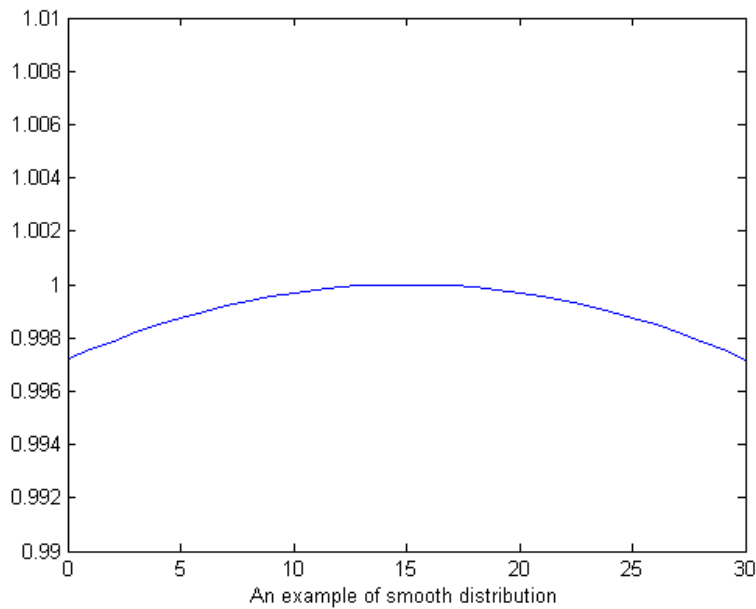


Figure 3: An example of smooth distribution

then large bins are best since the approximation error as well as the variance is small. However if the density is more irregular, say like this:

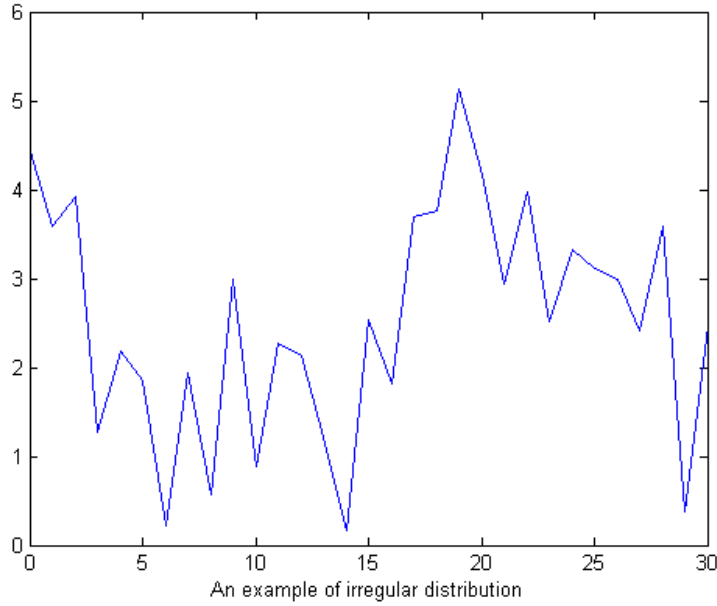


Figure 4: An example of irregular distribution

Then the approximation error will be very large unless we use many small bins. To move forward, we must make some assumptions about the smoothness of p . One of the least restrictive notions of smoothness is **Lipschitz regularity**.

We say that p is **Lipschitz smooth with constant $L > 0$** if:

$$|p(x) - p(y)| \leq L\|x - y\|_2 \quad \forall x, y \in [0, 1]^d$$

If $\|x - y\|_2$ is small, then $p(x) \approx p(y)$. Assuming p is Lipschitz we can bound the bias (approximation error) as follows:

$$\begin{aligned} \int |p(x) - p_m(x)|^2 dx &= \sum_{j=1}^m \int_{B_j} |p(x) - p_m(x)|^2 dx \\ &\leq \sum_{j=1}^m \int_{B_j} |p(x) - p(z_j)|^2 dx, \end{aligned}$$

where z_j is midpoint of bin B_j . The inequality follows from the fact that the best constant approximation (in a squared error sense) of p on B_j is the average value of p on B_j . So we have

$$\begin{aligned} \int |p(x) - p_m(x)|^2 dx &\leq \sum_{j=1}^m \int_{B_j} L^2 |x - z_j|^2 dx \\ &\leq \sum_{j=1}^m \int_{B_j} L^2 dm^{-\frac{2}{d}} dx = L^2 dm^{-\frac{2}{d}}, \end{aligned}$$

since the diameter of cube of sidelength $m^{-\frac{1}{d}}$ is $\sqrt{d} m^{-\frac{1}{d}}$ and $\int_{B_j} 1_{x \in B_j} dx = \frac{1}{m}$. So we now have two upper bounds

$$\|p - p_m\|_2^2 \leq d L^2 m^{-\frac{2}{d}}$$

and

$$\mathbb{E}[\|p_m - \hat{p}_m\|_2^2] \leq \frac{m}{n}.$$

Combining these yields

$$\text{MSE}(\hat{p}_m) = \mathbb{E}[\|p - \hat{p}_m\|_2^2] \leq dL^2 m^{-\frac{2}{d}} + \frac{m}{n}.$$

To minimize the upper bound we choose m so that both terms are roughly equal (since one is proportional to m and the other is inverse proportional to m). Ignoring constants

$$m^{-\frac{2}{d}} = \frac{m}{n} \Rightarrow m = n^{\frac{d}{2+d}}.$$

Plugging this choice back into the MSE bound yields

$$\text{MSE}(\hat{p}_m) \leq \text{constant} \times n^{-\frac{2}{2+d}}.$$

Note that the rate of convergence depends on the dimension d . Moreover, we can also prove that for the class of Lipschitz smooth density functions there does not exist any other form of estimator that has a convergence rate faster than $n^{-\frac{2}{2+d}}$. The rate is considerably slower in high dimensions, which is called ‘curse of dimensionality’.