ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** J. Jiao

Statistical Learning Theory

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

Example 1. *Hypothesis testing:*

$$H_0: \quad x_i \overset{iid}{\sim} \quad p_0$$
$$H_1: \quad x_i \overset{iid}{\sim} \quad p_1$$

Example 2. *Parametric estimation:*

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

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. *Parameter Estimation:*
Suppose that $x \in \mathbb{R}^d$ are 'features' that can be used to predict the class 'label':$y =$ 0 or 1.This is similiar to binary hypothesis testing if we have $p_0(x) = p(x|y = 0)$ and $p_1(x) = p(x|y = 1)$. Suppose we don't know p_0 and p_1, but we have a set of labeled examples $\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} p(x, y)$. 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.

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

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

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

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 (i.e. differentiable) but other unknown function?

Density Estimation

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

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

$$\hat{p}(x) \propto \# x_i \quad \text{falling in a small neighborhood about } x$$

If there are more/less x_i near x, then the probabilistic 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:

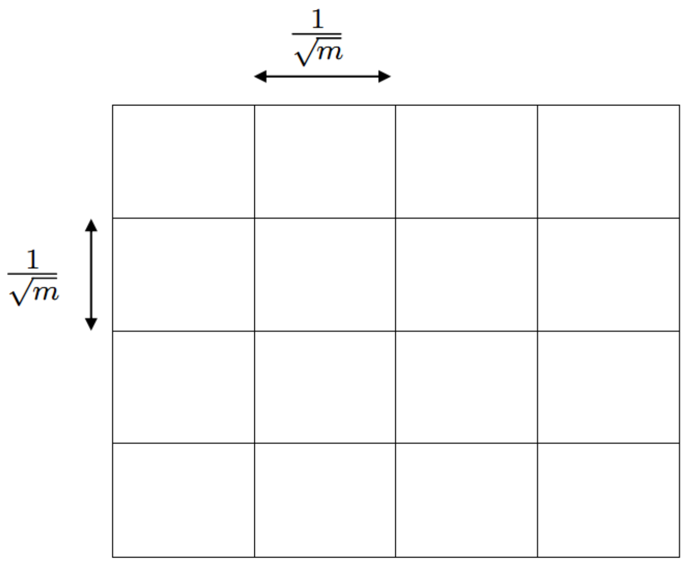

Figure 1: Partition of the hypercube

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

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

The quantity

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

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

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

Note that n_j is the number of samples out of a total of n in B_j and q_j is the probability of a sample falling in the bin, therefore:

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

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

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 $x_i \overset{iid}{\sim} p, i = 1, \ldots, n$. 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

$$\mathbb{E}[\|p - \hat{p}_m\|_2^2] \overset{n\to\infty}{\longrightarrow} 0$$

Specifically this is L_2 or MSE consistency.

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

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

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 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 following estimator of the density function p:

$$\hat{p}_m(x) = \sum_{j=1}^{m} m\hat{q}_j 1_{x\in B_j}$$

$$= m \times \frac{1}{n} \times \{\#x_i \text{ in bin containing } x\}$$

To analyze the MSE, consider the following decomposition into bias and variance.

$$\mathbb{E}[\|p - \hat{p}_m\|_2^2] = \mathbb{E}[\|p - p_m + p_m - \hat{p}_m\|_2^2]$$

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

Note that

$$\mathbb{E}[\|p - \hat{p}_m\|_2^2] = \|p - p_m\|_2^2 + \mathbb{E}[\|p_m - \hat{p}_m\|_2^2]$$

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

The quantity $\|p - p_m\|_2^2$ 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:

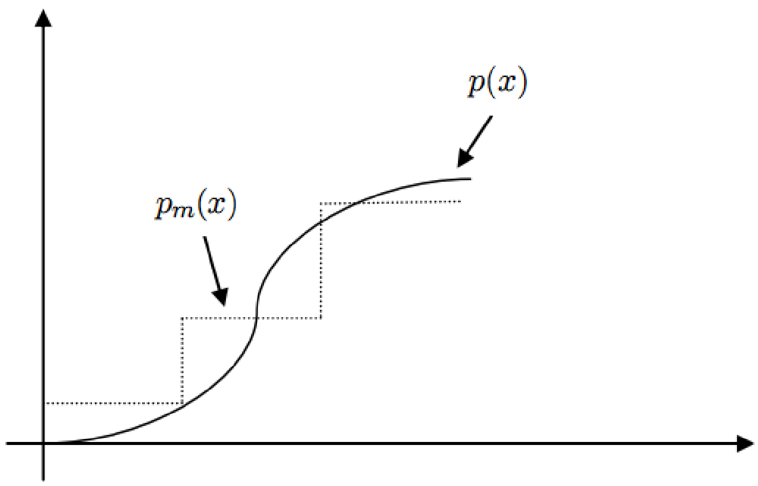

Figure 2: Picture of $p(x)$ and $p_m(x)$

It is clear that as $m \to \infty$, the approximation error $\|p - p_m\|_2^2 \to 0$.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.

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

Therefore,

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

As we know $\forall q_j \in [0,1]$, $q_j(1-q_j) \leq q_j$, so we have the upper bound for $\mathbb{E}[\|p_m - \hat{p}_m\|_2^2]$:

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

So we have shown that:

$$\mathbb{E}[\|p_m - \hat{p}_m\|_2^2] = C\frac{m}{n}, \text{for some constant } C > 0$$

If we want the variance to go to zero then we require:

$$\frac{m}{n} \xrightarrow{n \to \infty} 0$$

To conclude so far:

$$\begin{array}{rll} \text{The bias} & \|p - p_m\|_2^2 \to 0 & \text{if } m \to \infty \\ \text{The variance} & \mathbb{E}[\|p_m - \hat{p}_m\|_2^2] \to 0 & \text{if } \dfrac{m}{n} \to 0 \end{array}$$

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

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

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

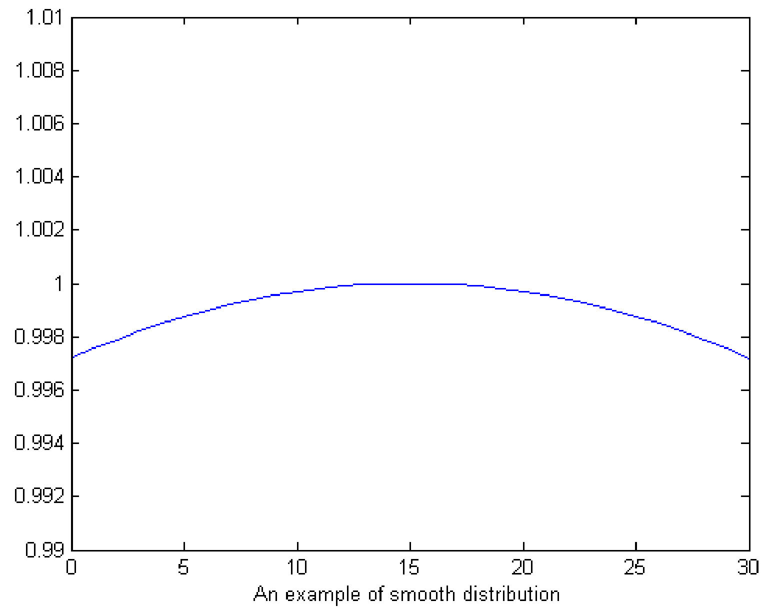

Figure 3: An example of smooth distribution

then large bins are best since the approximation error as well as the variance is small.

However if the density is more irregular, lay like this:

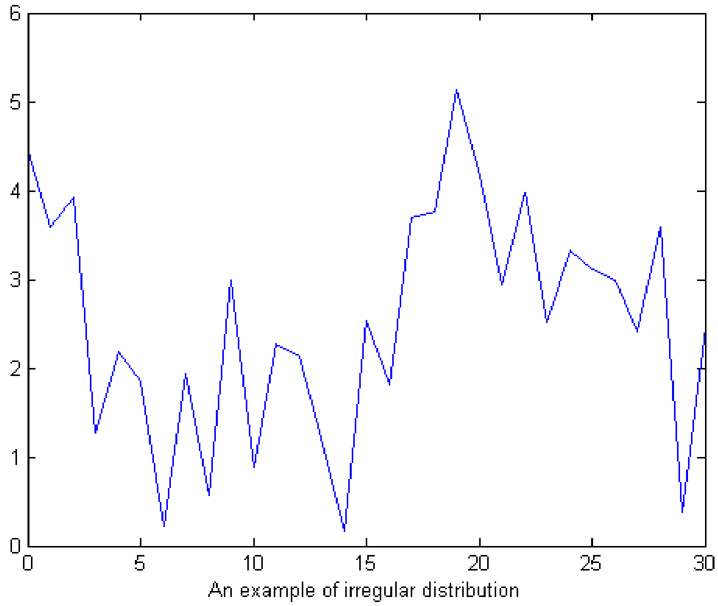

Figure 4: An example of irregular distribution

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 notion 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:

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

where z_j is midpoint of bin B_j.

Then we have:

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

since the diameter of cube of sidelength $m^{-\frac{1}{d}}$ is $dm^{-\frac{2}{d}}$ and $\int_{B_j} 1_{x \in B_j} dx = \frac{1}{m}$.

So we have:

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

and

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

as before, put them together, we have:

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

To minimize the upper bound we choose m so that both terms are 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 d.Moreover, we can also prove according to information theory that there is no Lipschitz smooth function that can perform better than the decreasing rate of $n^{-\frac{2}{2+d}}$.Obviously, the decreasing rate is considerably slower in high dimensions, which is called 'curse of dimensionality'.