ECE 901 Spring 2014 Statistical Learning Theory

instructor: R. Nowak

Minimax Lower Bounds

Lower Performance Bounds

Up to now in class, we've been analyzing estimators/predictors obtaining upper bounds on their performance. These bounds are of the form:

$$\min_{f \in \mathcal{F}} \mathbb{E}[d(\widehat{f}_n, f)] \leq C n^{-\gamma}$$

where $\gamma > 0$. We would like to know if these bounds are tight, in the sense that there is no other estimator that is significantly better. To answer this, we need lower bounds like

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}[d(\widehat{f}_n, f)] \geq c n^{-\gamma}$$

We assume we have the following ingredients:

* Class of models, $\mathcal{F} \subseteq \mathcal{S}$. $\mathcal{F}$ is a class of models containing the "true" model and is a subset of some bigger class $\mathcal{S}$. E.g. $\mathcal{F}$ could be the class of Lipschitz density functions or distributions P_{XY} satisfying the box-counting condition.

* An observation model, $\mathcal{P}_f$, indexed by $f \in \mathcal{F}$. $\mathcal{P}_f$ denotes the distribution of the data under model f. E.g. in regression and classification, this is the distribution of $Z = (X_1, Y_1, \ldots, X_n, Y_n) \subseteq \mathcal{Z}$. We will assume that $\mathcal{P}_f$ is a probability measure on the measurable space $(\mathcal{Z}, \mathcal{B})$.

* A performance metric $d(.,.) \geq 0$. If you have a model estimate $\widehat{f}_n$, then the performance of that model estimate relative to the true model f is $d(\widehat{f}_n, f)$. E.g.

$$\text{Regression:} \qquad d(\widehat{f}_n, f) = ||\widehat{f}_n - f||_2 = \left(\int (\widehat{f}_n(x) - f(x))^2 dx \right)^{1/2}$$

$$\text{Classification:} \qquad d(\widehat{f}_n, f) = R(\widehat{G}_n) - R^* = \int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| dP_X(x)$$

As before, we are interested in the risk of a learning rule, in particular the maximal risk given as:

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[d(\widehat{f}_n, f)] = \sup_{f \in \mathcal{F}} \int d(\widehat{f}_n(Z), f) d\mathcal{P}_f(Z)$$

where $\widehat{f}_n$ is a function of the observations Z and $\mathbb{E}_f$ denotes the expectation with respect to $\mathcal{P}_f$.

The main goal is to get results of the form

$$\mathcal{R}_n^* \stackrel{\Delta}{=} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}[d(\widehat{f}_n, f)] \geq c s_n$$

where $c > 0$ and $s_n \to 0$ as $n \to \infty$. The inf is taken over all estimators, i.e. all measurable functions $\widehat{f}_n : \mathcal{Z} \to \mathcal{S}$.

Suppose we have shown that

$$\liminf_{n\to\infty} s_n^{-1}\mathcal{R}_n^* \geq c > 0 \qquad \text{(A lower bound)}$$

and also that for a particular estimator $\bar{f}_n$

$$\limsup_{n\to\infty} s_n^{-1} \sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\bar{f}_n, f)] \leq C$$

$$\implies \limsup_{n\to\infty} s_n^{-1}\mathcal{R}_n^* \leq C,$$

We say that s_n is the optimal rate of convergence for this problem and that $\bar{f}_n$ attains that rate.

Note: Two rates of convergence Ψ_n and Ψ'_n are equivalent, i.e. $\Psi_n \equiv \Psi'_n$ iff

$$0 < \liminf_{n\to\infty} \frac{\Psi_n}{\Psi'_n} \leq \limsup_{n\to\infty} \frac{\Psi_n}{\Psi'_n} < \infty$$

General Reduction Scheme

Instead of directly bounding the expected performance, we are going to prove stronger probability bounds of the form

$$\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) \geq c > 0$$

These bounds can be readily converted to expected performance bounds using Markov's inequality:

$$\mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) \leq \frac{\mathbb{E}_f[d(\widehat{f}_n, f)]}{s_n}$$

Therefore it follows:

$$\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\widehat{f}_n, f)] \geq \inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} s_n \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) \geq c s_n$$

First Reduction Step

Reduce the original problem to an easier one by replacing the larger class $\mathcal{F}$ with a smaller finite class $\{f_0, \ldots, f_M\} \subseteq \mathcal{F}$. Observe that

$$\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) \geq \inf_{\widehat{f}_n} \sup_{f\in\{f_0,\ldots,f_M\}} \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n)$$

The key idea is to choose a finite collection of models such that the resulting problem is as hard as the original, otherwise the lower bound will not be tight.

Second Reduction Step

Next, we reduce the problem to a hypotheses test. Ideally, we would like to have something like

$$\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) \geq \inf_{\widehat{f}_n} \sup_{j\in\{0,\ldots,M\}} \mathcal{P}_{f_j}(\widehat{h}_n(Z) \neq j)$$

The inf is over all measurable test functions

$$\widehat{h}_n : \mathcal{Z} \to \{0, \ldots, M\}$$

and $\mathcal{P}_{f_j}(\widehat{h}_n(Z) \neq j)$ denotes the probability that after observing the data, the test infers the wrong hypothesis.

This might not always be true or easy to show, but in certain scenarios it can be done. Suppose $d(.,.)$ is a semi-distance, i.e. it satisfies

(i) $d(f,g) = d(g,f) \geq 0$ (Symmetric)

(ii) $d(f,f) = 0$

(iii) $d(f,g) \leq d(h,f) + d(h,g)$ (Triangle inequality)

E.g. with $f, g : \mathbb{R}^d \to \mathbb{R}, d(f,g) \triangleq ||f-g||_2$.

Lemma 1. *Suppose $d(.,.)$ is a semi-distance. Also suppose that we have constructed $f_0, \ldots, f_M$ s.t. $d(f_j, f_k) \geq 2s_n$, $\forall j \neq k$. Take any estimator $\widehat{f}_n$ and define the test: $\Psi^* \circ \widehat{f}_n : \mathcal{Z} \to \{0, \ldots, M\}$ as*

$$\Psi^*(\widehat{f}_n) = \arg\min_j d(\widehat{f}_n, f_j)$$

Then $\Psi^*(\widehat{f}_n) \neq j$, implies $d(\widehat{f}_n, f_j) \geq s_n$.

Proof. Suppose $\Psi^*(\widehat{f}_n) \neq j \iff \exists k \neq j : d(\widehat{f}_n, f_k) \leq d(\widehat{f}_n, f_j)$. Now

$$2s_n \leq d(f_j, f_k) \leq d(\widehat{f}_n, f_j) + d(\widehat{f}_n, f_k) \leq 2d(\widehat{f}_n, f_j)$$

$$\implies d(\widehat{f}_n, f_j) \geq s_n$$

■

The previous lemma implies that

$$\mathcal{P}_{f_j}(d(\widehat{f}_n, f_j) \geq s_n) \geq \mathcal{P}_{f_j}(\Psi^*(\widehat{f}_n) \neq j)$$

Therefore,

$$\begin{aligned}
\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathcal{P}_{f_j}(d(\widehat{f}_n, f_j) \geq s_n) &\geq \inf_{\widehat{f}_n} \max_{f \in \{f_0, \ldots, f_M\}} \mathcal{P}_{f_j}(d(\widehat{f}_n, f_j) \geq s_n) \\
&\geq \inf_{\widehat{f}_n} \max_{j \in \{0, \ldots, M\}} \mathcal{P}_{f_j}(\Psi^*(\widehat{f}_n) \neq j) \\
&\geq \inf_{\widehat{h}_n} \max_{j \in \{0, \ldots, M\}} \mathcal{P}_j(\widehat{h}_n \neq j) \\
&\triangleq P_{e,M}
\end{aligned}$$

The third step follows since we are replacing the class of tests defined by $\Psi^*(\widehat{f}_n)$ by a larger class of ALL possible tests $\widehat{h}_n$, and hence the inf taken over the larger class is smaller.

Now our goal throughout is going to be to find lower bounds for $P_{e,M}$.

So we need to construct $f_0, \ldots, f_M$ s.t. $d(f_j, f_k) \geq 2s_n$, $j \neq k$ and $P_{e,M} \geq c > 0$. Observe that this requires careful construction since the first condition necessitates that f_j and f_k are far from each other, while the second condition requires that f_j and f_k are close enough so that it is harder to distinguish them based on a given sample of data, and hence the prob of error $P_{e,M}$ is bounded away from 0.

We now try to lower bound the prob of error $P_{e,M}$. We first consider the case $M = 1$, corresponding to binary hypothesis testing.

M = 1: Let P_0 and P_1 denote the two probability measures, i.e. distributions of the data under models 0 and 1. Clearly if P_0 and P_1 are very "close", then it is hard to distinguish the two hypotheses, and so $P_{e,1}$ is large.

A natural measure between probability measures is the **total variation**, defined as:

$$V(P_0, P_1) = \sup_A |P_0(A) - P_1(A)| = \sup_A |\int_A p_0(Z) - p_1(Z) d\nu(Z)|$$

where p_0 and p_1 are the densities of P_0 and P_1 with respect to a common dominating measure ν and A is any subset of the domain. We will lower bound the prob of error $P_{e,1}$ using the total variation distance. But first, we establish the following lemma.

Scheffe's lemma

$$\begin{aligned} V(P_0, P_1) &= \frac{1}{2}\int |p_0(Z) - p_1(Z)| d\nu(Z) = \frac{1}{2}\int |p_0 - p_1| \\ &= 1 - \int \min(p_0, p_1) \end{aligned}$$

Proof. Recall the definition of the total variation distance:

$$V(P_0, P_1) = \sup_A |\int_A p_0 - p_1|$$

Observe that the set A maximizing the right hand side is given by either $\{Z \in \mathcal{Z} : p_0(Z) \geq p_1(Z)\}$ or $\{Z \in \mathcal{Z} : p_1(Z) \geq p_0(Z)\}$.

Let us pick $A_0 = \{Z \in \mathcal{Z} : p_0(Z) \geq p_1(Z)\}$. Then

$$V(P_0, P_1) = \int_{A_0} p_0 - p_1 = -\int_{A_0^c} p_0 - p_1 = \frac{1}{2}\int |p_0 - p_1|$$

For the second part, notice that

$$p_0(Z) - \min(p_0(Z), p_1(Z)) = \begin{cases} 0 & \text{if } p_0(Z) \leq p_1(Z) \\ p_0(Z) - p_1(Z) & \text{if } p_0(Z) \geq p_1(Z) \end{cases}$$

Now consider

$$1 - \int \min(p_0, p_1) = \int p_0(Z) - \min(p_0(Z), p_1(Z)) = \int_{A_0} p_0(Z) - p_1(Z) d\nu(Z) = V(P_0, P_1)$$

■

We are now ready to tackle the lower bound on $P_{e,1}$. In this case, we consider all tests $\widehat{h}_n(Z) : \mathcal{Z} \to \{0, 1\}$. Equivalently, we can define $\widehat{h}_n(Z) = 1_A(Z)$, where A is any subset of the domain.

$$\begin{aligned} P_{e,1} = \inf_{\widehat{h}_n} \max_{j \in \{0,\ldots,M\}} \mathcal{P}_j(\widehat{h}_n \neq j) &\geq \inf_{\widehat{h}_n} \left(\frac{1}{2} P_0(\widehat{h}_n \neq 0) + P_1(\widehat{h}_n \neq 1) \right) \\ &= \frac{1}{2} \inf_A P_0(1_A(Z) \neq 0) + P_1(1_A(Z) \neq 1) \\ &= \frac{1}{2} \inf_A P_0(A) + P_1(A^c) \\ &= \frac{1}{2} \inf_A 1 - (P_1(A) - P_0(A)) \\ &= \frac{1}{2}(1 - V(P_0, P_1)) \end{aligned}$$

So if P_0 is close to P_1, then $V(P_0, P_1)$ is small and the probability of error $P_{e,1}$ is large.

This is interesting, but unfortunately, it is hard to work with total variation, especially for multivariate distributions. Bounds involving the Kullback-Leibler divergence are much more convenient.

$$K(P_1||P_0) = \int \log \frac{p_1(Z)}{p_0(Z)} p_1(Z) d\nu(Z) = \int \log \frac{p_1}{p_0} p_1$$

The following Lemma relates total variation, affinity and KL divergence.

Lemma 2. $1 - V(P_0, P_1) \geq \frac{1}{2} A^2(P_0, P_1) \geq \frac{1}{2} \exp(-K(P_1||P_0))$

Proof. For the first inequality,

$$\begin{aligned}
A^2(P_0, P_1) &= \left(\int \sqrt{p_0 p_1} \right)^2 \\
&= \left(\int \sqrt{\min(p_0, p_1) \max(p_0, p_1)} \right)^2 \\
&= \left(\int \sqrt{\min(p_0, p_1)} \sqrt{\max(p_0, p_1)} \right)^2 \\
&\leq \int \min(p_0, p_1) \int \max(p_0, p_1) && \text{by Cauchy-Schwarz inequality} \\
&= \int \min(p_0, p_1) \left(2 - \int \min(p_0, p_1) \right) && \because \int \min(p_0,p_1) + \int \max(p_0,p_1) = \int p_0 + \int p_1 = 2 \\
&\leq 2 \int \min(p_0, p_1) \\
&= 2(1 - V(P_0, P_1))
\end{aligned}$$

For the second inequality,

$$\begin{aligned}
A^2(P_0, P_1) &= \left(\int \sqrt{p_0 p_1} \right)^2 \\
&= \exp\left(\log \left(\int \sqrt{p_0 p_1} \right)^2 \right) \\
&= \exp\left(2 \log \left(\int \sqrt{p_0 p_1} \right) \right) \\
&= \exp\left(2 \log \left(\int \sqrt{\frac{p_0}{p_1}} p_1 \right) \right) \\
&\geq \exp\left(2 \int \log \left(\sqrt{\frac{p_0}{p_1}} \right) p_1 \right) && \text{by Jensen's inequality} \\
&= \exp\left(- \int \log \left(\sqrt{\frac{p_1}{p_0}} \right) p_1 \right) \\
&= \exp(-K(P_1||P_0))
\end{aligned}$$

■

Putting everything together, we now have the following Theorem:

Theorem 1. *Let $\mathcal{F}$ be a class of models, and suppose we have observations Z distributed according to $\mathcal{P}_f$, $f \in \mathcal{F}$. Let $d(\widehat{f}_n, f)$ be the performance measure of the estimator $\widehat{f}_n(Z)$ relative to the true model f. Assume also $d(.,.)$ is a semi-distance. Let $f_0, f_1 \in \mathcal{F}$ be s.t. $d(f_0, f_1) \geq 2s_n$. Then*

$$\begin{aligned}\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) &\geq \inf_{\widehat{f}_n} \max_{j\in\{0,1\}} \mathcal{P}_{f_j}(d(\widehat{f}_n, f_j) \geq s_n) \\ &\geq \frac{1}{4}\exp(-K(P_{f_1}||P_{f_0}))\end{aligned}$$

How do we use this theorem?
Choose f_0, f_1 such that $K(P_1||P_0) \leq \alpha$, then $P_{e,1}$ is bounded away from 0 and we get a bound

$$\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathcal{P}_f(d(\widehat{f}_n, f) \geq s_n) \geq c > 0$$

or, after Markov's

$$\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\widehat{f}_n, f)] \geq cs_n$$

To apply the theorem, we need to design f_0, f_1 s.t. $d(f_0, f_1) \geq 2s_n$ and $\exp(-K(P_{f_1}||P_{f_0})) > 0$. To reiterate, the design of f_0, f_1 requires careful construction so as to balance the tradeoff between the first condition which requires f_0, f_1 to be far apart, and the second condition which requires f_0, f_1 to be close to each other.

Let's use this theorem in a problem we are familiar with. Let $X \in [0,1]$ and $Y|X = x \sim$ Bernoulli$(\eta(x))$, where $\eta(x) = P(Y = 1|X = x)$. Suppose $G^* = [t^*, 1]$. Under these assumptions and an upper bound on the density of X, the Chernoff bounding technique yields an expected error rate for ERM of

$$\mathbb{E}[R(\widehat{G}_n) - R^*] = O\left(\sqrt{\frac{\log n}{n}}\right)$$

Is this the best possible rate?

Construct two models in the above class (denote it by $\mathcal{P}$), $P_{XY}^{(0)}$ and $P_{XY}^{(1)}$. For both take $P_X \sim$ Uniform$([0,1])$ and $\eta_{(0)} = 1/2 - a$, $\eta_{(1)} = 1/2 + a$ $(a > 0)$, so $G_0^* = \emptyset$, $G_1^* = [0,1]$.

We are interested in controlling the excess risk

$$R(\widehat{G}_n) - R(G^*) = \int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| dP_X(x)$$

Note that if the true underlying model is either $P_{XY}^{(0)}$ or $P_{XY}^{(1)}$, we have:

$$R_j(\widehat{G}_n) - R_j(G_j^*) = \int_{\widehat{G}_n \Delta G_j^*} |2\eta_j(x) - 1| dx = 2a \int_{\widehat{G}_n \Delta G_j^*} dx = 2a d_\Delta(\widehat{G}_n, G_j^*)$$

Proposition 1. *$d_\Delta(.,.)$ is a semi-distance.*

Proof. It suffices to show that $d(G_1, G_2) = d(G_2, G_1) \geq 0$, $d(G, G) = 0$ $\forall G$ and $d(G_1, G_2) \leq d(G_1, G_3) + d(G_3, G_2)$. The first two statements are obvious. The last one (triangle inequality) follows from the fact that $G_1 \Delta G_2 \subseteq (G_1 \Delta G_3) \cup (G_3 \Delta G_2)$.

Suppose this was not the case, then $\exists x : x \in G_1 \Delta G_2$ s.t. $x \notin G_1 \Delta G_3$ and $x \notin G_2 \Delta G_3$. In other words,

$$x \in (G_1 \Delta G_2) \cap (G_1 \Delta G_3)^c \cap (G_2 \Delta G_3)^c$$

Since $S \Delta T = (S \cap T^c) \cup (S^c \cap T)$, we have:

$$\begin{aligned} x &\in [(G_1 \cap G_2^c) \cup (G_1^c \cap G_2)] \cap [(G_1^c \cup G_3) \cap (G_1 \cup G_3^c)] \cap [(G_2^c \cup G_3) \cap (G_2 \cup G_3^c)] \\ &\in [G_1 \cap (G_1^c \cup G_3) \cap G_2^c \cap (G_2 \cup G_3^c)] \cup [G_1^c \cap (G_1 \cup G_3^c) \cap G_2 \cap (G_2^c \cup G_3)] \\ &\in [G_1 \cap G_3 \cap G_2 \cap G_3^c] \cup [G_1^c \cap G_3^c \cap G_2 \cap G_3] \\ &\in \emptyset, \text{ a contradiction} \end{aligned}$$

■

Lets look at the first reduction step:

$$\begin{aligned}\inf_{\widehat{G}_n}\sup_{p\in\mathcal{P}} P(R(\widehat{G}_n)-R(G^*)\geq s_n) &\geq \inf_{\widehat{G}_n}\max_{j\in\{0,1\}} P_j(R_j(\widehat{G}_n)-R_j(G_j^*)\geq s_n)\\ &= \inf_{\widehat{G}_n}\max_{j\in\{0,1\}} P_j(d_\Delta(\widehat{G}_n,G_j^*)\geq s_n/2a)\end{aligned}$$

So we can work out a bound on d_Δ and then translate it to excess risk.

Lets apply Theorem 1. Note that $d_\Delta(G_0^*,G_1^*)=1$ and let $P_0 \triangleq P^{(0)}_{X_1,Y_1,\ldots,X_n,Y_n}$ and $P_1 \triangleq P^{(1)}_{X_1,Y_1,\ldots,X_n,Y_n}$.

$$\begin{aligned}K(P_1||P_0) &= \mathbb{E}_1\left[\log\frac{p^{(1)}_{X_1,Y_1,\ldots,X_n,Y_n}(X_1,Y_1,\ldots,X_n,Y_n)}{p^{(0)}_{X_1,Y_1,\ldots,X_n,Y_n}(X_1,Y_1,\ldots,X_n,Y_n)}\right]\\ &= \mathbb{E}_1\left[\log\frac{p^{(1)}_{X_1,Y_1}(X_1,Y_1)\ldots p^{(1)}_{X_n,Y_n}(X_n,Y_n)}{P^{(0)}_{X_1,Y_1}(X_1,Y_1)\ldots p^{(0)}_{X_n,Y_n}(X_n,Y_n)}\right]\\ &= \sum_{i=1}^n \mathbb{E}_1\left[\log\frac{p^{(1)}_{X_i,Y_i}(X_i,Y_i)}{p^{(0)}_{X_i,Y_i}(X_i,Y_i)}\right]\\ &= n\mathbb{E}_1\left[\log\frac{p^{(1)}_{Y|X}(Y_1|X_1)}{p^{(0)}_{Y|X}(Y_1|X_1)}\right]\end{aligned}$$

Now $p^{(1)}_{Y|X}(Y_1=1|X_1)=1/2+a$ and $p^{(0)}_{Y|X}(Y_1=1|X_1)=1/2-a$. Also under model 1, $Y_1\sim$ Bernoulli$(1/2+a)$. So we get:

$$\begin{aligned}K(P_1||P_0) &= n\left[(1/2+a)\log\frac{1/2+a}{1/2-a}+(1/2-a)\log\frac{1/2-a}{1/2+a}\right]\\ &= n\left[2a\log(1/2+a)-2a\log(1/2-a)\right]\\ &= 2na\log\frac{1/2+a}{1/2-a}\\ &\leq 2na\left(\frac{1/2+a}{1/2-a}-1\right)\\ &= 4na^2\frac{1}{1/2-a}\end{aligned}$$

Let $a=1/\sqrt{n}$ and $n\geq 16$, then $K(P_1||P_0)\leq 4n\frac{1}{n}\frac{1}{1/2-1/\sqrt{n}}\leq 16$.

Using Theorem 1, since $d_\Delta(G_0^*,G_1^*)=1$, we get:

$$\inf_{\widehat{G}_n}\max_j P_j(d_\Delta(\widehat{G}_n,G_j^*)\geq 1/2)\geq\frac{1}{4}e^{-16}$$

Taking $s_n=1/\sqrt{n}$, this implies

$$\inf_{\widehat{G}_n}\sup_{p\in\mathcal{P}} P(R(\widehat{G}_n)-R(G^*)\geq 1/\sqrt{n})\geq\frac{1}{4}e^{-16}$$

or, after Markov's inequality

$$\inf_{\widehat{G}_n}\sup_{p\in\mathcal{P}}\mathbb{E}[R(\widehat{G}_n)-R(G^*)]\geq\frac{1}{4}e^{-16}\frac{1}{\sqrt{n}}$$

Therefore, apart from the $\log n$ factor, ERM is getting the best possible performance.

Reducing the initial problem to a binary hypothesis testing does not always work. Sometimes we need M hypotheses, with $M\to\infty$ as $n\to\infty$. If this is the case, we have the following theorem:

Theorem 2. *Let $M \geq 2$. $\{f_0, \ldots, f_M\} \in \mathcal{F}$ be such that*

- *$d(f_j, f_k) \geq 2s_n$, where d is a semi-distance.*
- *$\frac{1}{M}\sum_{j=1}^{M} K(P_j||P_0) \leq \alpha \log M$, with $0 < \alpha < 1/8$.*

Then

$$\begin{aligned}\inf_{\widehat{f}_n} \sup_{f\in\mathcal{F}} P_f(d(\widehat{f}_n, f) \geq s_n) &\geq \inf_{\widehat{f}_n} \max_j P_j(d(\widehat{f}_n, f_j) \geq s_n) \\ &\geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log M}}\right) > 0\end{aligned}$$

We will use this theorem to show that the estimator of Lecture 4 is optimal. Recall the setup of Lecture 4. Let

$$\mathcal{F} = \{f : |f(t) - f(s)| \leq L|t - s| \forall t, s\}$$

i.e. class of Lipschitz functions with constant L. Let

$$x_i = i/n, \quad i = 1, \ldots, n$$

$$Y_i = f(x_i) + W_i$$

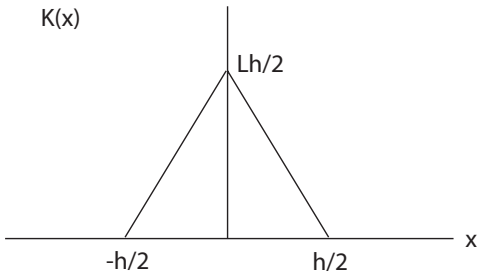
$\mathbb{E}[W_i] = 0, \mathbb{E}[W_i^2] = \sigma^2 < \infty, W_i, W_j$ are indepedent if $i \neq j$. In that lecture, we constructed an estimator $\widehat{f}_n$ such that

$$\sup_{f\in\mathcal{F}} \mathbb{E}[||\widehat{f}_n - f||^2] = O(n^{-2/3})$$

Is this the best we can do?

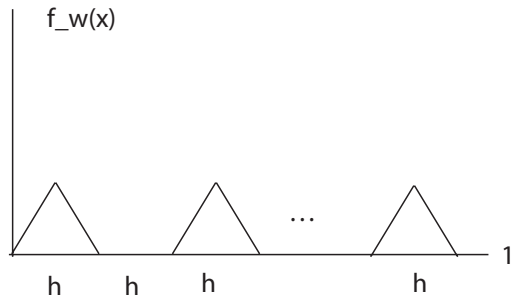
We are going to construct a collection $f_0, \ldots, f_M \in \mathcal{F}$ and apply Theorem 2. Notice that the metric of interest is $d(\widehat{f}_n, f) = ||\widehat{f}_n - f||$, a semi-distance. Let $W_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Let $m \in \mathbb{N}$, $h = 1/m$ and define

$$K(x) = \left(\frac{Lh}{2} - L|x|\right)\mathbb{I}_{|x|\leq h/2} = \frac{L}{2}|h - 2x|\mathbb{I}_{|x|\leq h/2}$$



Note that $|K(a) - K(b)| \leq L|a - b|$, $\forall a, b$. The subclass we are going to consider are functions of the form i.e. "bump" functions. Let $\Omega = \{0, 1\}^m$ be the collection of binary vectors of length m, e.g. $w = (1, 0, 1, \ldots, 0) \in \Omega$. Define

$$f_w(x) = \sum_{i=1}^{m} w_i K\left(x - \frac{h}{2}(2i - 1)\right)$$



Note that for $w, w' \in \Omega$,

$$\begin{aligned} d(f_w, f_{w'}) = ||f_w - f_{w'}|| &= \left(\int_0^1 \sum_{i=1}^m (w_i - w_i')^2 K^2 \left(x - \frac{h}{2}(2i-1) \right) \right)^{1/2} \\ &= \sqrt{\rho(w, w')} \sqrt{\int K^2(x) dx} \end{aligned}$$

where $\rho(w, w')$ is the Hamming distance, $\rho(w, w') = \sum_{i=1}^m |w_i - w_i'|^2 = \sum_{i=1}^m |w_i - w_i'|$. Now

$$\int K^2(x) = 2 \int_0^{h/2} L^2 x^2 dx = 2L^2 \frac{h^3}{3 \cdot 8} = \frac{L^2}{12} h^3$$

so

$$d(f_w, f_{w'}) = \sqrt{\rho(w, w')} \frac{L}{\sqrt{12}} h^{3/2}$$

Since $|\Omega| = 2^n$, the number of functions in our class is 2^n. Turns out, we do not need to consider all functions $f_w, w \in \Omega$, but only a select few. Using all the functions leads to a looser lower bound of the form n^{-1}, which corresponds to the parametric rate. The problem under consideration is non-parametric, and hence we expect a slower rate of convergence. To get a tighter lower bound, the following result is of use:

Lemma 3. *(Varshamov-Gilbert '62)*
Let $m \geq 8$. There exists a subset $\{w^{(0)}, \ldots, w^{(M)}\}$ of Ω such that $w^{(0)} = (0, 0, \ldots, 0)$,

$$\rho(w^{(j)}, w^{(k)}) \geq \frac{m}{8}, \quad \forall 0 \leq j < k \leq M \text{ and } M \geq 2^{m/8}.$$

What this lemma says is that there are many ($\sim 2^m$) sequences in Ω that are very different (i.e. $\rho(w^{(j)}, w^{(k)}) \sim m$). We are going to use the lemma to construct a useful set of hypotheses. Let $\{w^{(0)}, \ldots, w^{(M)}\}$ be the class of sequences in the lemma and define

$$f_j \triangleq f_{w^{(j)}}, \quad j \in \{0, \ldots, M\}$$

We now need to look at the conditions of Theorem 2 and choose m appropriately.

First note that for $j \neq k$,

$$d(f_j, f_k) = \sqrt{\rho(w^{(j)}, w^{(k)})} \frac{L}{\sqrt{12}} h^{3/2} \geq \sqrt{\frac{m}{8}} \frac{L}{\sqrt{12}} m^{-3/2} = \frac{L}{4\sqrt{6}} m^{-1}$$

Now let $P_j \triangleq P^{(j)}_{Y_1,\ldots,Y_m}, j \in \{0,\ldots,M\}$. Then

$$\begin{aligned} K(P_j||P_0) &= \mathbb{E}_j\left[\log \frac{p^{(j)}_{Y_1,\ldots,Y_m}}{p^{(0)}_{Y_1,\ldots,Y_m}}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}_j\left[\log \frac{p^{(j)Y_i}}{p^{(0)}_{Y_i}}\right] = \frac{1}{2\sigma^2}\sum_{i=1}^{n} f_j^2(x_i) \\ &\leq \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{Lh}{2}\right)^2 = \frac{L^2}{8\sigma^2}nh^2 = \frac{L^2}{8\sigma^2}nm^{-2} \end{aligned}$$

Now notice that $\log M \geq \frac{m}{8}\log 2$ (from Lemma 3). We want to choose m such that

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j||P_0) \leq \frac{L^2}{8\sigma^2}nm^{-2} < \alpha\frac{m}{8}\log 2 \leq \alpha \log M$$

This gives

$$m > \left(\frac{L^2}{\alpha\sigma^2 \log 2}\right)^{1/3} n^{1/3} := C_0 n^{1/3}$$

so take $m = \lfloor C_0 n^{1/3} + 1 \rfloor$. Now

$$d(f_j, f_k) \geq \frac{L}{4\sqrt{6}}m^{-1} \geq 2\text{const } n^{-1/3} \qquad \text{for } n \geq n_0(\text{const})$$

Therefore,

$$\inf_{\widehat{f}_n}\sup_{f\in\mathcal{F}} P_f(||\widehat{f}_n - f|| \geq \text{const } n^{-1/3}) \geq c > 0$$

or,

$$\inf_{\widehat{f}_n}\sup_{f\in\mathcal{F}} P_f(||\widehat{f}_n - f||^2 \geq \text{const } n^{-2/3}) \geq c > 0$$

or after Markov's inequality,

$$\inf_{\widehat{f}_n}\sup_{f\in\mathcal{F}} \mathbb{E}_f[||\widehat{f}_n - f||^2] \geq c \cdot \text{const } n^{-2/3}$$

Therefore, the estimator constructed in class attains the optimal rate of convergence.