Chapter 8: Nonparametric Estimators

May 20, 2022

- Examples of nonparametric models and problems:
- Example (1): Estimation of a probability density. Let $(X_1, ..., X_n)$ be i.i.d real valued random variables whose common distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$.
- The density of this distribution, denoted by p is a funtion from $\mathbb{R}$ to $[0, \infty)$ is the target of inference.
- An estimator of p is a function $x \to p_n(x) = p_n(x, X_1, ..., X_n)$ measureable with respect to the observations $\mathbf{X} = (X_1, ..., X_n)$.

- In nonparametrics statistics, one typically assumes that p belongs to some massive class $\mathcal{P}$ of densities
- For example, $\mathcal{P}$ can be the set of all the continuous probabilities on $\mathbb{R}$.
- or $\mathcal{P}$ can be the set of Lipschitz continuous probability densities on $\mathbb{R}$.
- Classe of such type will be called nonparametrix classes of functions.

- Example (2): Assume that we have n independent pairs of RVs $(X_1, Y_1), ..., (X_n, Y_n)$ such that

$$\begin{aligned} Y_i &= f(X_i) + \varepsilon_i, X_i \in [0, 1], \\ \mathbb{E}(\varepsilon_i | X_i) &= 0 \end{aligned}$$

  where the function f from $[0, 1]$ to $\mathbb{R}$ (called the regression function) is unknown.
- The problem of nonparametric regression is to estimate f given a priori that this function belongs to a nonparametric class of functions $\mathcal{F}$
- For example $\mathcal{F}$ can be the set of all continuous functions on $[0, 1]$.
- or the set of all convex functions on $[0, 1]$.
- An estimator of f is a function $x \to f_n(x) = f_n(x, \mathbf{X})$ defined on $[0, 1]$ and measureable with respect to the observations $\mathbf{X} = (X_1, ..., X_n, Y_1, ..., Y_n)$. In what follows, we will mainly focus on the particular case where $X_i = i/n$.

- **Is density pathwise differentiable?**
- Suppose one would like to use semiparametric theory to develop an estimator of a density evaluated at a point $p(x_0)$, in the nonparametric model where no restriction is a priori imposed on the density except certain regularity conditions such as continuity.
- Is $p(x_0)$ pathwise differentiable? If so what is the corresponding efficient gradient and influence function.
- Consider a regular submodel $p_\theta\left(x_0\right)$, then we seek a function $\delta_{x_0}(X)$ in L_2 such that

$$\frac{dp_\theta\left(x_0\right)}{d\theta} = \mathbb{E}\left\{\delta_{x_0}(X)S_\theta\left(X\right)\right\}$$

where $S_\theta\left(X\right)$ is the score of θ at zero.

- This is equivalent to

$$\frac{dp_\theta\left(x_0\right)}{d\theta} = \int \delta_{x_0}(x)\frac{dp_\theta\left(x\right)}{d\theta}dx$$

for all continuous functions $\frac{dp_\theta(x)}{d\theta}$.

- **Is density pathwise differentiable?**
- The only "function" $\delta_{x_0}(x)$ known to satisfy such an equation is the Dirac delta function at x_0.
- It is actually not a function per se but a measure which satisfies

$$f(0) = \int \delta_0(x)f(x)dx$$

for all continuous functions with compact support. Informally such a measure can be described as a function

$$\delta_0(x) = \begin{cases} \infty \text{ if } x = 0 \\ 0 \text{ otherwise} \end{cases}$$

which also satisfies the restriction

$$1 = \int \delta_0(x)dx$$

- **Is density pathwise differentiable?**
- While the Dirac delta at x_0 satisfies

$$\frac{dp_\theta\left(x_0\right)}{d\theta} = \int \delta_{x_0}(x)\frac{dp_\theta\left(x\right)}{d\theta}dx$$

it is not in L_2.

- To show this, note that if it were by the Cauchy-Schwartz inequality, then

$$f(0) = \int \delta_0(x)f(x)dx \leq C||f||_2$$

for some C. it suffices to construct a sequence $f_n(x)$ such that $||f_n||_2 \to 0$ but $f_n(0) \to \infty$ as $n \to \infty$, contradicting the inequality. For example you may check that $f_n(x) = \sqrt{n}\exp\left(-nx^2\right)$ has this property.

- We conclude that $p(x_0)$ is not pathwise differentiable, and therefore may not be $\sqrt{n}$ estimable.
- The idea behind kernel smoothing is to replace $\delta_0(x)$ by a "kernel" function in L_2 which mimics the behavior of the Dirac delta function but without blowing up the variance.

- **Kernel density estimation**
- Let $X_1, ..., X_b$ be independent identically distributed (i.i.d) random variables that have a probability density p wrt Lebesgue measure on $\mathbb{R}$.
- The corresponding distribution function is $F(x) = \int_{-\infty}^{x} p\left(t\right)dt$. Consider the empirical distribution function

$$F_n\left(x\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i \leq x\right)$$

- By the strong long of large numbers, we have $F_n\left(x\right) \to F\left(x\right)$ for all $x \in \mathbb{R}$ almost surely as $n \to \infty$.
- How can we estimate p?

- One of the first intuitive solutions is based on the following argument. For sufficiently small $h > 0$ we can write an approximation

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

- Replacing F by its empirical version, we define

$$\widehat{p}_n^R(x) \approx \frac{F_n(x+h) - F_n(x-h)}{2h}$$

- $\widehat{p}_n^R(x)$ is an estimator of p called the *Rosenblatt estimator*, which is equivalenty written

$$\begin{aligned} \widehat{p}_n^R(x) &= \frac{1}{2nh}\sum_{i=1}^{n} I(x-h < X_i \leq x+h) \\ &= \frac{1}{2nh}\sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

where $K_0(u) = 1/2I(-1 \leq u \leq 1)$.

- An immediate generalization of Rosenblatt's estimator is given by

$$\widehat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u)\,du = 1$.

- Such a function K is called a kernel and the parameter h is called a bandwidth of the estimator $\widehat{p}_n(x)$
- $\widehat{p}_n(x)$ is called the kerrnel density estimator or the *Parzen-Rosenblatt estimator*.
- In the asymptotic framework, as $n \to \infty$, we will consider the bandwidth h_n indexed by n, such that $h_n \to 0$ along a sequence as $n \to \infty$.

- **Classical examples of kernels**
  - The rectangular kernel $K(u) = 1/2I(|u| \leq 1)$
  - The triangular kernel $K(u) = (1 - |u|)I(|u| < 1)$
  - The parabolic or Epanechnikov kernel:$K(u) = 3/4(1-u^2)I(|u| \leq 1)$
  - Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$.
- Note that if the kernel K takes only nonnegative values, then conditional on $X_1, ..., X_n$, $\widehat{p}_n(x)$ is a probability density
- Further note that the Parzen-Rosenblatt estimator is easily generalized to the multidimensional case

$$\widehat{p}_n(x, z) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Z_i - z}{h}\right)$$

- **Mean squared error of kernel estimators**
- A basic measure of the accuracy of $\widehat{p}_n(x)$ is its mean squared risk at an arbitrary point x_0

$$\begin{aligned} MSE(x_0) &= \mathbb{E}\left\{[\widehat{p}_n(x_0) - p(x_0)]^2\right\} \\ &= \left\{[\mathbb{E}(\widehat{p}_n(x_0)) - p(x_0)]^2\right\} \\ &\quad +\mathbb{E}\left\{[\widehat{p}_n(x_0) - \mathbb{E}(\widehat{p}_n(x_0))]^2\right\} \\ &= b(x_0)^2 + \sigma^2(x_0) \end{aligned}$$

- $b(x_0) = \mathbb{E}(\widehat{p}_n(x_0)) - p(x_0)$ is the bias of $\widehat{p}_n(x_0)$ while

$$\sigma^2(x_0) = \mathbb{E}\left\{[\widehat{p}_n(x_0) - \mathbb{E}(\widehat{p}_n(x_0))]^2\right\}$$

is its variance.

- To evaluate the MSE requires a separate analysis of these terms.

- **Variance of the estimator**
  Result 8.1:Suppose that the density p satisfies $p(x) \leq p_{\max} < \infty$ for all $x \in \mathbb{R}$. Let K be a function such that

$$\int K^2(u)\, du < \infty.$$

  Then for any $x_0 \in \mathbb{R}$, $h > 0$ and $n \geq 1$ we have that

$$\sigma^2(x_0) \leq \frac{C_1}{nh}$$

  where $C_1 = p_{\max} \int K^2(u)\, du$.

- Proof of Result 8.1.
  Put

$$\eta_i(x_0) = K\left(\frac{X_i - x_0}{h}\right) - \mathbb{E}\left[K\left(\frac{X_i - x_0}{h}\right)\right]$$

  The random variables $\eta_i(x_0), i = 1, ..., n$ are i.i.d with mean and variance

$$\begin{aligned}
\mathbb{E}\left[\eta_i(x_0)^2\right] &\leq \mathbb{E}\left[K^2\left(\frac{X_i - x_0}{h}\right)\right] \\
&= \int K^2\left(\frac{z - x_0}{h}\right) p(z)dz \\
&\leq p_{\max} h \int K^2(u)\, du
\end{aligned}$$

- Then

$$\begin{aligned}
\sigma^2(x_0) &= \mathbb{E}\left[\left(\frac{1}{nh}\sum_{i=1}^{n} \eta_i(x_0)\right)^2\right] \\
&= \frac{1}{nh^2}\mathbb{E}\left[\eta_i(x_0)^2\right] \leq \frac{C_1}{nh}.
\end{aligned}$$

  We conclude that if the bandwidth $h = h_n$ is such that $nh \to \infty$ as $n \to \infty$, then the variance $\sigma^2(x_0)$ goes to zero as $n \to \infty$. This also implies that $\sigma^2(x_0)$ converges to zero at a rate $1/nh$ which will typically be much slower than the parametric rate $1/n$.

- **Bias of the kernel estimator:** Recall that the bias of the kernel density estimator has the form:

$$\begin{aligned}
b(x_0) &= \mathbb{E}\left(\widehat{p}_n(x_0)\right) - p(x_0) \\
&= \frac{1}{h}\int K\left(\frac{z - x_0}{h}\right) p(z)dz - p(x_0)
\end{aligned}$$

  We now analyze the behavior of $b(x_0)$ under some regularity conditions on the density p and on the kernel K.

- Definition 8.1 Let T be an interval in $\mathbb{R}$ and let β and L be two positive numbers. The Hölder class $\Sigma(\beta, L)$ on T is defined as the set of $l = \lfloor \beta \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose derivative $f^{(l)}$ satisfies

$$|f^{(l)}(x) - f^{(l)}(x')| \leq L|x - x'|^{\beta - l}, \text{ for all } x, x' \in T.$$

- Definition 8.2 Let $l \geq 1$ be an integer. We say that K is a kernel of order l if the functions $K(u)\, u^j \quad j = 0, 1, ..., l$,are integrable and satisfy

$$\int K(u)du = 1, \int K(u)u^j du = 0, \; j = 1, ..., l.$$

- Suppose that p belongs to the class of densities $\mathcal{P} = \mathcal{P}(\beta, L)$ defined as followed:

$$\mathcal{P}(\beta, L) = \left\{ p | p \geq 0, \int p(x)dx = 1 \text{ and } p \right\}$$

and assume that K is a kernel of order l. Then

- Result 8.2:Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta |K(u)| du < \infty$$

Then for all $x_0 \in \mathbb{R}$,$h > 0$ and $n \geq 1$ we have

$$|b(x_0)| \leq C_2 h^\beta$$

where

$$C_2 = \frac{L}{l!} \int |u|^\beta |K(u)| du$$

Proof of Result 8.2. We have that

$$\begin{aligned} b(x_0) &= \frac{1}{h} \int K\left(\frac{z - x_0}{h}\right) p(z)dz - p(x_0) \\ &= \int K(u) \left[p(x_0 + uh) - p(x_0)\right] du \\ &= \int K(u) \frac{(uh)^l}{l!} \left[p^{(l)}(x_0 + \tau uh) - p^{(l)}(x_0)\right] du \end{aligned}$$

where we use the expansion

$$p(x_0 + uh) = p(x_0) + p^{(1)}(x_0)uh + \ldots + p^{(l)}(x_0 + \tau uh)\frac{(uh)^l}{l!}$$

for $0 \leq \tau \leq 1$ and we further use the fact that K is of order l. We conclude that

$$\begin{aligned} b(x_0) &\leq \int |K(u)| \frac{|uh|^l}{l!} \left|p^{(l)}(x_0 + \tau uh) - p^{(l)}(x_0)\right| du \\ &\leq L \int |K(u)| \frac{|uh|^l}{l!} |\tau uh|^{\beta - l} du \leq C_2 h^\beta. \end{aligned}$$

- **Upper bound on the mean squared error**
- From Results 8.1 and 8.2 we see that the upper bound on the bias and variance behave in opposite ways as the h varies. The variance decreases as h grows, whereas the bound on the bias increases. The choice of small h corresponding to a large variance is called *undersmoothing*.
- Alternatively, with a large h the bias cannot be reasonably controlled which leads to *oversmoothing*.
- An optimal value of h that balances bias and variance is located between these two extremes.
- if p and K satisfy the assumptions of Results 1.1. and 1.2 we obtain

$$MSE \leq C_2^2 h^{2\beta} + \frac{C_1}{nh}$$

- The minimum with respect to h of the right hand side is attained at

$$h_n^* = \left(\frac{C_1}{2\beta C_2^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$

Therefore, the choice $h = h_n^*$ gives

$$MSE(x_0) = O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$$

- Result 8.3: Assume that $\|K\|_{2,\mu} < \infty$ and the assumptions of Result 8.2 are satisfied. Fix $\alpha > 0$ and take $h = \alpha n^{-\frac{1}{2\beta+1}}$. then for $n \geq 1$ the kernel estimator $\widehat{p}_n$ satisfies

$$\sup_{x_0 \in \mathbb{R}} \sup_{p \in \mathcal{P}(\beta, L)} \mathbb{E}_p \left\{ \left[\widehat{p}_n(x_0) - p(x_0)\right]^2 \right\} \leq C n^{-\frac{2\beta}{2\beta+1}}$$

where $C > 0$ is a constant depending on β, L, α and on the kernel K.

- Proof: The proof relies on an application of Result 8.1 which requires proving that there exists a constant $p_{\max} < \infty$ satisfying

$$\sup_{x_0 \in \mathbb{R}} \sup_{p \in \mathcal{P}(\beta, L)} p(x) \leq p_{\max}$$

that is we need to show that functions in the Hölder ball $\mathcal{P}(\beta, L)$ are bounded by a universal constant $p_{\max}$. We will show this in a homework. Then the proof follows from the upper bound of the MSE.

- Under the assumptions of Result 8.3, the rate of convergence of the estimator $p_n(x_0)$ is $\psi_n = n^{-\frac{\beta}{2\beta+1}}$, meaning that for a finite constant C and for all $n \geq 1$ we have that

$$\sup_{p \in \mathcal{P}(\beta,L)} \mathbb{E}_p \left\{ \left[\widehat{p}_n(x_0) - p(x_0)\right]^2 \right\} \leq C\psi_n^2$$

- Two questions arise:
  - Can we improve the rate ψ_n by using other density estimators?
  - what is the best possible rate of convergence
- In order to answer these questions, it is useful to consider the minimax risk R_n^* associated to the class $\mathcal{P}(\beta, L)$.

$$R_n^*\left(\mathcal{P}(\beta, L)\right) \triangleq \inf_{T_n} \sup_{p \in \mathcal{P}(\beta,L)} \mathbb{E}_p \left\{ \left[\widehat{p}_n(x_0) - p(x_0)\right]^2 \right\}$$

where the infinum is over all estimators .

- We will establish later that a lower bound on the minimax risk is given by

$$R_n^*\left(\mathcal{P}(\beta, L)\right) \geq C'\psi_n^2$$

with some constant $C' > 0$.

- This implies that under the assumptions of result 8.3, the kernel estimator attains the optimal rate of convergence ψ_n associated with the class of densities (β, L).

- Remark on positivity constraint:
- It follows easily from Definition 8.2 that kernels of order $l \geq 2$ must take negative values on a set of positive Lebesgue measure. The estimators $\widehat{p}_n$ based on such kernels can also take negative values
- This property is sometimes emphasized as a drawback of estimators iwth higher order kernels since the density p itself is nonnegative.
- However, this remark is of minor relevance because one can always use the positive part estimator

$$\widehat{p}_n^+ \triangleq \max\{0, \widehat{p}_n\}$$

whose risk is smaller than or equal to the risk of $\widehat{p}_n$.

$$\mathbb{E}_p \left\{ \left[\widehat{p}_n^+(x_0) - p(x_0)\right]^2 \right\} \leq \mathbb{E}_p \left\{ \left[\widehat{p}_n(x_0) - p(x_0)\right]^2 \right\} \text{ for all } x_0 \in \mathbb{R}$$

- In particular, Result 8.3 remains valid if we replace there $\widehat{p}_n(x_0)$ by $\widehat{p}_n^+$. Thus, the estimator $\widehat{p}_n^+$ is nonnegative and attains fast convergence rates associated with higher order kernels.

- **Integrated squared risk of kernel estimator**:
- We have previously studied the behavior of the kernel estimator at a given point x_0. Next we analyze its global risk.
- We will consider the *mean integrated squared error (MISE):*

$$MISE = \mathbb{E} \int \left(\widehat{p}_n(x) - p(x)\right)^2 dx$$

It is straightforward to show that

$$MISE = \int MSE(x)dx = \int b^2(x)dx + \int \sigma^2(x)\, dx$$

- Therefore we can obtain an upper bound for MISE by bounding the integrated pointwise squared bias and variance previously obtained.

- Result 8.4: Suppose that the kernel function K satisfying

$$\int K^2(u)\,du < \infty.$$

Then for any $h > 0, n \geq 1$ and any probability density p we have that

$$\int \sigma^2(x)dx \leq \frac{1}{nh}\int K^2(u)du$$

Proof: As in the proof Result 8.1 we obtain

$$\begin{aligned}
\int \sigma^2(x)dx &= \frac{1}{nh^2}\int \mathbb{E}\left[\eta_i(x)^2\right] \\
&\leq \frac{1}{nh^2}\int\left[\int K^2\left(\frac{z-x}{h}\right)p(z)dz\right]dx \\
&= \frac{1}{nh^2}\int p(z)\left[\int K^2\left(\frac{z-x}{h}\right)dx\right]dz \\
&= \frac{1}{nh}\int K^2(u)\,du.
\end{aligned}$$

- The variance bound does not require any condition on p. The result holds for any density. For the bias term the situation is quite different.
- We can only control it on a restricted subset of densities with sufficient smoothness.
- Because the MISE is a risk corresponding to the L_2-norm, it is natural to assume that p is smooth wrt this norm.
- Definition 8.3 (Nikol'ski):Let $\beta > 0$ and $L > 0$. The *Nikol'ski class* $\mathcal{H}(\beta, L)$ is defined as the set of functions $f : \mathbb{R} \to \mathbb{R}$ whose derivatives $f^{(l)}$ of order $l = \lfloor \beta \rfloor$ exist and satisfy

$$\left[\int \left(f^{(l)}(x+t) - f^{(l)}(x)\right)^2 dx\right]^{1/2} \leq L|t|^{\beta - l} \text{ for all } t \in \mathbb{R}$$

- Definition 8.4(Sobolev) :Let $\beta > 0$ be an integer and $L > 0$. The *Sobolev class* $\mathcal{S}(\beta, L)$ is defined as the set of functions $f : \mathbb{R} \to \mathbb{R}$ that have $\beta - 1$ absolutely continuous derivatives and satisfies

$$\int \left(f^{(\beta)}(x)\right)^2 dx \leq L^2$$

- We will now give an upper bound on the bias term $\int b^2(x)dx$ when p belongs to the class of probability densities that are smooth in the sense of Nikol'ski.

$$\mathcal{P}_{\mathcal{H}}(\beta, L) = \left\{p \in \mathcal{H}(\beta, L)\,|p > 0 \text{ and } \int p(x)dx = 1\right\}$$

- The bound also applies to the corresponding Sobolev class.

- Result 8.5: Assume that $p \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ and let K be a kernel of order $l = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta\, |K(u)|\, du < \infty.$$

Then, for any $h > 0$ and $n \geq 1$,

$$\int b^2(x)\,dx \leq C_2^2 h^{2\beta}$$

where

$$C_2 = \frac{L}{l!}\int |u|^\beta\, |K(u)|\, du.$$

- To prove the result, we will need the following well-known Lemma (proof will be given in Lab)
- Lemma 8.1(Generalized Minkowski inequality) :For any (Borel) function g on $\mathbb{R} \times \mathbb{R}$, we have

$$\int \left(\int g(u,x)\,du\right)^2 dx \leq \left[\int \left(\int g^2(u,x)\,dx\right)^{1/2} du\right]^2$$

- Proof of Result 8.5: Take any x, $u, h > 0$ and write the Taylor expansion

$$p(x+hu) = p(x) + p'(x)uh + ... + \frac{(uh)^l}{(l-1)!}\int_0^1 (1-\tau)^{l-1}\, p^{(l)}(x+\tau uh)d\tau.$$

Since the kernel K is of order $l = \lfloor \beta \rfloor$, we obtain

$$\begin{aligned} b(x) &= \int K(u)\frac{(uh)^l}{(l-1)!}\left[\int_0^1 (1-\tau)^{l-1}\, p^{(l)}(x+\tau uh)d\tau\right] du \\ &= \int K(u)\frac{(uh)^l}{(l-1)!}\left[\int_0^1 \left((1-\tau)^{l-1}\, p^{(l)}(x+\tau uh) - p^{(l)}(x)\right) d\tau\right] du \end{aligned}$$

- Applying twice the generalized Minkowski inequality and using the fact that p belongs to the class $\mathcal{H}(\beta, L)$, we get the following upper bound for the bias term

$$\begin{aligned} & \int b(x)^2 dx \\ \leq & \int \left(\int |K(u)| \frac{|uh|^l}{(l-1)!}\left[\int_0^1 (1-\tau)^{l-1} \left|\left(p^{(l)}(x+\tau uh) - p^{(l)}(x)\right)\right| d\tau du\right]\right)^2 dx \\ \leq & \left(\int |K(u)| \frac{|uh|^l}{(l-1)!}\left[\int \left(\int_0^1 (1-\tau)^{l-1}\left|p^{(l)}(x+\tau uh) - p^{(l)}(x)\right| d\tau\right)^2 dx\right]^{1/2} du\right)^2 \\ \leq & \left(\int |K(u)| \frac{|uh|^l}{(l-1)!}\left[\int_0^1 (1-\tau)^{l-1}\left(\int \left(p^{(l)}(x+\tau uh) - p^{(l)}(x)\right)^2 dx\right)^{1/2} d\tau\right] du\right)^2 \\ \leq & \left(\int |K(u)| \frac{|uh|^l}{(l-1)!}\left[\int_0^1 (1-\tau)^{l-1}\, L\,|uh|^{\beta - l}\, d\tau\right] du\right)^2 \\ \leq & C_2^2 h^{2\beta} \end{aligned}$$

- Under the assumptions of Results 8.4 and 8.5 we obtain

$$MISE \leq C_2^2 h^{2\beta} + \frac{1}{nh}\int K^2(u)du$$

and the minimizer $h = h_n^*$ of the right hand-side is

$$h_n^* = \left(\frac{\int K^2(u)du}{2\beta C_2^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$

Taking $h = h_n^*$ we get

$$MISE = O\left(n^{-\frac{2\beta}{2\beta+1}}\right), n \to \infty.$$

We see that the behavior of the MISE is analogous to that of the mean squared risk at a fixed point (MSE).

- We can summarize the above argument in the following way:
- Result 8.6: Suppose that the assumptions of Results 8.4 and 8.5 hold. Fix $\alpha > 0$ and take $h = \alpha n^{-\frac{1}{2\beta+1}}$. Then for any $n \geq 1$ the kernel estimator $\widehat{p}_n$ satisfies

$$\sup_{p \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{E}_p \int \left\{[\widehat{p}_n(x) - p(x)]^2\right\} dx \leq C n^{-\frac{2\beta}{2\beta+1}}$$

- An analogous result can be shown to hold for Sobolev smoothness classes although the argument for the bias bound is slightly different. (See homework)

- **Lack of asymptotic optimality for fixed density:**
- How to choose the kernel K and the bandwidth h for the kernel density estimators in an optimal way?
- An old and still popular approach is based on minimizatiion in K and h of the asymptotic MISE for fixed density p.
- We now argue that this does not lead to a consistent concept of optimality. We state the main result without proof.
- Result 8.7:Assume that (i) the function K is a kernel of order 1 satisfying the conditions

$$\int K^2(u)\,du < \infty, \int u^2 |K(u)|\,du < \infty, S_K = \int K(u)u^2 du \neq 0$$

(ii)the density of p is differentiable on $\mathbb{R}$, the first derivative p' is absolutely continuous on $\mathbb{R}$ and the second derivaive satisfies

$$\int (p''(x))^2 dx < \infty$$

Then for all $n \geq 1$, the mean integrated squared error of the kernel estimator satisfies

$$\begin{aligned} MISE &= \mathbb{E}_p \int \left\{ [\widehat{p}_n(x) - p(x)]^2 \right\} dx \\ & \left[\frac{1}{nh} \int K^2(u)du + \frac{h^4}{4} S_K^2 \int \left(p''(x)\right)^2 dx \right] (1 + o(1)) \end{aligned}$$

where the $o(1)$ term is independent of n (but depends on p) and tends to 0 as $h \to 0$.

- **Lack of asymptotic optimality for fixed density:**
- The main term of the MISE is

$$\left[\frac{1}{nh} \int K^2(u)du + \frac{h^4}{4} S_K^2 \int \left(p''(x)\right)^2 dx \right]$$

Note that if K is a nonnegative kernel, this rate coincides with the nonasymptoti upper bound for the MISE given in Result 8.6 which holds for all n and h when $\beta = 2$.

- The approach we will criticize minimizes this expression in h and in nonnegative h which yields the optimal bandwidth and nonnegative kernel:

$$\begin{aligned} h^{MISE}(K) &= \left(\frac{\int K^2(u)du}{nS_K^2 \int \left(p''(x)\right)^2 dx} \right)^{1/5} \\ K^* &= \frac{3}{4}\left(1 - u^2\right)_+, \\ & \text{therefore} \\ h^{MISE}(K^*) &= \left(\frac{15}{n \int \left(p''(x)\right)^2 dx} \right)^{1/5} \end{aligned}$$

- **Lack of asymptotic optimality for fixed density:**
- Note that the optimal kernel K^* is the Epanechnikov kernel (or parabolic kernel) previously described
- However, the optimal bandwidth $h^{MISE}(K^*)$ is not a feasible one, as it depends on the unknown second derivative of the density.
- The estimator using the optimal kernel and bandwitdh is known as an oracle estimator or the Epanechnikov oracle. The Oracle MISE (the best achievable MISE by this logic) is given by

$$n^{4/5} \lim_{n\to\infty} \mathbb{E}_p \int \left\{ [\widehat{p}_n(x) - p(x)]^2 \right\} dx = \frac{3^{4/5}}{5^{1/5}} \left(\int \left(p''(x)\right)^2 dx \right)^{1/5}$$

- This argument is often exhibited as a benchmark for the optimal choice of kernel K and bandwidth h with the above constant an efficiency bound for kernel estimation, attainable by substituting an estimator of $\left(\int \left(p''(x)\right)^2 dx\right)$ from the observed sample. We now explain why such an approach to optimality is misleading.

- **Lack of asymptotic optimality for fixed density:**
- The main issue is that one can show that for fixed p satisfying the above assumptions

$$\inf_{T_n} \lim \sup_{n\to\infty} n^{4/5} \mathbb{E}_p \int \left\{ [\widehat{p}_n(x) - p(x)]^2 \right\} dx = 0$$

where $\inf_{T_n}$ is the infimum over all the kernels estimators (or all the positive part kernel estimators). That is for a fixed p, there are many kernel estimators with strictly smaller MISE than the oracle estimator!!!

- For example, one can choose a kernel K of second order, such that $S_K = 0, ||K||_2 < \infty$, then for any $\varepsilon > 0$ with bandwidth

$$h = n^{-1/5} \varepsilon^{-1} \int K^2(u)du$$

satisfies

$$\lim \sup_{n\to\infty} n^{4/5} \mathbb{E}_p \int \left\{ [\widehat{p}_n(x) - p(x)]^2 \right\} dx \leq \varepsilon$$

- **Lack of asymptotic optimality for fixed density:**
- Thus this estimator is guaranteed in sufficiently large sample to outperform the oracle estimator and therefore a counter-example of the claimed optimality result.
- However, we do not necessarily recommend this latter estimator for use, the main point is that such an estimator can be obtained which has smaller variance than the oracle estimator (controlled by ε) and strictly smaller bias controlled by the fact that K is second order.
- That is the fact that K is chosen such that $S_K = 0$ eliminates the leading bias term for n large enough.
- This elimination of the main bias term is possible for fixed p since it is equal to $\frac{h^4}{4}\left(\int u^2 K(u)du\right)^2 \int (p''(x))^2\, dx$, but not uniformly over p in $\mathcal{P}_{\mathcal{H}}(\beta = 2, L)$ or even $\mathcal{P}_{\mathcal{S}}(\beta = 2, L)$. This is because, at least in the case of $\mathcal{P}_{\mathcal{H}}(\beta = 2, L)$, the bias can at most be shown to be no larger than $\frac{h^4 L}{4}\left(\int |u|^2 |K(u)|\, du\right)^2$ which cannot be reduced further by choosing a kernel of higher order.

- **Lack of asymptotic optimality for fixed density:**
- To summarize, the approach based on fixed p asymptotics does not lead to a consistent concept of optimality.
- In particular, saying that "the choice of h and K is optimal" does not make much sense.
- This explains why instead of studying the asymptotics for fixed density p, we focus on the uniform bounds on the risk over smoothness classes of densities. We compute the behavior of estimators in a minimax sense over these classes.
- This lead to a valid concept of optimality (among all estimators).

- **Unbiased risk estimation using cross-validation.**
- In this section, the kernel K is fixed, and we are interested in choosing the bandwidth h. Therefore, we wish to find

$$h_{opt} = \arg\min_{h>0} MISE(h)$$

- Unfortunately, this value remains purely theoretical since MISE(h) depends on the unknown p.
- An approach would be to estimate $MISE(h)$ and to minimize an approximately unbiased or unbiased estimator to obtain an estimator of h_{opt}.
- We briefly describe a popular implementation of this idea given by cross-validation

- **Unbiased risk estimation using cross-validation.**
- First, note that

$$MISE(h) = \mathbb{E}_p \int (\widehat{p}_n - p)^2 = \mathbb{E}_p \left[\int \widehat{p}_n^2 - 2\int \widehat{p}_n p\right] + \int p^2$$

we will write for brevity $\int f$ for $\int f(x)dx$.

- Since $\int p^2$ does not depend on h, the minimizer of $MISE(h)$ also minimizes the function

$$J(h) = \mathbb{E}_p \left[\int \widehat{p}_n^2 - 2\int \widehat{p}_n p\right]$$

therefore it suffices to obtain an unbiased estimator of $J(h)$, that is, of $\int \widehat{p}_n p$.

- **Unbiased risk estimation using cross-validation.**
- Consider the estimator of $p(X_i)$

$$\widehat{p}_{n,-i}(X_i) = \frac{1}{(n-1)\,h}\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right)$$

then it is straightforward to verify that an unbiased estimator of $\int \widehat{p}_n p$ is given by

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^{n}\widehat{p}_{n,-i}\left(X_i\right)$$

- Indeed,

$$\begin{aligned}\mathbb{E}\left[\widehat{G}\right] &= \mathbb{E}\left[\widehat{p}_{n,-i}\left(X_i\right)\right] \\ &= \mathbb{E}\left[\frac{1}{(n-1)\,h}\sum_{j\neq i}\int K\left(\frac{X_j - x}{h}\right)p(x)dx\right] \\ &= \frac{1}{h}\int p(z)\int K\left(\frac{z-x}{h}\right)p(x)dxdz\end{aligned}$$

- **Unbiased risk estimation using cross-validation.**
- on the other hand,

$$\begin{aligned}G &= \mathbb{E}\left[\int \widehat{p}_n p\right] \\ &= \mathbb{E}\left[\frac{1}{(n)\,h}\sum_{i=1}^{n}\int K\left(\frac{X_i - x}{h}\right)p(x)dx\right] \\ &= \frac{1}{h}\int p(z)\int K\left(\frac{z-x}{h}\right)p(x)dxdz\end{aligned}$$

summarizing our argument, an unbiased estimator fo $J(h)$ can be written as follows

$$CV(h) = \int \widehat{p}_n^2 - \frac{2}{n}\sum_{i=1}^{n}\widehat{p}_{n,-i}\left(X_i\right)$$

where CV stands for "cross-validation".
- The function CV$(\cdot)$ is called the leave-one-out cross-validation criterion or simply the cross validatiton criterion.

- **Unbiased risk estimation using cross-validation.**
- The cross-validation estimator $\widehat{p}_{n,CV}$ of p is defined ast the kernel estimator

$$\widehat{p}_{n,CV}\left(x\right) = \frac{1}{nh_{CV}}\sum_{i} K\left(\frac{X_i - x}{h_{CV}}\right)$$

where

$$h_{CV} = \arg\min_{h>0} CV\left(h\right)$$

- It can be proved (beyond the scope of this course) that $\widehat{p}_{n,CV}\left(x\right)$ is asymptotically equivalent to that of the oracle estimator !!!
- Such an estimator is known as adaptive wrt to the oracle bandwidth.

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- Nonparametric regression with random design:
- Let (X, Y) be a pair of real valued random variables such that $E\left|Y\right| < \infty$. The function

$$f(x) = \mathbb{E}\left(Y|X = x\right)$$

is called a regression function of Y on X. Suppose that we have an i.i.d sample (X_i, Y_i), $i = 1, ..., n$.
- The goal is to recover inferences about $f(\cdot)$ in a model that makes no assumption about the conditional density of $\varepsilon = Y - f(X)|X$ and also allows the marginal density of X to be unrestricted.

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- Nonparametric regression with fixed design:
- The quantity of interest is still

$$f(x) = \mathbb{E}\left(Y|X = x\right),$$

The conditional density the conditional density of $\varepsilon_i = Y_i - f(X_i)|X_i$ is unrestricted but X_i are now fixed instead of random and i.i.d.

- For example, in the case of regular design $X_i = i/n$. We will mainly focus on this design.

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- Given a kernel K and bandwidth h, the most celebrated kernel estimator of regression function is the Nadaraya-Watson etimator:

$$\begin{aligned} f_n^{NW}(x) &= \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} \text{ if } \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0 \\ f_n^{NW}(x) &= 0 \text{ otherwise.} \end{aligned}$$

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- The Nadaraya-Watson etimator with rectangular kernel takes $K(u) = \frac{1}{2} I(|u| < 1)$, then $f_n^{NW}(x)$ is the average of such Y_i that $X_i \in [x - h, x + h]$. For fixed n, the two extreme cases for the bandwidth are
  - $h \to \infty$, then $f_n^{NW}(x)$ tends to $n^{-1} \sum_{i=1}^n Y_i$ which is a constant independent of x. The bias can be too large, this is the situation of *oversmoothing*.
  - $h \to 0$, then $f_n^{NW}(X_i) = Y_i$ whenever $h <$ $\min_{i,j} |X_i - X_j||$ and $\lim_{h \to 0}(x) = 0$ if $x \neq X_i$.which is a constant independent of x. The bias can be too large, this is the situation of *oversmoothing*.

$$\begin{aligned} f_n^{NW}(x) &= \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} \text{ if } \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \neq 0 \\ f_n^{NW}(x) &= 0 \text{ otherwise.} \end{aligned}$$

  The estimator is too oscillating and reproduces the data Y_i at points X_i and vanishes elsewhere. This makes the stochastic error (varianve) too large. In other wods *undersmoothing* occurs.

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- An optimal bandwitdh h yielding a balance between bias and variance is situated between these two extremes.
- The Nadaraya-Watson estimator can be represented as a weighted sum of the Y_i

$$f_n^{NW}(x) = \sum_{i=1}^n Y_i W_{ni}^{NW}(x)$$

where the weights are

$$W_{ni}^{NW}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} I\left(\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right) \neq 0\right)$$

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- Definition 8.5:An estimator $f_n(x)$ of $f(x)$ is called a linear nonparametric regression estimator if it can be written in the form

$$f_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

where the weights $W_{ni}(x) = W_{ni}(x, X_1 ... X_n)$ depend only on n, x and $X_1, ... X_n$.

- The weights are typically chosen such that

$$\sum_{i=1}^{n} W_{ni}(x) = 1$$

for all x.

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- An intuitive motivation for the Nadaraya-Watson estimator is to note that

$$f(x) = \mathbb{E}(Y|X = x) = \frac{\int yp(y,x)dy}{p(x)}$$

therefore if we replace $p(y,x)$ with a kernel estimator $p_n(y,x)$ of the density of (Y,X) and use the kernel estimator $p_n(x)$ of $p(x)$, we obtain the Nadaraya-Watson estimator in view of the folowing result.

- Result 8.8:Let $p_n(x)$ and $p_n(x,y)$ be the kernel density estimators of $p(x)$ and $p(x,y)$ previously defined, w kernel K of order 1. Then

$$f_n^{NW}(x) = \frac{\int yp_n(y,x)dy}{p_n(x)}$$

if $p_n(x) \neq 0$.

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- Proof of Result 8.8: We have

$$\int yp_n(y,x)dy = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \int yK\left(\frac{Y_i - y}{h}\right) dy$$

Since K is of order 1,

$$\begin{aligned} & \frac{1}{h} \int yK\left(\frac{Y_i - y}{h}\right) dy \\ = & \int \frac{y - Y_i}{h} K\left(\frac{Y_i - y}{h}\right) dy \\ & + \frac{Y_i}{h} \int K\left(\frac{Y_i - y}{h}\right) dy \\ = & -\int K(u)udu + Y_i \int K(u)du = Y_i \end{aligned}$$

- **Nonparametric regression. The Nadaraya-Watson estimator.**
- If the marginal density of X is known, we can use $p(x)$ instead of $p_n(x)$. The we get the following estimator which is slightly different from the NW estimator

$$\begin{aligned} f_n(x) &= \frac{\int yp_n(y,x)dy}{p(x)} \\ &= \frac{1}{nhp(x)} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right) \end{aligned}$$

- In particular, if p is uniform on [0,1]

$$\begin{aligned} & f_n(x) \\ = & \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right) \end{aligned}$$

This estimator can therefore be used in both random uniform design or regular fixed design $(X_i = i/n)$.

- **Nonparametric regression. Local Polynomial estimators**
- If the kernel K is nonnegative, the NW estimator satisfies

$$f_n^{NW}(x) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right)$$

  Thus $f_n^{NW}(x)$ may be viewed as a local constant weighted least squares approximation of outcome Y_i
- The degree of locality is determined by the kernel K, that downweights obs with X that are not close to x, whereas θ plays the role of a local constant to be fitted.
- To further reduce the bias, we may define a local polynomial least square approximation to exploit smoothness in $f(x)$, by replacing the constant function w a polynomial of given degree.

- **Nonparametric regression. Local Polynomial estimators**
- If f is $\Sigma(\beta, L), \beta > 1$ then for z sufficently close to x

$$\begin{aligned} f(z) &= f(x) + f'(x)(z-x) + ... + \frac{f^{(l)}(x)}{l!}(z-x)^l \\ &= \theta^T(x) U\left(\frac{z-x}{h}\right) \end{aligned}$$

  where $l = \lfloor \beta \rfloor$

$$\begin{aligned} U(u) &= \left(1, u, u^2/2!, ..., u^l/l!\right)^T \\ \theta^T(x) &= \left(f(x), f'(x)h + f^{''}(x)h^2 ... + f^{(l)}(x)h^l\right)^T \end{aligned}$$

- **Nonparametric regression. Local Polynomial estimators**
- Definition 8.6:Let K be a kernel, $h > 0$ be a bandwidth and $l \geq 0$ be an integer. A vector $\theta_n(x) \in \mathbb{R}^{l+1}$ defined by

$$\theta_n(x) = \arg\min_{\theta \in \mathbb{R}^{l+1}} \sum_{i=1}^{n} \left(Y_i - \theta^T U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right)$$

  is called a **local polynomial of order** l of $\theta(x)$ or LP(l) estimator of $f(x)$ for short.
- Note that $f_n(x)$ is the first coordinate of the vector $\theta_n(x)$.
- Also note that the NW estimator with $K \geq 0$ is the LP(0) estimator.
- Furthermore, note that properly normalized coordinates of $\theta_n(x)$ provide estimators of the derivatives $f'(x), f^{''}(x), ..., f^{(l)}(x)$.

- **A note on the Curse of Dimensionality**
- If X takes values in a high-dimensional space (i.e. $X \in \mathbb{R}^d$ for large d), estimating the regression function can be especially difficult.
- The main reason for this is athat in the case of large d, in general it is not possible to densely pack the space of X with finitely many sample points, even if the sample size is very large.
- This fact is often referred to as the "curse of dimensionality"

- **A note on the Curse of Dimensionality**
- An illustration of COD. Let $X_1, ..., X_n$ be i.i.d. uniform$\left([0,1]^d\right)$
- Denote the expected supremum-norm distance of X to its nearest neighbor in $X_1, ..., X_n$ by

$$D_\infty(d,n) = \mathbb{E}\left\{\min_{i=1,...n} ||X - X_i||_\infty\right\}$$

where $||x||_\infty$ is the supremum norm of a vector $x = (x^{(1)}, ..., x^{(d)})$ defined by

$$||x||_\infty = \max_{l=1,...,d} |x^{(l)}|$$

- Then

$$\begin{aligned} D_\infty(d,n) &= \int_0^\infty \Pr\left\{\min_{i=1,...,n} ||X - X_i||_\infty > t\right\} dt \\ &= \int_0^\infty 1 - \Pr\left\{\min_{i=1,...,n} ||X - X_i||_\infty \le t\right\} dt \end{aligned}$$

- **A note on the Curse of Dimensionality**
- The bound

$$\begin{aligned} \Pr\left\{\min_{i=1,...,n} ||X - X_i||_\infty \le t\right\} &\le n \Pr\{||X - X_1||_\infty \le t\} \\ &\le n\,(2t)^d \end{aligned}$$

implies that

$$\begin{aligned} D_\infty(d,n) &\ge \int_0^{1/\left(2n^{1/d}\right)} (1 - n\,(2t)^d) dt \\ &= \frac{d}{2\,(d+1)} \frac{1}{n^{1/d}} \end{aligned}$$

- **A note on the Curse of Dimensionality**
-

	n=100	n=1000	n=100000
$D_\infty(1,n)$	≥ 0.0025	≥ 0.0025	≥ 0.0000025
$D_\infty(10,n)$	≥ 0.28	≥ 0.18	≥ 0.14
$D_\infty(20,n)$	≥ 0.37	≥ 0.30	≥ 0.26

- This table shows values of this lower bound for various values of d and n. For $d = 10, 20$, this lower bound is not close to zero even if the sample size is extremely large
- So for most values of x one only has data points (X_i, Y_i) where X_i is not close to x.At such data points, $m(X_i)$ will in general, not be close to $m(x)$ even for very smooth regression functions

- **A note on the Curse of Dimensionality**
- The only way to overcome the COD is to incorporate additional assumptions about the regression function besides the sample.
- This is implicitly done by nearly all multivarate estimation procedures.
- A similar pb occurs if one replaces the sup norm by the euclidean norm.
- The arguments above are no longer valid if the components of X are not independent, they are approx correct if one replaces d wwith the "intrisic" dimension of the X, i.e. the number of independent components of X.

- **Pointwise risk of local polynomial estimators**
- Consider again the case of $d = 1$, Recall that the local polynomial estimator is defined as :

$$\theta_n(x) = \arg\min_{\theta \in \mathbb{R}^{l+1}} \sum_{i=1}^{n} \left(Y_i - \theta^T U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right)$$

- A unique solution exists

$$f_n(x) = \sum_{i=1}^{n} Y_i W_{ni}^{LP}(x)$$

where
$$W_{ni}^{LP}(x) = \frac{1}{nh} U^T(0)\, \mathcal{B}_{n,x}^{-1} U^T\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

Provided the matrix

$$\mathcal{B}_{n,x} = \frac{1}{nh} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^T\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

is positive definite.

- **Pointwise risk of local polynomial estimators**
- You will show in lab that the local polynomial of order l reproduces polynomials of degree $\leq l$. That is,

$$\sum_{i=1}^{n} Q(X_i)\, W_{ni}^{LP}(x) = Q(x)$$

for Q a polynomial of degree $\leq l$. In particular

$$\sum_{i=1}^{n} W_{ni}^{LP}(x) = 1; \sum_{i=1}^{n} (X_i - x)^k\, W_{ni}^{LP}(x) = 0, k \leq l$$

- **Pointwise risk of local polynomial estimators**
- Assumptions (LP):
- **(LP1)** There exist a real number λ_0 and a positive integer n_0 such that the smallest eigenvalue $\lambda_{\min}(\mathcal{B}_{n,x})$ satisfies

$$\lambda_{\min}(\mathcal{B}_{n,x}) \geq \lambda_0$$

for all $n \geq n_0$ and any $x \in [0, 1]$

- This assumption is stronger than requiring positive definiteness for given x and n as it is uniform in both. The assumption is natural in the case where the matrix $\mathcal{B}_{n,x}$ has an asymptotic limit
- **(LP2)**There exist a real number $a_0 > 0$ such that for any interval $A \subseteq [0, 1]$ and all $n \geq 1$

$$\frac{1}{n} \sum_i I(X_i \in A) \leq a_0 \max(\text{Leb}(A), 1/n)$$

- This second assumption means the points X_i are sufficiently dense in the interval $[0, 1]$. one can verify that it is satisfied for $X_i = i/n$.

- **Pointwise risk of local polynomial estimators**
- Assumptions (LP):
- **(LP3)**The kernel K has compact support belonging to [-1,1] and there exist a number $K_{\max} < \infty$ such that $|K(u)| \leq K_{\max}$ for all $u \in \mathbb{R}$.
- this last assumption is not much of a restriction since we are free to pick K.
- Because the matrix $\mathcal{B}_{n,x}$ is symmetric, LP1 implies that for all $n \geq n_0, x \in [0, 1]$ and $v \in \mathbb{R}^{l+1}$

$$\left\|\mathcal{B}_{n,x}^{-1} v\right\| \leq \|v\| / \lambda_0$$

where $\|t\|$ is the Euclidean norm.

- **Pointwise risk of local polynomial estimators**
- We will use the following result:
- Result 8.9 :Under (LP1)-(LP3), for all $n \geq n_0, h > 1/2n$ and $x \in [0,1]$, the weights $W_{ni}^{LP}(x)$ of the LP(l) estimator satisfy the following:

$$\begin{aligned}(i) \sup_{i,x} |W_{ni}^{LP}(x)| &\leq \frac{C_*}{nh} \\ (ii) \sum_{i}^{n} |W_{ni}^{LP}(x)| &\leq C_* \\ (iii)\ W_{ni}^{LP}(x) &= 0 \text{ if } |X_i - x| > h\end{aligned}$$

where C^* only depends on λ_0, a_0 and $K_{\max}$.

- **Pointwise risk of local polynomial estimators**
- Result 8.10 :Suppose that f belongs to $\Sigma(\beta, L)$ on $[0,1]$. Let f_n be the LP(l) estimator of f with $l = \lfloor \beta \rfloor$. Furthermore, suppose that
- The design points $X_1, ..., X_n$ are deterministic.
- Assumptions (LP1)-(LP3) hold.

The random variables $\varepsilon_i = Y_i - f(X_i)$ are independent and such that for all $i = 1, ..., n$.

$$\mathbb{E}(\varepsilon_i) = 0,\ \mathbb{E}(\varepsilon_i^2) \leq \sigma_{\max}^2 < \infty$$

Then for all $x_0 \in [0,1]$, $n \geq n_0, h > 1/2n$, the following upper bounds hold:

$$\begin{aligned}(i)\, |b(x_0)| &= |\mathbb{E}[f_n(x_0)] - f(x_0)| \leq q_1 h^\beta \\ (ii) \sigma^2(x_0) &= \mathbb{E}\left\{ (f_n(x_0) - \mathbb{E}[f_n(x_0)])^2 \right\} \leq \frac{q_2}{nh}\end{aligned}$$

where $q_1 = C_* L/l!$ and $q_2 = \sigma_{\max}^2 C_*^2$.

- **Pointwise risk of local polynomial estimators**
- Proof:

$$\begin{aligned}b(x_0) &= \mathbb{E}[f_n(x_0)] - f(x_0) \\ &= \sum_{i=1}^{n} (f(X_i) - f(x_0)) W_{ni}^{LP}(x_0) \\ &= \sum_{i=1}^{n} \frac{(f^{(l)}(x_0 + \tau_i (X_i - x_0)) - f^{(l)}(x_0))}{l!} (X_i - x_0)^l W_{ni}^{LP}(x_0)\end{aligned}$$

for $0 \leq \tau_i \leq 1$. Therefore

$$\begin{aligned}|b(x_0)| &\leq \sum_{i=1}^{n} \frac{L}{l!} |(X_i - x_0)|^\beta \left|W_{ni}^{LP}(x_0)\right| \\ &\leq \sum_{i=1}^{n} \frac{L}{l!} h^\beta \left|W_{ni}^{LP}(x_0)\right| = q_1 h^\beta\end{aligned}$$

- **Pointwise risk of local polynomial estimators**
- Proof:The variance satisfies

$$\begin{aligned}\sigma^2(x_0) &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \varepsilon_i W_{ni}^{LP} \right)^2 \right] \\ &= \sum_{i=1}^{n} \left(W_{ni}^{LP}\right)^2 \mathbb{E}\left[(\varepsilon_i)^2\right] \\ &\leq \sigma_{\max}^2 \sup_{i,n} \left|W_{ni}^{LP}\right| \sum_{i=1}^{n} \left|W_{ni}^{LP}\right| \leq \sigma_{\max}^2 \frac{C^{*2}}{nh}\end{aligned}$$

- **Pointwise risk of local polynomial estimators**
- The result implies that
$$MSE \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}$$
and the minimizer h^* is
$$h^* = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$
which gives the upper bound
$$\lim_{n\to\infty}\sup \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbb{E}_f|\psi_n^{-2}\left(f_n\left(x_0\right)-f\left(x_0\right)\right)^2 \leq C<\infty,$$
$$\psi_n = n^{-\frac{\beta}{2\beta+1}}$$
a similar bound applies to the MISE.

- **Cross validation of regression function estimator.**
- ψ_n is analogous to the nonparametric rate we have proviously attained w kernel estimation of a density
- We will show later that this rate is minimax, in the sense that it is not possible to find a nonparametric estimator that converges at a faster rate uniformely over the assumed Hölder ball.
- In practice, how to choose the bandwidth.
- A good way is by cross-validation, similar to kernel density estimation

- **Cross validation of regression function estimator.**
- To understand how well cross-validation works in the regression context, let
$$\Delta_n^h = \mathbb{E}||f_n^h - f||^2$$
the expected mean squared error of a kernel estimator f_n^h based on n samples and bandwidth h.
- The cross validation selection of h is
$$H_n = \arg\min_{h\in Q_n} \frac{1}{n}\sum_{i=1}^n \left(f_{n,i}^h\left(X_i\right)-Y_i\right)^2$$
where,Q_n is a discrete finite set indexed by n, $f_{n,i}^h$ is the leave-i-out LP estimator of f using the sample of size $n-1$ from which unit i has been deleted.
- The corresponding cross validation regression estimator is $f_n^{H_n}\left(x\right)$.

- **Cross validation of regression function estimator.**
- Let
$$\mathbb{E}\left(\Delta_{n-1}^{H_n}\right) = \mathbb{E}||f_{n-1}^{H_n} - f||^2$$
- The following result compares the above risk of the selected estimator to that of the oracle
$$\Delta_{n-1}^{\overline{h}_{n-1}} = \min_{h\in Q_n} \Delta_{n_{-1}}^h$$
with $\overline{h}_{n-1}$ the best deterministic choice for sample size $n-1$, and shows that the two risks are in fact quite close.

- **Cross validation of regression function estimator.**
- Assuming $|Y| < L^* < \infty$, for any $\delta > 0$

$$\mathbb{E}\left(\Delta_{n-1}^{H_n}\right) \leq (1+\delta)\,\Delta_{n-1}^{\overline{h}_{n-1}} + c\frac{|Q_n|}{n}\log n$$

where c depends only on δ, L^* and an upper bound for $\sum_i^n |W_{ni}^{LP}(x)|$

- This is remarkable result as it shows that for small δ, the risk of the cross validated LP estimator $\mathbb{E}\left(\Delta_{n-1}^{H_n}\right)$ and that of the oracle $\Delta_{n-1}^{\overline{h}_{n-1}}$ are close to each other, within a correction term $c\frac{|Q_n|}{n}\log n$.
- One is interested in settings where the correction term is of smaller order than $\Delta_{n-1}^{\overline{h}_{n-1}}$ since both are shrinking to zero.

- **Cross validation of regression function estimator.**
- Note that if f is a d dimensional function in $\Sigma(\beta = 1, L)$, then the oracle local polynomial satisfies

$$\Delta_{n-1}^{\overline{h}_{n-1}} = O\left(n^{-\frac{2}{d+2}}\right)$$

- Therefore the correction is of small order than $\Delta_{n-1}^{\overline{h}_{n-1}}$ only if $|Q_n|$ is not too large, i.e. only if $|Q_n|$ is roughly (ignoring the log term) $O\left(n^k\right)$, $k < d/(d+2)$.
- In other words, cross validation works well provided the number of candidates bandwiths is not too large relative to the effective smoothness $2/d$ in this case.
- What is also remarkable is that the error incured for model selection can be quite smaller compared to the nonparametric risk of estimation essentially $\frac{|Q_n|}{n}$.

- **Lower bounds on the minimax risk.**
- The pb of nonparametric inference is characterized by the following$g$:
  - A nonparametric class of functions Θ containing the function θ that we aim to estimate. e.g. $\Theta = \Sigma(\beta, L)$ the Hölder class.
  - A family $\{P_\theta : \theta \in \Theta\}$ of probability measures, indexed by θ. For example in the density model P_θ is the probability measure associated with a sample $X_1, ..., X_n$ of size n with density function $p(\cdot) = \theta$.
  - A distance (or semi-distance) d on a Θ used to define risk. The key property of d is that it is positive and satisfies the triangle inequality $d(\theta, \theta') = d(\theta', \theta)$; $d(\theta, \theta') + d(\theta', \theta'') \geq d(\theta, \theta'')$ and $d(\theta, \theta) = 0$.
  - e.g. $d(f, g) = |f(x_0) - g(x_0)|$ for fixed $x_0; d(f, g) = ||f - g||_2; d(f, g) = ||f - g||_\infty$.

- **Lower bounds on the minimax risk.**
- The maximum risk of an esitmator θ_n of θ is defined as

$$r(\theta_n) = \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[d^2(\theta_n, \theta)\right]$$

- We established upper bounds on the maximum risk for several estimators in nonparametric problems,

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\left[d^2(\theta_n, \theta)\right] \leq C\psi_n^2$$

for certain estimators θ_n, certain positive sequences $\psi_n \to 0$ and constants C.

- We will now consider complementing these upper bounds with corresponding lower bounds

$$\text{For all } \theta_n : \quad \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[d^2(\theta_n, \theta)\right] \geq c\psi_n^2$$

for n sufficiently large where c is a positive constant.

- **Lower bounds on the minimax risk.**
- Minimax risk associated with model $\{P_\theta : \theta \in \Theta\}$ and w semidistance d:

$$\mathcal{R}_n = \inf_{\theta_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2 (\theta_n, \theta) \right]$$

where the infimum is over all estimators. The upper bounds established in previous lectures imply that there exist a constant $C < \infty$ such that

$$(i) \lim \sup_{n \to \infty} \psi_n^{-2} \mathcal{R}_n \leq C$$

- The corresponding claim is that the there exists a constant $c > 0$ such that, for the same sequence ψ_n,

$$(ii) \lim \inf_{n \to \infty} \psi_n^{-2} \mathcal{R}_n \geq c$$

- A positive sequence ψ_n satisfying (i) and (ii) is called an optimal rate of convergence and an estimator θ_n^* satsifying

$$\mathcal{R}_n = \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2 (\theta_n, \theta) \right] \leq C'' \psi_n^2$$

is called a rate optimal estimator

- **Lower bounds on the minimax risk.**
- A general scheme for obtaining lower bounds is based on the following three remarks:
  1. Reduction to bounds on probabilities: By Markov inequality, for any monotone increasing function w such $w(0) = 0$ and for any A such that $w(A) > 0$ we have

$$\begin{aligned} \mathbb{E}_\theta \left[w \left(\psi_n^{-1} d (\theta_n, \theta) \right) \right] &\geq w(A) P_\theta \left\{ \psi_n^{-1} d (\theta_n, \theta) \geq A \right\} \\ &= w(A) P_\theta \left\{ d (\theta_n, \theta) \geq s \right\} \end{aligned}$$

with $s = \psi_n A$. Therefore, instead of searching for a lower bound on the minimax risk $\mathcal{R}_n$, it is sufficient to find a lower bound on the minimax probabilities of the form

$$\inf_{\theta_n} \sup_{\theta \in \Theta} P_\theta \left\{ d (\theta_n, \theta) \geq s \right\}$$

giving a first simplification.

- **Lower bounds on the minimax risk.**
  2. Reduction to a finite number of hypotheses. We note that

$$\inf_{\theta_n} \sup_{\theta \in \Theta} P_\theta \left\{ d (\theta_n, \theta) \geq s \right\} \geq \inf_{\theta_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta \left\{ d (\theta_n, \theta) \geq s \right\}$$

for any finite set $\{\theta_0, ..., \theta_M\}$ of *hypotheses* contained in Θ. We will select the $M + 1$ hypotheses carefully to obtain lower bounds on the minimax risk . We will also call any function $\Psi : (X_1, ..., X_n) \to \{0, 1, ...M\}$ a test to decide which hypothesis generated the data.

- **Lower bounds on the minimax risk.**
  3. Choice of $2s-$separated hypotheses. If

$$d (\theta_j, \theta_k) \geq 2s \text{ for all } j \neq k$$

then for any estimator θ_n

$$P_{\theta_j} \left\{ d (\theta_n, \theta_j) \geq s \right\} \geq P_{\theta_j} \left\{ \Psi \neq j \right\} \; j = 0, 1, ...M$$

where Ψ is the minimum distance test defined by

$$\Psi = \arg \min_{0 \leq k \leq M} d (\theta_n, \theta_k)$$

The inequality follows form the $2s$ separation between hypotheses and the triangle inequality

- **Lower bounds on the minimax risk.**
- We conclude that if we can construct $M+1$ hypotheses satisfying

$$\begin{aligned}\inf_{\theta_n}\sup_{\theta\in\Theta} P_\theta\left\{d\left(\theta_n,\theta\right)\geq s\right\} &\geq \inf_{\theta_n}\max_{\theta\in\{\theta_0,\ldots,\theta_M\}} P_\theta\left\{d\left(\theta_n,\theta\right)\geq s\right\}\geq p_{e,M}\\ p_{e,M} &= \inf_{\Psi}\max_{0\leq j\leq M} P_\theta\left\{\Psi\neq j\right\}\end{aligned}$$

  where $\inf_\Psi$ is over all tests. The lower bound is obtained if one can find a constant $c'>0$ independent of n.
- The probability $p_{e,M}$ may be interpreted as minimax probability of error for testing $M+1$ hypotheses $\{\theta_0,\ldots,\theta_M\}$.

- **Lower bounds on the minimax risk.**
- Lower bounds based on two hypotheses: Consider $M=1$ and denote $P_0=P_{\theta_0}$ and $P_1=P_{\theta_1}$. We find a lower bound for the minimax probability of error $p_{e,1}$ based on the likelihood ratio $\frac{dP_0}{dP_1}$ (assuming the two measures are absolutely continuous, this assumption can be relaxed)
- Result 8.11:

$$p_{e,1}\geq\sup_{\tau>0}\left\{\frac{\tau}{1+\tau}P_1\left\{\frac{dP_0}{dP_1}>\tau\right\}\right\}$$

- **Lower bounds on the minimax risk.**
- Thus in order to obtain a lower bound it suffices to find constants $\tau>0$ and $0<\alpha<1$ independent of n and satisfying

$$P_1\left\{\frac{dP_0}{dP_1}>\tau\right\}\geq 1-\alpha$$

  this means that the two laws P_0 and P_1 are not very far apart, and the closer they are to each other as controlled by $d\left(\theta_0,\theta_1\right)$, the greater is the lower bound

$$\inf_{\theta_n}\sup_{\theta\in\Theta} P_\theta\left\{d\left(\theta_n,\theta\right)\geq s\right\}\geq\sup_{\tau>0}\left\{\frac{\tau\left(1-\alpha\right)}{1+\tau}\right\}$$

- **Lower bounds on the minimax risk.**
- Proof of Result 8.11.

$$\begin{aligned}P_0\left(\Psi=1\right) &= \int I(\Psi=1)\frac{dP_0}{dP_1}dP_1\\ &\geq \tau\int I(\{\Psi=1\}\cap\left\{\frac{dP_0}{dP_1}\geq\tau\right\})\frac{dP_0}{dP_1}dP_1\\ &\geq \tau\left(p-\alpha_1\right)\end{aligned}$$

  where $p=P_1\left(\Psi=1\right)$ and $\alpha_1=P_1\left(\frac{dP_0}{dP_1}<\tau\right)$. Then

$$\begin{aligned}p_{e,1} &= \inf_{\Psi}\max_{j=0,1} P_j\left(\Psi\neq j\right)\geq\min_{0\leq p\leq 1}\max\left\{\tau\left(p-\alpha_1\right),1-p\right\}\\ &= \frac{\tau}{1+\tau}\left(1-\alpha_1\right)\end{aligned}$$

- **Lower bounds on the minimax risk.**
- Other Distance between probabilities: The Kullback divergence

$$K(P,Q) = \begin{cases} \int \log \frac{dP}{dQ} dP \text{ if } P << Q \\ +\infty \text{ otherwise} \end{cases}$$

- We will use the following result without proof (see Tsybakov for proof)
- Result 8.12:If $K(P_0, P_1) \le \alpha < \infty$, then

$$p_{e,1} \ge \max \left(\frac{1}{4} \exp(-\alpha), \frac{1 - \sqrt{\alpha/2}}{2} \right)$$

- Similar bounds can be obtained using total variation distance, Hellinger distance and chi-squared divergence.

- **Lower bounds on the minimax risk.**
- Risk of regression estimator at a point:
- Assumption(A) The statistical model is that of nonparametric regresssion
  - (i)

$$Y_i = f(X_i) + \varepsilon_i, \; i = 1, ..., n$$

  - (ii)The random variables ε_i are $i.i.d$ having a density p_ε st there exist $p_* > 0$ and $v_0 > 0$:

$$\int p_\varepsilon(u) \log \frac{p_\varepsilon(u)}{p_\varepsilon(u) + v} du \le p_* v^2$$

    for all $|v| \ge v_0$
  - (iii) The variables $X_i \in [0,1]$ are deterministic.

- **Lower bounds on the minimax risk.**
- Risk of regression estimator at a point:
- Assumption (A) part (ii) can be shown to hold for the normal density $N(0, \sigma^2)$.
- We will also assume that **(LP2)** holds, which states that there exist a real number $a_0 > 0$ such that for any interval $A \subseteq [0,1]$ and all $n \ge 1$

$$\frac{1}{n} \sum_i I(X_i \in A) \le a_0 \max(\mathrm{Leb}(A), 1/n)$$

- Our aim is to obtain a lower bound for the minimax risk on (Θ, d) when $\Theta = \Sigma(\beta, L)$ and d is a distance at a fixed point x_0 in the unit interval $[0,1]$

$$d(f,g) = |f(x_0) - g(x_0)|$$

- The rate that we wish to establish as lower bound is

$$\psi_n = n^{-\frac{\beta}{1+2\beta}}$$

- **Lower bounds on the minimax risk.**
- Risk of regression estimator at a point:
- By the general scheme we have outlined above, it suffices to prove that:

$$\inf_{\theta_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta(d(\theta_n, \theta) \ge s) \ge c' > 0$$

  where $s = A\psi_n$ with a constant $A > 0$.
- Using two hypotheses $M - 1$ we wish to establish

$$\inf_{f_n} \sup_{f \in \{f_0, f_1\}} P_\theta(|f_n(x_0) - f(x_0)| \ge A\psi_n) \ge c' > 0$$

  where $f_0 = \theta_0; f_1 = \theta_1$.
- In order to apply this latter bound, we work with the Kullback bound for the minimax error probability for the choice of hypotheses:

$$f_0(x) \equiv 0 \text{ and } f_1^n(x) = L h_n^\beta K\left(\frac{x - x_0}{h_n}\right), \; x \in [0,1]$$

  where

$$h_n = c_0 n^{-\frac{1}{2\beta+1}}, c_0 > 0$$

- **Lower bounds on the minimax risk.**
- Risk of regression estimator at a point:
- The function K is assumed to satisfy conditions

$$K \in \Sigma(\beta, 1/2), \; K(u) > 0 \text{ only if } u \in \left(-\frac{1}{2}, \frac{1}{2}\right)$$

- A convenient choice of such function is given by

$$\begin{aligned} K(u) &= aK_0(u) \\ K_0(u) &= \exp\left(-\frac{1}{1-u^2}\right) I(|u| \leq 1) \end{aligned}$$

  for sufficiently small $a > 0$.
- Before obtaining the sought bound, we first need to ensure that
  - (a)$f_j^n \in \Sigma(\beta, L)$ $j = 0, 1$
  - (b)$d(f_0^n, f_1^n) \geq 2s$
  - (c) $K(P_0, P_1) \leq \alpha < \infty$.

- **Lower bounds on the minimax risk.**
- (a) For $l = \lfloor \beta \rfloor$ the lth order derivative of f_1^n is

$$f_1^n(x) = Lh_n^{\beta-l} K^{(l)}\left(\frac{x - x_0}{h_n}\right)$$

  Then

$$\begin{aligned} \left|f_1^n(x) - f_1^n(x')\right| &= Lh_n^{\beta-l}\left|K^{(l)}(u) - K^{(l)}(u')\right| \\ &\leq Lh_n^{\beta-l}\left|K^{(l)}(u) - K^{(l)}(u')\right| \\ &= Lh_n^{\beta-l}\left|u - u'\right|^{\beta-l}/2 = L|x - x'|^{\beta-l}/2 \end{aligned}$$

  where $u = (x - x_0)/h$ and $u' = (x' - x_0)/h$. Proving the result
  - (b)$d(f_0^n, f_1^n) \geq 2s$:

$$\begin{aligned} d(f_0^n, f_1^n) &= |f_1^n(x_0)| = Lh_n^{\beta} K(0) = Lc_0^{\beta} n^{-\frac{\beta}{2\beta+1}} \\ \text{proving the result for } s &= \frac{1}{2} Lc_0^{\beta} n^{-\frac{\beta}{2\beta+1}} = A\psi_n \end{aligned}$$

- **Lower bounds on the minimax risk.**
- (c) $K(P_0, P_1) \leq \alpha < \infty$.

$$\begin{aligned} K(P_0, P_1) &= \int \log \frac{dP_0}{dP_1} dP_0 \\ &= \int \log \prod_{i=1}^n \frac{p_\varepsilon(Y_i)}{p_\varepsilon(Y_i - f_1^n(X_i))} \prod_{i=1}^n p_\varepsilon(Y_i) dY_i \\ &= \sum_{i=1}^n \int \log \frac{p_\varepsilon(y)}{p_\varepsilon(y - f_1^n(X_i))} p_\varepsilon(y) dy \leq p_* \sum_{i=1}^n f_1^n(X_i)^2 \\ &= p_* L^2 h^{2\beta} \sum_{i=1}^n K^2\left(\frac{X_i - x_0}{h_n}\right) \\ &\leq p_* L^2 K_{\max}^2 h^{2\beta} \sum_{i=1}^n I\left(\left|\frac{X_i - x_0}{h_n}\right| \leq \frac{1}{2}\right) \\ &\leq p_* L^2 K_{\max}^2 h^{2\beta} a_0 \max(nh_n, 1) \\ &= p_* L^2 K_{\max}^2 n h^{2\beta+1} a_0 \\ &= c_0^{2\beta+1} p_* L^2 K_{\max}^2 \end{aligned}$$

  proving the result $K(P_0, P_1) \leq \alpha = c_0^{2\beta+1} p_* L^2 K_{\max}^2$

- **Lower bounds on the minimax risk.**
- We have therefore established the following result:
- Result 8.13:

$$\lim_{n\to\infty}\inf \inf_{f_n} \sup_{f \in \Sigma(\beta, L)} \mathbb{E}(n^{\frac{2\beta}{2\beta+1}} \left|f_n(x_0) - f(x_0)\right|^2 \geq \alpha$$

- That is no estimator of $f(x_0)$ can achieve a uniform rate of convergence over $\Sigma(\beta, L)$ faster than $n^{-\frac{2\beta}{2\beta+1}}$ wrt risk $\left|f_n(x_0) - f(x_0)\right|^2$.
- However, we have also shown that LP(l) achieves this rate since its risk is bounded above by $Cn^{-\frac{2\beta}{2\beta+1}}$ for a constant C and therefore we may conclude that LP(l) is rate minimax!!!!