Lecture Notes on Non-Parametric Econometrics

ETE spirit

These notes have been last updated on February 6, 2025

Contents

1 Notes on Probability

1.1 On measure space

Let Σ be a set, we want to make sense of $\mathbb{P}(F)$ the probability of the event F, where $F \subseteq \Sigma$.

Definition 1.1. A collection of subsets of Σ is a σ-algebra if:

- $\Sigma \in F$
- $f \in F \implies F^c = \{w \in \Sigma : w \notin F\} \in F$
- For every sequence (F_n) of elements of F, $\bigcup_{n\in N} F_n \in F$. Note that this property is the same as: $\bigcap_{n\in N}(F_n^c)^c$

Remark 1.1. In measure theory, one axiomatizes the notion of 'measurable set', insisting that the union of a countable collection of measurable sets is measurable, and that the intersection of a countable collection of measurable sets is also measurable. Also, the complement of a measurable set must be mea- surable, and the whole space must be measurable. Thus the measurable sets form a sigma-algebra, a structure stable (or 'closed') under countably many set operations. Without the insistence that 'only countably many operations are allowed', measure theory would be self-contradictory - a point lost on certain philosophers of probability [1].

Example 1.1. If $\Sigma = \mathbb{R}^d$, the Borel sigma-algebra, denoted $\mathbb{B}(\mathbb{R}^d)$ is the smallest sigma-algebra of $\mathbb{R}^d$ which contains all open sets.

Definition 1.2. $\mu : F \to [0, \infty]$ is a measure on Σ, F if:

- $\mu(\emptyset) = 0$
- if for all sequences $(F_n)_{n\in N}$ of disjoint sets of F we have that $\mu(\bigcap)_{n\in N} = \sum_{n\in N} \mu(F_n)$

Definition 1.3. $\mathbb{P}$ is a probability if it is a measure and $\mathbb{P}(\Sigma) = 1$

Definition 1.4. $(\Sigma, F, \mathbb{P})$ is a probability triple if F is a sigma-algebra on Σ and $\mathbb{R}$ a probability on (Σ, F).

Definition 1.5. $X : (\Sigma, F, \mathbb{P}) \to (\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$ is a random vector if $\forall \mathbb{B} \in \mathbb{B}(\mathbb{R}^d), X^{-1}(\mathbb{B} \in F)$. That is, X is f-measurable.

Example 1.2. Measurability of a Random Variable: A Simple Example

Let us consider a sample space $\Omega = \{HH, HT, TH, TT\}$, representing the outcomes of flipping a fair coin twice. Define a random variable $X : \Omega \to \{0, 1, 2\}$, which takes the following values based on the number of heads in the outcome:

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = TT, \\ 1 & \text{if } \omega = HT \text{ or } \omega = TH, \\ 2 & \text{if } \omega = HH. \end{cases}$$

To determine whether X is measurable with respect to a given σ-algebra $\mathcal{F}$, recall that measurability means for every Borel set $B \subseteq \{0, 1, 2\}$, the preimage $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ must belong to $\mathcal{F}$.

Suppose $\mathcal{H}$ is the trivial σ-algebra $\{\emptyset, \Omega\}$. For example, consider the preimage $\{X = 2\} = \{HH\}$. Clearly, $\{HH\} \notin \mathcal{H}$, so X is not $\mathcal{H}$-measurable.

Define a σ-algebra $\mathcal{F}$ where X is measurable Now, let us construct a σ-algebra $\mathcal{F}$ such that X is measurable. Consider the following σ-algebra:

$$\mathcal{F} = \{\emptyset, \Omega, \{HH\}, \{TH, HT\}, \{TT\}, \{HH, TT\}, \{TH, HT, TT\}, \{HH, TH, HT\}\}.$$

We verify that $\mathcal{F}$ is a σ-algebra (it is closed under complements and countable unions). Next, we check whether X is measurable with respect to $\mathcal{F}$.

[1] straight copy pasting from the glorious and much helpful book "Probability with martingales" by Williams"

We calculate the preimages of the sets $\{0\}$, $\{1\}$, and $\{2\}$:

$$\{X = 0\} = \{TT\} \in \mathcal{F},$$
$$\{X = 1\} = \{HT, TH\} \in \mathcal{F},$$
$$\{X = 2\} = \{HH\} \in \mathcal{F}.$$

Since the preimages of all singleton sets in the range of X belong to $\mathcal{F}$, and σ-algebras are closed under unions, the preimages of all Borel subsets of $\{0, 1, 2\}$ also belong to $\mathcal{F}$. Therefore, X is $\mathcal{F}$-measurable.

Definition 1.6. Let X be a random vector in $\mathbb{R}^d$, the sigma-algebra generated by X, denoted $\sigma(X)$ is the smallest sigma-algebra on Σ s.t: $X : (\Sigma, F, \mathbb{P}) \to (\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$ is measurable.

Example 1.3. Consider a probability space $(\Sigma, \mathcal{F}, \mathbb{P})$ where:

- $\Sigma = \{a, b, c\}$,
- $\mathcal{F} = \{\emptyset, \{a\}, \{b, c\}, \{a, b, c\}\}$,
- $\mathbb{P}(\{a\}) = 0.5$, $\mathbb{P}(\{b, c\}) = 0.5$.

Define a random variable $X : \Sigma \to \mathbb{R}^2$ as:

$$X(a) = (1, 2), \quad X(b) = (3, 4), \quad X(c) = (5, 6).$$

We claim X is measurable. To see why, consider any Borel set $B \in \mathbb{B}(\mathbb{R}^2)$:

- If $B = \{(1, 2)\}$, then $X^{-1}(B) = \{a\} \in \mathcal{F}$.
- If $B = \{(3, 4), (5, 6)\}$, then $X^{-1}(B) = \{b, c\} \in \mathcal{F}$.
- If $B = \mathbb{R}^2$, then $X^{-1}(B) = \{a, b, c\} \in \mathcal{F}$.

Since $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathbb{B}(\mathbb{R}^2)$, X is measurable. This demonstrates how random variables map events in the sample space to measurable subsets of $\mathbb{R}^2$.

Remark 1.2.

- $x \in \mathbb{R}^d \to [0, \infty)$
- $x \mapsto \|x\|_2^\rho$, where $\rho > 1$ $\left(\|x\|_2 = \sqrt{\sum_{k=1}^d x_k^2}\right)$ is continuous, as a consequence we can show that $\|x\|_2^\rho$ is a random variable if X is a random vector **Exercise 1**

Definition 1.7. We write $X \in L^\rho(\Sigma, F, \mathbb{P})$ if X is F-measurable and $\mathbb{E}[\|x\|_2^\rho] < \infty$

Definition 1.8. The distribution law of X, denoted $\mathbb{P}_x$ is the probability on $(\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$ such that $\forall \mathbb{B} \in \mathbb{B}(\mathbb{R}^d)$, $\mathbb{P}(X \in \mathbb{B}) = \mathbb{P}_x(\mathbb{B})$

Definition 1.9. X has density f_x with respect to the Lebesque measure (dx) on $\mathbb{R}$ if $\mathbb{P}_x(B) = \int_{\mathbb{R}^d} \mathbb{1}\{x \in B\} f_x dx, \forall B \in \mathbb{B}(\mathbb{R}^d)$, where $f_x dx$ is $d\mathbb{P}_x(x)$

Definition 1.10. The support of X (or of $\mathbb{P}_x$) is defined as: $\text{Supp(X)} = \left\{x \in \mathbb{R}^d : \forall r > 0, \mathbb{P}(X \in B(X, r)) > 0\right\}$ where $B(x, r) := \{z : \|x - z\|_2 < r\}$ is an open ball.

Remark 1.3.

- $\text{Supp(X)}^c := \left\{x \in \mathbb{R}^d : \forall r > 0, \mathbb{P}(X \in B(X, r)) = 0\right\}$ is an open set and hence the Supp(X) is a closed set.
- Supp(X) is the longest? $C \in \mathbb{B}(\mathbb{R}^d)$ such that for all ball B such that $B \cap C \neq \emptyset \implies \mathbb{P}_x(B \cap C) > 0$

Exercise: let $Y = Y^* \mathbb{1}\left\{Y^* \geq \frac{1}{2}\right\} + \frac{1}{3}\mathbb{1}\left\{Y^* < \frac{1}{2}\right\}$ show that:

1. $supp(Y^*) = [-1, 1]$
2. $supp(Y) = \frac{1}{3} \cup [\frac{1}{2}, 1]$

Remark 1.4. Note that: $\mathbb{E}[|y|] \leq \sqrt{\mathbb{E}(y^2)}$ by Cauchy-Schwarz so $L^2(\Sigma, \mathcal{F}, \mathbb{P})$, where we can define: $\mathbb{E}(|y|) = \mathbb{E}(Y \cdot sign(Y))$.

Exercise: If X is a random vector and X_1 the rth entry, then $\sigma(X_1) \subset \sigma(X)$

1.2 Conditional probability

Definition 1.11. For a random variable $X \in L^1(\Omega, \mathcal{F}, P)$, we define the *expectation* $\mathbb{E}(X)$ of X by

$$\mathbb{E}(X) := \int_\Omega X \, dP = \int_\Omega X(\omega) P(d\omega).$$

We also define $\mathbb{E}(X) \leq \infty$ for $X \in (\Updownarrow \mathcal{F})^+$. In short, $\mathbb{E}(X) = P(X)$.

Definition 1.12. Let $Y \in L^1((\Sigma, \mathcal{F}, \mathbb{P}))$ and X be a random vector in $\mathbb{R}^d$. We define $z := \mathbb{E}[Y|X]$ is the unique[2] random v then:

- $z \in L^1(\Sigma, \sigma(x), \mathbb{P})$, where z is both integrable and σ-x-measurable.
- $\forall B \in \mathbb{B}(\mathbb{R}^d) \mathbb{E}(z \cdot \mathbb{1}\{x \in B\}) = \mathbb{E}(y \cdot \mathbb{1}\{x \in B\})$
- The second condition is true $\iff \mathbb{E}((y-z) \cdot \mathbb{1}\{x \in B\}) = 0 \iff \mathbb{E}((y-z) \cdot \Phi(x))$

Remark 1.5. z is $\sigma(x)$ measurable $\iff z = \varphi(x)$ for some $\varphi_i(\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$, which in words it means that z is a function of x which in turn is measurable.

Another definition we can find it in the book:

Definition 1.13 (Fundamental Theorem and Definition (Kolmogorov, 1933))**.** Let $(\Omega, \mathcal{F}, P)$ be a triple, and X a random variable with $\mathbb{E}(|X|) < \infty$. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then there exists a random variable Y such that:

1. Y is $\mathcal{G}$-measurable,
2. $\mathbb{E}(|Y|) < \infty$,
3. for every set $G \in \mathcal{G}$ (equivalently, for every set G in some π-system which contains Ω and generates $\mathcal{G}$), we have
$$\int_G Y \, dP = \int_G X \, dP, \quad \forall G \in \mathcal{G}.$$

Moreover, if $\tilde{Y}$ is another random variable with these properties, then $\tilde{Y} = Y$, a.s., that is, $P[\tilde{Y} = Y] = 1$. A random variable Y with properties $(a) - (c)$ is called a version of the $\mathbb{E}(X \mid \mathcal{G})$ of X given $\mathcal{G}$, and we write $Y = \mathbb{E}(X \mid \mathcal{G})$, a.s.

Remark 1.6. There are two particular cases:

1. $(y, x')'$ with discrete countable support, the definition then coincides with: $\mathbb{E}(Y|X=x) = \sum_{y_j \in \text{supp}_y} y_j \cdot \mathbb{P}(Y = y_j | X = x)$. If x is such that $\P(X = x) > 0$.
2. $(y, x')'$ and X have a density then $\mathbb{E}(y|X=x) = \int_{\mathbb{R}} y \cdot \frac{\{_{y,x}(y,x)}{\{_x(x)} dy$

exercise Let X and B be random vectors in $\mathbb{R}^d$ such that X and B are independent. Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ be continuous and bounded and $Y = g(x_i' B_i)$, show that then $\mathbb{E}[Y|X=x] = \mathbb{E}[x_i' B_i] \forall x \in \text{supp}(X)$

Exercise: provide the proof of this properties.

1.2.1 Some properties of the conditional expectation function

Note: the idea of the proof of this properties can be found in pages 88 and 89 of the masterful book Probailities with martingales. The development and errors are due only to myself.

1. Let $Y \in L^1((\Sigma, \mathcal{F}, \mathbb{P}))$[3] then $\mathbb{E}[\mathbb{E}[y|x]] = \mathbb{E}[y]$.

[2] unique means that if z and z' satisfy 1 and 2 $\implies z = z' a.s$

[3] we assume this throughout this exercise

Proof. By the definition of conditional expectation, $\mathbb{E}[Y|X]$ satisfies:

$$\int_A \mathbb{E}[Y|X]\, d\mathbb{P} = \int_A Y\, d\mathbb{P}, \quad \forall A \in \sigma(X).$$

Setting $A = \Omega$, we have:

$$\int_\Omega \mathbb{E}[Y|X]\, d\mathbb{P} = \int_\Omega Y\, d\mathbb{P}.$$

The left-hand side is $\mathbb{E}[\mathbb{E}[Y|X]]$, and the right-hand side is $\mathbb{E}[Y]$. Thus, we conclude:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y].$$

□

2. Let $g(\mathbb{R}^d, \mathbb{B}(R^d)) \to (\mathbb{R}^d, \mathbb{B}(R^d))$ measurable such that $\mathbb{E}[g(x)] < \infty$ then $\mathbb{E}[g(x)|x] = g(x) a.s.$

Proof. From Remark 1.5, we know that a random variable Z is $\sigma(X)$-measurable if and only if $Z = \varphi(X)$ for some measurable function $\varphi : (\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d)) \to (\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$.

Let $Z = g(X)$. By assumption, g is measurable and $\mathbb{E}[|g(X)|] < \infty$, so Z is $\sigma(X)$-measurable.

It is a property of conditional expectation that if Z is $\sigma(X)$-measurable and $\mathbb{E}[|Z|] < \infty$, then $\mathbb{E}[Z|X] = Z$ almost surely. Applying this property to $Z = g(X)$, we conclude:

$$\mathbb{E}[g(X)|X] = g(X) \quad \text{a.s.}$$

□

3. $\forall a_1, a_2 \in \mathbb{R}$ and $\mathbb{E}[|Y_1| + |Y_2|] < \infty$ then $\mathbb{E}[a_1 Y_1 + a_2 Y_2] = a_1 E[Y_1] + a_2[Y_2] a.s$

Proof. Given that $\mathbb{E}[|Y_1| + |Y_2|] < \infty$, the random variables Y_1 and Y_2 are integrable, and their conditional expectations are well-defined. Consider the random variable $a_1 Y_1 + a_2 Y_2$. By the definition of conditional expectation, we have:

$$\mathbb{E}[a_1 Y_1 + a_2 Y_2|X] = \int_A (a_1 Y_1 + a_2 Y_2)\, d\mathbb{P},$$

for all $A \in \sigma(X)$.

Using the linearity of the integral, this can be expressed as:

$$\int_A (a_1 Y_1 + a_2 Y_2)\, d\mathbb{P} = \int_A a_1 Y_1\, d\mathbb{P} + \int_A a_2 Y_2\, d\mathbb{P}.$$

Taking the constants a_1 and a_2 outside their respective integrals, we get:

$$\mathbb{E}[a_1 Y_1 + a_2 Y_2|X] = a_1 \int_A Y_1\, d\mathbb{P} + a_2 \int_A Y_2\, d\mathbb{P}.$$

By the definition of conditional expectation, this becomes:

$$\mathbb{E}[a_1 Y_1 + a_2 Y_2|X] = a_1 \mathbb{E}[Y_1|X] + a_2 \mathbb{E}[Y_2|X].$$

Thus, the result follows:

$$\mathbb{E}[a_1 Y_1 + a_2 Y_2|X] = a_1 \mathbb{E}[Y_1|X] + a_2 \mathbb{E}[Y_2|X] \quad \text{a.s.}$$

□

4. if $y \geq 0$ a.s then $\mathbb{E}[Y|X] \geq 0$ a.s.

   Proof. Let $Z = \mathbb{E}[Y|X]$. Suppose $\mathbb{P}(Z < 0) > 0$. Then there exists $G := \{Z < -n^{-1}\}$ for some $n \in \mathbb{N}$ with $\mathbb{P}(G) > 0$ and $G \in \sigma(X)$. By the definition of conditional expectation:

$$\int_G Y \, d\mathbb{P} = \int_G Z \, d\mathbb{P}.$$

   Since $Y \geq 0$ a.s., $\int_G Y \, d\mathbb{P} \geq 0$, but $\int_G Z \, d\mathbb{P} < -n^{-1}\mathbb{P}(G) < 0$[4], a contradiction. Thus, $\mathbb{P}(Z < 0) = 0$, so $Z \geq 0$ a.s. □

5. Let $c : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\mathbb{E}[|c(y)|] < \infty$ then $\mathbb{E}[c(y)|x] \geq c[\mathbb{E}[y|x]]$ a.s.

   We will use the following definition - which requires that function c is continuous (6.6.a in Williams): $c(x) = sup_{q \in G} \left[D_{-c}(q)(x - q) + c(q)\right] = sup_n(a_n x + b_n)$ for some sequences in $a_n b_n \in \mathbb{R}$[5].

   Proof. If we apply this definition to this problem we know that $c(x) = sup_n(a_n x + b_n)$ and $c(x) \geq (a_n x + b_n)$ for a fixed $n \in \mathbb{N}$. Then we take expectations in both sides: $\mathbb{E}(c(x)|\zeta) \geq (a_n \mathbb{E}(x|\zeta) + b_n)$, where $\zeta \in \sigma(x)$. Then for all n, this is also true by countability property[6] : $\mathbb{E}(c(x)|\zeta) \geq sup_n(a_n \mathbb{E}(x|\zeta) + b_n) = c(\mathbb{E}(x|\zeta))$. □

6. Let $\mathcal{H}, \xi$ be the sigma-algebra and that $\mathcal{H} \subset \xi$. Then $\mathbb{E}[\mathbb{E}[y|\xi]|\mathcal{H}] = \mathbb{E}[Y|\mathcal{H}]$ a.s.

   Proof. By the definition of conditional expectation, $\mathbb{E}[Y|\xi]$ satisfies $\int_A \mathbb{E}[Y|\xi] \, d\mathbb{P} = \int_A Y \, d\mathbb{P}$ for all $A \in \xi$. Now consider $\mathbb{E}[\mathbb{E}[Y|\xi]|\mathcal{H}]$. By the definition of conditional expectation, for any $B \in \mathcal{H}$,

$$\int_B \mathbb{E}[\mathbb{E}[Y|\xi]|\mathcal{H}] \, d\mathbb{P} = \int_B \mathbb{E}[Y|\xi] \, d\mathbb{P}.$$

   Since $\mathcal{H} \subset \xi$, every $B \in \mathcal{H}$ is also in ξ, so we have

$$\int_B \mathbb{E}[Y|\xi] \, d\mathbb{P} = \int_B Y \, d\mathbb{P}.$$

   Thus, by the uniqueness of conditional expectation, $\mathbb{E}[\mathbb{E}[Y|\xi]|\mathcal{H}] = \mathbb{E}[Y|\mathcal{H}]$ almost surely. □

7. Let z be a bounded random variable (i.e $\exists M \in \mathbb{R} : |z| \leq Ma.s$) that is y measurable. Then $\mathbb{E}[zy|\xi] = z\mathbb{E}[y|\xi]$ a.s. [7]

   Proof. Fix $G \in \xi$ and let $W = \mathbb{E}[Y|\xi]$. We aim to prove that

$$\mathbb{E}[ZY \mid G] = \mathbb{E}[Z \cdot W \mid G].$$

[4]Since $Z(\omega) \leq -n^{-1}$ for all $\omega \in G$, it follows that $Z(\omega) \leq -n^{-1}$ on G. Therefore, the integral satisfies:

$$\int_G Z \, d\mathbb{P} \leq \int_G (-n^{-1}) \, d\mathbb{P} = -n^{-1}\mathbb{P}(G).$$

[5]The definition expresses $c(x)$ as the supremum of affine functions $a_n x + b_n$ for sequences a_n and b_n, which correspond to the subgradients $D_{-c}(q_n)$ and values $c(q_n)$ at points q_n. This means that $c(x)$ can be constructed as the pointwise upper bound of all linear approximations of the function based on its subgradients at various points. Intuitively, each linear function $a_n x + b_n$ is a "supporting hyperplane" to the convex function at a specific point, and the supremum ensures that $c(x)$ is the tightest convex function that satisfies these inequalities.

[6]The countability property allows the supremum over all $n \in \mathbb{N}$ to replace the fixed n case, ensuring the inequality applies to the entire sequence.

[7]**this is taken from the glorious book that was recommended in class**

If Z is the indicator function of a set $A \in \xi$, then by the definition of conditional expectation, the result holds:

$$\mathbb{E}[\mathbf{1}_A Y \mid G] = \mathbf{1}_A \mathbb{E}[Y \mid G].$$

Linearity of the conditional expectation implies that the result holds for any simple function $Z \in SF^+(\Omega, \xi, \mathbb{P})$. By the monotone convergence theorem[8], the result extends to non-negative, ξ-measurable random variables. Finally, for a general bounded ξ-measurable random variable Z, we decompose Z into positive and negative parts and apply linearity (proof given above). Thus, the property is proven. □

Remark 1.7. If $z \in L^p(\Sigma, \xi, \mathbb{P})$ and $y \in L^q((\Sigma, \mathcal{F}, \mathbb{P}))$ where $\xi \subset \Sigma$ and $\frac{1}{p} + \frac{1}{q} = 1$ and $p, q \geq 1$ then $\mathbb{E}[zy|\xi] = z\mathbb{E}[y|\xi]$ a.s.

This is an application of the **Hühler inequality**, which is defined as follows: $\mathbb{E}[|zy|] \leq \mathbb{E}[|z|^p]^{\frac{1}{p}} \mathbb{E}[|y|^q]^{\frac{1}{q}}$ and this is defined $\forall p, q \geq 0 : \frac{1}{p} + \frac{1}{q} = 1$[9]

1.3 Some notes on identification (from lecture 2)

Data consists of draws from $\mathbb{P}_{y,x}$ where y and x are vectors that are observed by the econometrician. X is determined outside the model. **A structural econometrics model** consists of:

1. An equation $v(y, \gamma, X, \epsilon; \zeta) = 0$ where v is a vector of functions, γ is a vector of unobserved variables determined inside the model, ϵ unobserved variables determined outside the model and ζ which corresponds vector's function distribution??? **check notes because this makes little sense to me**

2. Restriction: $\zeta \in \mathcal{R}$: for example, then our draw would be $\mathbb{P}_{y,x;\zeta}$ is the distribution of observables generated by ζ.

Example 1.4. 1. $Y = f(X) + \varepsilon$, where $\Sigma = \{f, P_X, \varepsilon\}$. This is a type of regression.

2. $\mathbb{E}(|f(X)| + |\varepsilon|) < \infty$, and $\mathbb{E}[\varepsilon \mid X] = 0$, with f continuous on $\text{supp}(X)$.

   In this case, we have:

   $$\mathbb{E}[Y \mid X] = \mathbb{E}[f(X) \mid X] + \mathbb{E}[\varepsilon \mid X] = f(X),$$

   where f is the conditional expectation function (i.e., the regression function).

Example 1.5. 1. $Y = f(X) + \varepsilon$, where $\Sigma = \{f, P_{X,Z,\varepsilon}\}$, and the distribution of the observed data is $P_{Y,X,Z}$. Here, Z is determined outside the model.

2. $\mathbb{E}(|f(X)| + |\varepsilon|) < \infty$, $\mathbb{E}[\varepsilon \mid Z] = 0$, and f is continuous on $\text{supp}(X)$, with additional restrictions.

Example 1.6. We have these two equations:

1. $Y_1 = \alpha_{12} Y_2 + g_1(X_1) + \varepsilon_1$,

2. $Y_2 = \alpha_{21} Y_1 + g_2(X_2) + \varepsilon_2$,

where:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix},$$

and $\alpha_{12}, \alpha_{21} \in \mathbb{R}$, while g_1 and g_2 are unknown functions.

The model can be written as:

$$\gamma(Y, \gamma, X, \varepsilon, \zeta) = 0,$$

[8] $if X_n \geq 0$ then if $\mathbb{E}(X_n|\zeta)$ increases, $\mathbb{E}(X_n|\zeta)$ also increases.

[9] when p=q=2 this is the Cauchy-Schwarz inequality.

where $\zeta = \begin{pmatrix} \alpha_{12} \\ \alpha_{21} \\ g_1 \\ g_2 \\ P_{X,\varepsilon} \end{pmatrix}$ and explicitly:

$$\gamma = \begin{pmatrix} \alpha_{12}Y_2 + g_1(X_1) + \varepsilon_1 - Y_1 \\ \alpha_{21}Y_1 + g_2(X_2) + \varepsilon_2 - Y_2 \end{pmatrix} = 0$$

Case 2: If α_{12} and α_{21} are random, then we can assume ε satisfies one of the following restrictions:

$$\mathbb{E}[\varepsilon \mid X] = 0, \quad \mathbb{E}[\Sigma \mid X] = 0, \quad \text{or} \quad \varepsilon \perp X.$$

And now the parameters are where $\zeta = \begin{pmatrix} g_1 \\ g_2 \\ P_{X,\varepsilon} \end{pmatrix}$ $\varepsilon = \begin{pmatrix} a_{12} \\ g_{21} \\ \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$

Case 3: If $\gamma = \begin{pmatrix} \alpha_{12} \\ \alpha_{21} \end{pmatrix}$, we can be in a setup where the distribution of the observed data is $P_{Y,X,Z}$ and the assumptions $Z \perp Y$ hold in R.

Example 1.7. This is an example of non-parametric IV that was used by the professor, with not much success, to introduce and convey in us a passion for the assumption of completeness:

1. $Y_1 = f(x) + \epsilon$ and the distribution of the observed data is $\mathbb{P}_{y,x,z}$. We are interested in $\varphi = f$.
2. Restrictions:
   - f is continuous on the $supp(x)$.
   - $\mathbb{E}\left[|f(x)| + |\epsilon|\right] < \infty$
   - $\mathbb{E}[\epsilon|z] = 0$
   - **Completeness**: if φ is continuous in the $supp(x)$ s.t $\mathbb{E}|\varphi(x)| < \infty$ then $\mathbb{E}(\varphi(x)|z) = 0 \implies x \in \text{supp}(x), \varphi(x) = 0$.

Example 1.8. Exercise 4.1 $(x', z')'$ has a finite support such that supp(x)=$\{x_1, \cdots, x_k\}$ and supp(z)=$\{z_1, \cdots, z_l\}$ then show that $\mathbb{E}\left[\varphi(x)|z = z_l\right] = 0$.
We can define $\mathbb{E}\left[\varphi(x)|z = z_l\right] = \sum_{k=1}^{k} \varphi(x_k)\mathbb{P}\left(x = x_k|z = z_l\right)$. This is equal to 0 iff $\varphi \cdot M_{k,l} = 0$, we can apply the rank theorem[10] to show that this holds only if the k vector φ is equal to 0.

Exercise 4.2 In this exercise we are interested in finding a sufficient condition under which completeness holds. **To complete in another occasion**

Example 1.9. Let x, β be two random variables in $\mathbb{R}^d$ s.t $x \perp \beta$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be continuous and bounded. We define $y = g(x'\beta)$, we need to show that $\mathbb{E}(y|X = x) = \mathbb{E}[g(x'\beta)]$. For simplicity we define $\varphi(x) = g(x'\beta)$. The argument needs to proof these two points:

1. $\varphi(x) \in L^1((\Sigma, \mathcal{F}, \mathbb{P}))$. The function is x-measurable.
2. Let $B \in \mathbb{B}(\mathbb{R}^d)$, then $\mathbb{E}\left[y\mathbf{1}\left\{x \in B\right\}\right] = \mathbb{E}\left[\mathbb{E}\left[g(x'\beta)\right]\mathbf{1}\left\{x \in B\right\}\right]$.

To verify the first point, by definition:

$$\varphi(x) = \int_{\mathbb{R}^d} g(x'\beta)\, d\mathbb{P}_\beta(b).$$

Since g is continuous and bounded, $g(x'\beta)$ is continuous as a function of x. The integral of a continuous function over β is also continuous in x. This implies that $\varphi(x)$ is a continuous function of x, and thus $\varphi^{-1}(C)$

[10] $\dim\ker(\mu) + \text{rank}(\mu) = k$

is in $\mathbb{B}(\mathbb{R}^d)$. Hence, $\varphi(x)$ is measurable.

For the second part, consider $\mathbb{E}_{x,\beta}[g(x'\beta)\mathbb{1}\{x \in B\}]$, where $B \in \mathbb{B}(\mathbb{R}^d)$. By the definition of joint expectation:

$$\mathbb{E}[g(x'\beta)\mathbb{1}\{x \in B\}] = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(x'\beta)\mathbb{1}\{x \in B\}\, d\mathbb{P}_x(x)\, d\mathbb{P}_\beta(b).$$

Here, the independence of x and β plays a critical role. Independence implies that the joint probability measure $d\mathbb{P}_{x,\beta}(x, b)$ can be written as $d\mathbb{P}_x(x) \cdot d\mathbb{P}_\beta(b)$. This factorization allows us to separate the expectation as:

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(x'\beta)\mathbb{1}\{x \in B\}\, d\mathbb{P}_x(x)\, d\mathbb{P}_\beta(b).$$

Using Fubini's theorem[11], we can swap the order of integration:

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(x'\beta)\mathbb{1}\{x \in B\}\, d\mathbb{P}_\beta(b)\, d\mathbb{P}_x(x).$$

The inner integral depends only on β, and the outer integral captures the expectation over x.The outer integral then accounts for the distribution of x, weighted by $\mathbb{1}\{x \in B\}$

By evaluating the inner integral, it follows that:

$$\mathbb{E}[g(x'\beta)\mathbb{1}\{x \in B\}] = \mathbb{E}[\mathbb{E}[g(x'\beta)]\mathbb{1}\{x \in B\}],$$

which completes the proof.

1.3.1 Definitions of Identification

We often care about $\Psi^* = \Psi(\zeta)$ for a certain $\Psi : \mathcal{R} \to \zeta$.

Remark 1.8. $\Gamma_{Y,X}(\Psi, \mathcal{R}) = \{P_{Y,X,\zeta} \mid \zeta \in \mathcal{R},\ \Psi(\zeta) = \Psi^*\}$. It is the set of all distributions of the observables generated by ζ consistent with $\mathcal{R}$ and Ψ.

Definition 1.14. Let $\Psi, \Psi' \in \zeta$. Ψ and Ψ' are directionally equivalent if: $\Gamma_{Y,X}(\Psi, \mathcal{R}) \cap \Gamma_{Y,X}(\Psi', \mathcal{R}) \neq \emptyset$.

Definition 1.15. Ψ^* is identified in model $(\mathcal{R})$ if:

$$\forall \Psi \in \zeta, \quad \Gamma_{Y,X}(\Psi, \mathcal{R}) \cap \Gamma_{Y,X}(\Psi^*, \mathcal{R}) \neq \emptyset \implies \Psi = \Psi^*.$$

1.3.2 Completeness (from class 4)

Supose (x, z) has a density and we define $x = x - \eta$ where $z \perp \eta$ and z and η have densities. We may think of x as miss-measured. Completeness in this situation can be defined as: $\forall \varphi$ continuou such that $\forall z \in \mathbb{R}$:

$$\int f_\eta(z - x)\varphi(x)dx = 0 \implies \forall x\ \varphi(x) = 0$$

.

Example 1.10. Example (Economic Context):
Suppose we are interested in estimating a structural model where the true relationship is:

$$y = \beta x + \epsilon,$$

but the observed covariate $z = x + \eta$ is a mismeasured version of x, where η represents classical measurement error. Here, z is the observed variable, and x and η are unobserved. To identify β, we often rely on an instrumental variable w that satisfies the following conditions: 1. $w \perp \epsilon$ (instrument exogeneity), 2. $w \not\perp x$ (relevance), 3. $w \perp \eta$ (independence from the measurement error).

[11]This step is valid because the integrand $g(x'\beta)\mathbb{1}\{x \in B\}$ is bounded and the measures $\mathbb{P}_x$ and $\mathbb{P}_\beta$ are finite.

Completeness plays a role in ensuring that x can be uniquely identified from the joint distribution of (z, w). Specifically, the completeness assumption can be formulated as follows:

Let $\varphi(x)$ be any measurable function such that:

$$\mathbb{E}[\varphi(x) \mid z, w] = 0 \quad \text{a.s.}$$

Completeness implies that $\varphi(x) = 0$ almost surely. This ensures that no non-trivial function of x is orthogonal to all functions of (z, w), thereby allowing x to be fully recovered from the observed data (z, w).

Why is this important?

Without completeness, the conditional expectation $\mathbb{E}[y \mid z, w]$ might not provide unique information about x, leading to non-identification of the structural parameter β. Completeness guarantees that the instrument w and the observed covariate z together contain enough information to uniquely identify the true signal x.

Connection to Measurement Error:

In the classical measurement error setting, we have:

$$z = x + \eta,$$

where $\eta \perp x$. The density $f_\eta(z - x)$ "smooths out" the true x, making it harder to identify without additional information (e.g., from an IV). Completeness ensures that the mapping $M : \varphi \mapsto \left(z \mapsto \int f_\eta(z - x)\varphi(x)dx\right)$ is injective, preventing loss of information about x.

Remark 1.9. Completeness $\iff$ M is injective, where $M : \varphi \mapsto x \mapsto \int f_\eta(z - x)\varphi(x)dx$. f is injective if $f(x) = f(y) \implies x = y$. If f is linear, checking $f(0) = 0$ is enough to check $f(x) = 0 \implies x = 0$

Remark 1.10. What is the density of $z + \eta$ when $z \perp \eta$ and z and η have a density? Let $B \in \mathbb{B}(\mathbb{R})$ then $\mathbb{P}(z + \eta) \in B = \int_{\mathbb{R}^2} \mathbb{1}(z + \epsilon \in B)\, d\mathbb{P}_{z,\eta}$, which by independence assumption is the same as: $\int_{\mathbb{R}^2} \mathbb{1}(z + \epsilon \in B)\, f_z(z) f_\eta(\eta) dz d\eta$. If we let $u = z + \epsilon$ then we can express this integral as: $\int_{\mathbb{R}^2} \mathbb{1}(u \in B)\, f_\eta(u - z) f_z(z) du dz$ where we have also applied Fubini to reverse the order of integration. Finally, we can rewrite the last expression as this: $\int_{\mathbb{R}^2} \mathbb{1}(u \in B)\, int_{\mathbb{R}^2} f_\eta(u - z) f_z(z) dz du$ where $int_{\mathbb{R}^2} f_\eta(u - z) f_z(z) = f_{z+\eta}(u)$ by definition.

Definition 1.16. Characteristic Function: The characteristic function of a random vector x is defined as:

$$\phi_x(t) = \mathbb{E}\left[\exp(it'x)\right],$$

where $i^2 = -1$ and $\exp(it) = \cos(t) + i \sin(t)$.

Applying the characteristic function to the previous exercise, we find:

$$\mathbb{E}\left[\exp(it(z + \eta))\right] = \mathbb{E}\left[\exp(itz)\right] \cdot \mathbb{E}\left[\exp(it\eta)\right],$$

because z and η are independent. This simplifies a potentially complex integral into a product of two simpler terms, making it significantly easier to work with.

The characteristic function is useful because it relates expectations to the underlying distributions. Specifically:

$$\mathbb{E}\left[\exp(itx)\right] = \mathbb{E}\left[\exp(itz)\right] \varphi(t).$$

This allows us to estimate:

$$\mathbb{E}[\exp(itz)] = \frac{\hat{\exp}(itx)}{\hat{\varphi}(t)}.$$

This relationship demonstrates how characteristic functions can simplify inference problems, particularly when working with convolution or noisy data.

Note: This is the last example where convolution and characteristic functions are specifically useful in our econometric applications.

Example 1.11. Consider the following model:

1. $y_i = x_i'\beta_i + \epsilon_i$,
2. $x_i \perp (\beta_i', \epsilon_i)$.

One feature of interest in this model is f_{β_i}, the distribution of β_i (e.g., to study the distributional effects of a treatment). The distribution f_{β_i} can be expressed as:

$$f_{\beta_i} = \int_{\mathbb{R}^{d-1}} f_\beta(b)\, db_2 \cdots db_d,$$

where d is the dimension of β_i.

Alternatively, this distribution can be estimated using the characteristic function. For a given $t \in \mathbb{R}$, we have:

$$\mathbb{E}\left[\exp(itx) \mid X = x\right] = \mathbb{E}\left[\exp\left(i(\beta' tx + t\epsilon)\right)\right].$$

This equality follows because $\mathbb{E}\left[\exp(itx) \mid X = x\right]$ is the characteristic function of $(\beta', \epsilon)'$ evaluated at (tx, t).

Exercise: As a suggested exercise, prove the equality above. Hint: Consider how the independence assumption $x_i \perp (\beta_i', \epsilon_i)$ simplifies the decomposition of the expectation. Good luck!

Definition 1.17. $L^p(\mathbb{R}^d) = \left(\text{measurable functions from } (\mathbb{R}^d, B(\mathbb{R}^d)) \text{ to } (\mathbb{R}, B(\mathbb{R}))\right) : \int_{\mathbb{R}^d} |f(x)|dx < \infty$

Remark 1.11. we can also have the following mapping: $(\mathbb{R}^d, B(\mathbb{R}^d)) \mapsto (\mathbf{C}, B(\mathbf{C}))$

Definition 1.18. If $f \in L^1(\mathbb{R}^d)$ the Fourier transform of f $F[f]$ is obtained as $\forall \omega \in \mathbb{R}^d$, $F(f)(\omega) = \int_{\mathbb{R}^d} \exp(iw'x) f(x) dx$.

Definition 1.19. If f and g belong to $L^1(\mathbb{R}^d)$, then the **convolution** of f g, denoted $(f * g)(z) = \int_{\mathbb{R}^d} f(x) g(z - x) dx$.

Remark 1.12. The previous expression is also equal to $\int_{\mathbb{R}^d} f(z - x) g(x) dx$ by change of variables.

Proposition 1.1. *If f and g belong to $L^1(\mathbb{R}^d) \implies (f * g)(z) \in L^1(\mathbb{R}^d)$ and $F[f * g](\omega) = F[f](\omega) F[g](\omega)$*

Proposition 1.2. *if f belong to $L^1(\mathbb{R}^d) \implies F[f]$ is continuous at $\lim_{||\omega||_2 \to \infty} F(f)(\omega) = 0$*

Example 1.12. Gaussian Kernel: let $k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ then given that $k \in L^1(\mathbb{R}^d)$ and $F[k](\omega) = exp\left(-\frac{\omega^2}{2}\right)$ is positive

Example 1.13. Box function: Let $k(x) = \frac{1}{2}\mathbb{1}[|x| \leq 1]$ then if $k \in L^1$ and $F[k](\omega) = \frac{1}{2}\int_{-1}^{1} \cos(wx) dx + \frac{i}{2}\int_{-1}^{1} sin(wx) dx$. Where the second integral is 0 and so this becomes: $\frac{1}{2\omega}\left[sin(wx)\right]_{-1}^{1}$.

1.3.3 Properties of the Fourier transform

Suppose f and $F[f]$ belong to L^1 and to L^2[12]. Then, the following properties apply:

1. **Plancherel equality:** $\frac{1}{(2\pi)^d}||F(f)||_2^2 = ||f||_2^2$ by Cauchy-Schwarz.
2. **Fourier's inversion formula:** $\forall x \in \mathbb{R}^d$, $f(x) = \frac{1}{(2\pi)^d} F\left[[F(f)](-x)\right] = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-iw'x} F[f](w) dw$

[12]Important to notice that if we talk about r.v then if $x \in L^2 \subseteq L^1$. This does not apply to functions belonging in L^1.

1.4 class 5

Some more notes to motivate completeness, as a titanic effort made by the professor to generate some enthusiasm in class.

Proposition 1.3. *Let φ be a continuous function such that $\varphi \in L^1$, $x = z - \eta$, $z \perp \eta$, and we assume that η has a density. Then $\forall z \in \mathbb{R}\ \int_{\mathbb{R}} \varphi(x) f_\eta(z-x)dx = 0 \implies \forall x \in \mathbb{R}\ \varphi(x) = 0$.*

This result shows that the interaction between $\varphi(x)$ and $f_\eta(x)$ vanishes entirely for all z. The convolution integral being zero everywhere implies that $\varphi(x)$ has no "effect" in the domain, forcing it to be identically zero. Completeness ensures no non-zero $\varphi(x)$ can "hide" under such conditions.

Also, $\int_{\mathbb{R}} \varphi(x) f_\eta(z-x)dx = 0 \implies F[\varphi](\omega)F[f_\eta](\omega) = F[0] = 0\ \forall\omega$ by properties of the Fourier transform defined above.

By the convolution theorem, the Fourier transform of the convolution is the product of the transforms. If this product is zero, and $F[f_\eta](\omega) \neq 0$, then $F[\varphi](\omega)$ must be zero for all ω. This directly implies $\varphi(x)$ is zero everywhere, given the Fourier transform uniquely determines a function.

Example 1.14. f_η is the density of a mean-zero random variable. Then $[f_\eta](\omega) > 0\ \forall\omega$.

The Fourier transform of a mean-zero density does not vanish anywhere, ensuring that any $\varphi(x)$ convolved with $f_\eta(x)$ will reflect entirely in the product $F[\varphi](\omega)F[f_\eta](\omega)$. If $F[\varphi](\omega) = 0$, this enforces that $\varphi(x)$ is zero everywhere.

This implies that $\forall\omega\ F[\varphi](\omega) = 0 \implies F[\varphi] = 0$. So $\varphi' = 0$ almost everywhere. But given that $\varphi(x)$ is continuous, then $\varphi(x) = 0$.

Example 1.15. $F[f_\eta](w) = \frac{\sin(w)}{w}$, $\forall\omega \in \mathbb{R}$. Here we also have that $\eta \sim U[-1,1]$.

Intuition: For the uniform distribution on $[-1,1]$, the Fourier transform $F[f_\eta](\omega)$ has zeros at $\omega = k\pi$ for $k \in \mathbb{Z} \setminus \{0\}$, corresponding to points where $\sin(w) = 0$.

Let $\zeta = [f_k, k \in \mathbb{Z}]$ be the set of zeros of $F[f_\eta]$. Then by the second property of the previous theorem, $\forall\omega \in \mathbb{R} \setminus \zeta$, $F[\varphi](w) = 0$.

Outside the zeros of $F[f_\eta](\omega)$, the Fourier transform forces $F[\varphi](\omega) = 0$. Continuity of $F[\varphi](\omega)$ ensures this behavior extends across all ω.

Or, similarly, let $\omega = k\pi$ for some $h \in \mathbb{Z}$ and $(\omega_n)_{n\in\mathbb{N}}$ a sequence in $\mathbb{R} \setminus \zeta$ such that $\omega_n \to \omega$. Because $F[\varphi]$ is continuous, then $\lim_{n\to\infty} F[\varphi](\omega_n) = F[\varphi](\omega)$. Hence $\forall\omega\ F[\varphi] = 0$, and we can conclude using a similar argument as before that $\varphi(x) = 0$.

This shows that the zeros of $F[f_\eta]$ do not allow $\varphi(x)$ to escape the conclusion of being identically zero. Continuity bridges gaps at points of zero, solidifying the result.

2 Non-Parametric Estimation

2.1 Density estimation (Lebesgue)

Let $x_1, ..., x_n$ be identically distributed random variables. Imagine we want to estimate the density f_x of x which we assume to lie in a large class of densities.

Example 2.1. We have two examples:

1. Densities such that $\forall x, x' \in \mathbb{R}$, $|f_x(x) - f_x(x')| \geq L|x - x'|$ for a certain $L > 0$.
2. Densities which are monothically increasing on $[0, 1]$

A nonparametric estimator is the histogram. Assume $f_x : [0, 1] \mapsto \mathbb{R}$ and divide $[0, 1]$ into non-overlapping intervals that we denote $(A_j)_{j=1}^m$ of length h. We define $n_j = \sum_{i=1}^n \mathbb{1}\{ X_i \in A_j\}$. Then the histogram estimator is just $f_x(x) = \sum_{j=1}^m \frac{n_j}{n \cdot h} \mathbb{1}\{ x \in A_j\}$ which depends on h (we want the smaller h as possible...). Then we also have this expression: $f_x(x) = \sum_{j=1}^m \frac{n_j}{n \cdot h} \mathbb{1}\{ x \in A_j\} = \frac{1}{n \cdot h} \sum_{i=1}^n \sum_{j=1}^m \mathbb{1}\{ x_i \in A_j; X \in A_j\}$

Definition 2.1. Kernel Density estimator Let $K \in L^1(\mathbb{R})$ s.t $\int_{\mathbb{R}} K(u)du = 1$ and $\hat{f}_x(x) = \frac{1}{nh} \sum_{i=1}^m K\left(\frac{X_i - x}{h}\right)$ is called a kernel density estimator with kernel K and bandwidth h.[13]

Example 2.2. We can start with an estimator of the CDF. Let's for example define F_x of x as $F_x(x) = \mathbb{E}[\mathbb{1}\{ X_i \leq x\}]$, then $\hat{F}_x(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{ X_i \leq x\}$. We define f_x as the pdf. $\hat{f}_x(x) = \frac{\hat{F}_x\left(x + \frac{h}{2}\right) - \hat{F}_x\left(x - \frac{h}{2}\right)}{h}$ for a given $h > 0$. This is equal to this expression: $\frac{1}{nh} \sum_{i=1}^n \mathbb{1}\left\{ x - \frac{h}{2} < X_i \leq x + \frac{h}{2}\right\}$. This function is a generalization of the Rosenblatt estimator. For example, we have the following expression:

Remark 2.1. If we take expectation of the Kernel function we have the following expression: $\mathbb{E}(\hat{f}_x(x)) = \frac{1}{hn} \sum_{i=1}^n \mathbb{E}\left(\left(\frac{x_i - x}{h}\right)\right) = \frac{1}{h} \mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)\right) = \frac{1}{h} \int_{\mathbb{R}} \left(\frac{y - x}{h}\right) f_x(y)dy$

2.1.1 Mean squared error of kernel estimators

This definition is taken directly from the book by Tsybakov A basic measure of the accuracy of estimator $\hat{p}_n$ is its *mean squared risk* (or *mean squared error*) at an arbitrary fixed point $x_0 \in \mathbb{R}$:

$$\mathrm{MSE} = \mathrm{MSE}(x_0) \triangleq \mathbb{E}_p \left[\left(\hat{p}_n(x_0) - p(x_0) \right)^2 \right].$$

Here, MSE stands for "mean squared error" and $\mathbb{E}_p$ denotes the expectation with respect to the distribution of $(X_1, \ldots, X_n)$:

$$\mathbb{E}_p \left[\left(\hat{p}_n(x_0) - p(x_0) \right)^2 \right] \triangleq \int \cdots \int \left(\hat{p}_n(x_0, x_1, \ldots, x_n) - p(x_0) \right)^2 \prod_{i=1}^n \left[p(x_i) dx_i \right].$$

We have

$$\mathrm{MSE} = b^2(x_0) + \sigma^2(x_0) \tag{1.4}$$

where[14]

$$b(x_0) = \mathbb{E}_p[\hat{p}_n(x_0)] - p(x_0)$$

and

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

[13] Some examples of Kernels are given in the class notes, but are omitted here for brevity.

[14] this is found by plugging in $\mathbb{E}(\hat{f}_n)$ inside the first equations and developing.

Definition 2.2. *The quantities $b(x_0)$ and $\sigma^2(x_0)$ are called the bias and the variance of the estimator $\hat{p}_n$ at a point x_0, respectively.*

We will develop here the same reasoning that was developed in the TD.

Bounds of the bias
We have the following expression for the bias:

$$b(x_0) = \mathbb{E}\left(\hat{f}'_n(x_0)\right) - f'(x_0).$$

Using the definition of the kernel estimator, we write:

$$b(x_0) = \mathbb{E}\left[-\frac{1}{nh^2}\sum_{i=1}^{n} k'\left(\frac{x_i - x_0}{h}\right)\right] - f'(x_0).$$

Since the observations x_i are i.i.d., the expression simplifies to:

$$b(x_0) = -\frac{1}{h^2}\mathbb{E}\left[k'\left(\frac{x_i - x_0}{h}\right)\right] - f'(x_0).$$

Now, applying the definition of expectation, we have:

$$b(x_0) = \int_{\mathbb{R}} -\frac{1}{h^2}k'\left(\frac{z - x_0}{h}\right) f(z)\, dz - f'(x_0).$$

Using integration by parts for the term $\int_{\mathbb{R}} k'\left(\frac{z-x_0}{h}\right) f(z)\, dz$, we recall the integration by parts formula:

$$\int u\, dv = uv\Big|_{-\infty}^{+\infty} - \int v\, du,$$

and set: $u = f(z)$, so $du = f'(z)\, dz$, $dv = k'\left(\frac{z-x_0}{h}\right) dz$, so $v = k\left(\frac{z-x_0}{h}\right) \cdot \frac{1}{h}$.

Thus, the integral becomes:

$$\int_{\mathbb{R}} k'\left(\frac{z - x_0}{h}\right) f(z)\, dz = \left[f(z) \cdot k\left(\frac{z - x_0}{h}\right) \cdot \frac{1}{h}\right]_{-\infty}^{+\infty} - \int_{\mathbb{R}} k\left(\frac{z - x_0}{h}\right) \cdot \frac{1}{h} f'(z)\, dz.$$

Substituting this result back:

$$b(x_0) = -\frac{1}{h^2} \cdot h \cdot f(z) \cdot k\left(\frac{z - x_0}{h}\right)\Big|_{-\infty}^{+\infty} + \frac{1}{h^2} \cdot h \int_{\mathbb{R}} k\left(\frac{z - x_0}{h}\right) f'(z)\, dz - f'(x_0).$$

The first term vanishes because the kernel k and the function $f(z)$ are assumed to decay sufficiently fast at infinity (they are integrable), i.e.,

$$\lim_{z \to \pm\infty} f(z) \cdot k\left(\frac{z - x_0}{h}\right) = 0.$$

Thus, the bias simplifies to:

$$b(x_0) = \frac{1}{h^2} \cdot h \int_{\mathbb{R}} k'\left(\frac{z - x_0}{h}\right) f'(z)\, dz - f'(x_0).$$

Now we define $\mu = \frac{z-x_0}{h}$, so that $z = \mu h + x_0$ and $dz = h \cdot d\mu$. Replacing these in the previous expression, we obtain:

$$b(x_0) = \frac{1}{n} \int_{\mathbb{R}} h \cdot k(\mu) \cdot f'(\mu h + x_0)\, d\mu - f'(x_0).$$

By assumption, f is l-time differentiable. Hence, we can use the $(l-1)$ Taylor expansion for $f'(\mu h + x_0)$ around x_0, we write:

$$f'(\mu h + x_0) = f'(x_0) + f''(x_0)\mu h + \frac{f'''(x_0)}{2!}(\mu h)^2 + \cdots + \frac{f^{(l)}(x_0 + \tau\mu h)}{(l-1)!}(\mu h)^{l-1},$$

where $\tau \in (0, 1)$. Substituting this expansion into the integral gives:

$$b(x_0) = \int_{\mathbb{R}} k(\mu) \left[f'(x_0) + f''(x_0)\mu h + \frac{f'''(x_0)}{2!}(\mu h)^2 + \cdots + \frac{f^{(l)}(x_0 + \tau\mu h)}{(l-1)!}(\mu h)^{l-1} \right] d\mu - f'(x_0).$$

Using the property of the kernel function that $\int_{\mathbb{R}} \mu^s k(\mu)\, d\mu = 0$ for all $s = 1, \ldots, l-1$ (where l is the order of the kernel), we simplify the expression to:

$$b(x_0) = \int_{\mathbb{R}} k(\mu) \cdot \frac{f^{(l)}(x_0 + \tau\mu h)}{(l-1)!}(\mu h)^{l-1}\, d\mu.$$

Now we introduce the Hölder condition for the function f, which holds by assumption. A function f belongs to the Hölder class of order (l) if:

$$|f^{(l)}(x) - f^{(l)}(x')| \leq L|x - x'|^{\beta - l},$$

which conveys that the norm is bounded and provides information on the smoothness of the function. Using this, we write:

$$b(x_0) = \int_{\mathbb{R}} k(\mu) \left[f^{(l)}(x_0 + \tau\mu h) - f^{(l)}(x_0) \right] \cdot \frac{(\mu h)^{l-1}}{(l-1)!}\, d\mu.$$

By the triangle inequality, we have:

$$\leq \int_{\mathbb{R}} |k(\mu)| \cdot \left| f^{(l)}(x_0 + \tau\mu h) - f^{(l)}(x_0) \right| \cdot \frac{(\mu h)^{l-1}}{(l-1)!}\, d\mu.$$

By the Hölder condition:

$$\leq L(\tau\mu h)^{\beta - l}.$$

Thus:

$$b(x_0) < \int_{\mathbb{R}} |k(\mu)| \cdot L(\mu h)^{\beta - l} \cdot \frac{(\mu h)^{l-1}}{(l-1)!}\, d\mu.$$

Finally, we simplify:

$$b(x_0) = C_1 \cdot h^{\beta - 1},$$

where C_1 is a constant depending on the kernel function and L.

Bound of the Variance

This is somewhat easier to show. Again, we start with the initial definition of the variance:

$$\text{Var} = \mathbb{E}\left[\left(\hat{f}'_n(x) - \mathbb{E}(\hat{f}'_n(x)) \right)^2 \right].$$

Using the definition of the kernel derivative estimator, this becomes:

$$\text{Var} = \frac{1}{n}\mathbb{E}\left[\left(\frac{1}{h^2} k'\left(\frac{x_i - x}{h} \right) - \frac{1}{h^2}\mathbb{E}\left(k'\left(\frac{x_i - x}{h} \right) \right) \right)^2 \right].$$

Applying the properties of the variance[15], we can bound the variance as:

[15] $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$

$$\text{Var} \leq \frac{1}{n}\mathbb{E}\left[\left(\frac{1}{h^2}k'\left(\frac{x_i - x}{h}\right)\right)^2\right].$$

Now, using the definition of expectation, we write:

$$\text{Var} \leq \frac{1}{n}\int_{\mathbb{R}} \frac{1}{h^4}\left[k'\left(\frac{z-x}{h}\right)\right]^2 f(z)\,dz.$$

As before, we make the substitution $\mu = \frac{z-x}{h}$, which implies $z = \mu h + x$ and $dz = h\,d\mu$. Substituting these into the integral gives:

$$\text{Var} = \frac{1}{n}\int_{\mathbb{R}} \frac{1}{h^4}\left[k'(\mu)\right]^2 f(\mu h + x)\cdot h\,d\mu.$$

Simplifying, we have:

$$\text{Var} \leq \frac{1}{nh^3}\int_{\mathbb{R}} \left[k'(\mu)\right]^2 f(\mu h + x)\,d\mu.$$

Next, we use the assumption that $f(z)$ is bounded by $f_{\max}$ for all $z \in \mathbb{R}$, so:

$$\text{Var} \leq \frac{1}{nh^3}\int_{\mathbb{R}} \left[k'(\mu)\right]^2 f_{\max}\,d\mu.$$

Factoring $f_{\max}$ outside of the integral:

$$\text{Var} \leq \frac{f_{\max}}{nh^3}\int_{\mathbb{R}} \left[k'(\mu)\right]^2 d\mu.$$

Finally, defining $C_2 = f_{\max}\int_{\mathbb{R}} \left[k'(\mu)\right]^2 d\mu$ as a constant dependent on $f_{\max}$ and the kernel function, we obtain:

$$\text{Var} = C_2\frac{1}{nh^3}.$$

Conclusion
Thus, the MSE is bounded by: Var+ Bias$^2 \leq C_2\frac{1}{nh^3} + C_1^2\cdot h^{2(\beta-1)}$.

2.1.2 Optimal h

How do we estimate the optimal h? To gain insight into the trade-off, we examine the upper bound on the mean squared risk. The variance decreases as h grows, whereas the bound on the bias increases (cf. Figure 4). The choice of a small h, corresponding to a large variance, is called *undersmoothing*. Alternatively, with a large h, the bias cannot be reasonably controlled, leading to *oversmoothing*. An optimal value of h that balances bias and variance is located between these two extremes.

Figure 2 illustrates three key scenarios in kernel density estimation:

- **Undersmoothing:** The bandwidth h is too small, resulting in a noisy estimate that overfits the sample points. This leads to high variance.
- **Oversmoothing:** The bandwidth h is too large, causing the estimator to oversimplify the density and miss important features such as modes. This leads to high bias.
- **Correct smoothing:** The bandwidth h is chosen optimally, balancing the trade-off between bias and variance. The estimate captures the essential structure of the density.

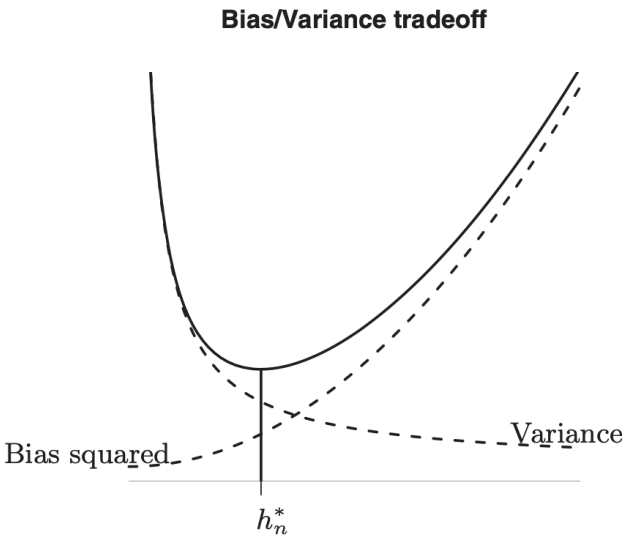

Figure 1.1. Squared bias, variance, and mean squared error (solid line) as functions of h.

Figure 1: Squared bias, variance, and mean squared error (solid line) as functions of h.

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

$$h_n^* = \left(\frac{3C_2}{2(\beta - 1)C_1^2} \right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

This differs from the result in the book, which presents:

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

16

However, it is important to note that the *rate of convergence* remains the same in both cases. Specifically, the MSE is given by:

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

uniformly in x_0.

Theorem 2.1. *Assume that condition (1.5) holds and the assumptions of Proposition 1.2 are satisfied. Fix $\alpha > 0$ and take $h = cn^{-\frac{1}{2\beta+1}}$. Then for $n \geq 1$, the kernel estimator p_n satisfies:*

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

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

[16]This is taken from the book, but we achieve different results because we are analyzing the derivative of the Kernel function.

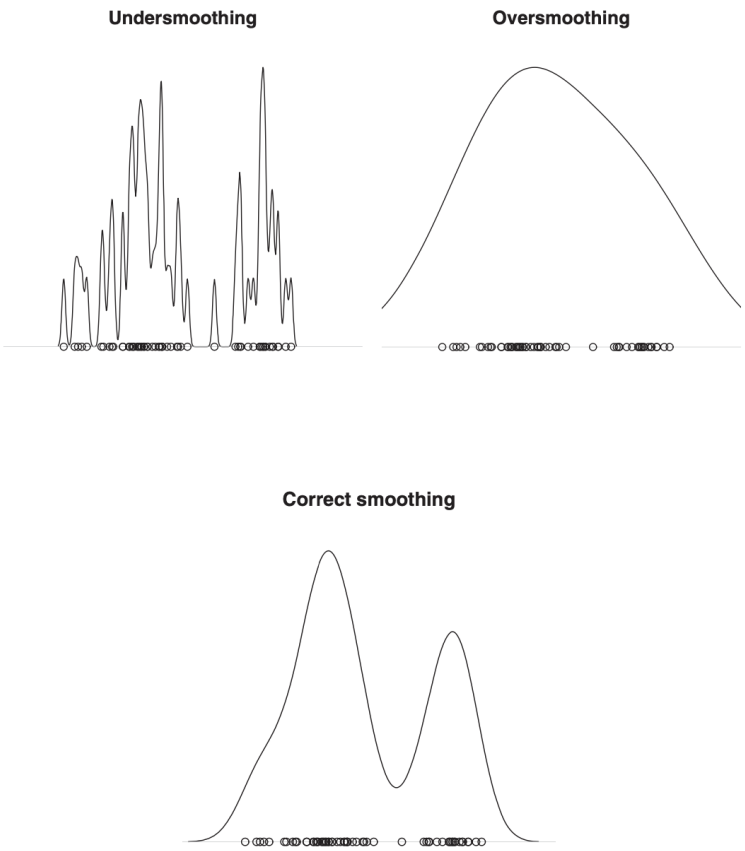

Figure 1.2. Undersmoothing, oversmoothing, and correct smoothing.
The circles indicate the sample points X_i.

Figure 2: Undersmoothing, oversmoothing, and correct smoothing. The circles indicate the sample points X_i.

Proof. Using previous equation and bounding the supremum norm of the kernel density estimate

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

where $p_{\max}$ is derived from boundedness properties of the kernel K and its derivatives □

2.1.3 Some definitions from class 6

These are definitions and theorems that were not used in the seminar. Most of the theorems, definitions and proofs given in class 6 were given by the great and informative TD so I omit their repetition.

Proposition 2.2. *Let $f(x) \in \mathbb{P}(\beta, l)$. With $\beta > 0$ and k of order $L > 0$ and $l = \lfloor\beta\rfloor$[17] such that $\int_{\mathbb{R}} |\mu|^\beta k(\mu)d\mu < \infty$ Then $\forall x \in \mathbb{R}$ and $\forall h > 0$ $|f(x)| \leq C_1 h^\beta$ where $C_1 := \frac{L}{!} \int_{\mathbb{R}} |\mu|^\beta k(\mu)d\mu$*

Theorem 2.3. *Assume k satisfies the assumptions of the previous proposition and of proposition xx* [18] *and*

[17] In this case, the floor function holds strictly.
[18] **I need a way to keep track of citations...add labels, lazy boy**

choose $h = \alpha n^{-\frac{1}{2\beta+1}}$ *for some* $\alpha \geq 0$ *then:*

$$\sup_{x\in\mathbb{R}} \sup_{f(x)\in\mathbb{P}(\beta,l)} \mathbb{E}\left[\left(\hat{f}_x(x) - f_x(x)\right)^2\right] \leq Cn^{\frac{-2\beta}{2\beta+1}}$$

Note that the more variables we have in the analysis the worse it gets...unbounded...

Remark 2.2. Note that the previous theorem can also be expressed as

$$\sup_{\hat{f}_x} \sup_{f(x)\in\mathbb{P}(\beta,l)} \mathbb{E}\left[\left(\hat{f}_x(x) - f_x(x)\right)^2\right] \geq Cn^{\frac{-2\beta}{2\beta+1}}$$

It is optimal up to a constant:

$$\sup_{\hat{f}_x} \sup_{f(x)\in\mathbb{P}(\beta,l)} \mathbb{E}\left[\left(\hat{f}_x(x) - f_x(x)\right)^2\right] n^{\frac{2\beta}{2\beta+1}} > 0$$

2.2 some notes on lecture 7: MISE

Previously we have studied the behavior of the Kernel function at an arbitrary fixed point x_0. Another measure of risk we can use is the *Mean integrated squared error* (MISE), where the expectation is defined over all the available data:

$$\textbf{MISE}(h) = \mathbb{E}\left[\int_{\mathbb{R}} \left(\hat{f}_x(x) - f_x(x)\right)^2 dx\right]$$

By the Tonelli-Fubini theorem and by the fact that we can express MSE as the sum of the vias and variance (as done before many times), we have this property:

$$\textbf{MISE} = \int \text{MSE}(x)dx = \int b^2(x)dx + \int \sigma^2(x)dx$$

To obtain a bound on these two objects the method is the same as before and is explained in pages 12-13 of Tsybakov.

Definition 2.3. Let $\beta > 0$ and $L > 0$. The Nikol'skiĭ class $\mathcal{H}(\beta, L)$ is defined as the set of functions $f : \mathbb{R} \to \mathbb{R}$ whose derivatives $f^{(\ell)}$ of order $\ell = \lfloor\beta\rfloor$ exist and satisfy

$$\left(\int \left(f^{(\ell)}(x+t) - f^{(\ell)}(x)\right)^2 dx\right)^{1/2} \leq L|t|^{\beta-\ell}, \quad \forall t \in \mathbb{R}.$$

Definition 2.4. Let $\beta \geq 1$ be an integer and $L > 0$. The Sobolev class $\mathcal{S}(\beta, L)$ is defined as the set of all $\beta - 1$ times differentiable functions $f : \mathbb{R} \to \mathbb{R}$ having absolutely continuous derivative $f^{(\beta-1)}$ and satisfying

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

In the context of density estimation using kernel methods, the smoothness of the true density function f plays a critical role in analyzing the Mean Integrated Squared Error (MISE). Sobolev functions provide a structured way to quantify this smoothness through derivatives and integrability conditions.

Example 2.3. Suppose we have $f_x \in L^2(\mathbb{R}^d)$ and $F[f_x] = P_x \in L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$

Theorem 2.4. *Let* $f_x \in L^2(\mathbb{R}^d)$ *and* $k \in L^2(\mathbb{R}^d)$ *be a symmetric kernel* ($k(x) = k(-x)$) *such that:*

$$\sup_{\omega\in\mathbb{R}^d \setminus (0)} \frac{|1 - F[k](\omega)|}{|\omega|^\beta} \leq A < \infty$$

Then

$$\sup_{f_x \in P_{\mathbb{S}}(\beta,L)} \mathbb{E}\left[||\hat{f}_x - f_x||_2^2\right] \leq C \cdot n^{\frac{-2\beta}{2\beta+d}}$$

Where C *is a constant that only depends on* L, α, K

Remark 2.3. We can also use the kernel such that

$$F[k](\omega) = \begin{cases} 0 & \text{if } |\omega| > a, \\ 1 & \text{if } |\omega| \leq b. \end{cases}$$

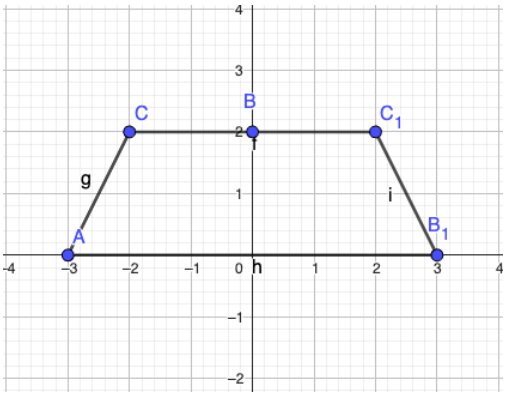

Figure 3: $F[k](w)$

This kernel acts as a simple filter in the frequency domain. It ensures that only frequencies within $[-b, b]$ are retained, while higher frequencies ($|\omega| > a$) are suppressed. Graphically, this corresponds to a rectangular function, where the flat top represents the frequencies that pass through unaltered ($|\omega| \leq b$), and the sides drop to zero, indicating suppression of higher frequencies.

This is the definition of a parallelogram. In fact, we can visualize it by drawing one using two orthonormal vectors with $b = 2$ and $a = 4$. Then, we define:

$$F[k_1] * F[k_1](\omega) = \int_{\mathbb{R}} \mathbf{1}\{|\omega - x| \leq 1\}\mathbf{1}\{|x| \leq 1\}\, dx = \int_{-1}^{1} \mathbf{1}\{\omega - 1 \leq x \leq \omega + 1\}\, dx = T_1(\omega).$$

The convolution $F[k_1] * F[k_1](\omega)$ measures the overlap between two shifted kernels. As ω varies, the interval of overlap changes, which leads to the triangular function $T_1(\omega)$. The graph of $T_1(\omega)$ (attached as the second image) shows this clearly: it peaks at the center and tapers off linearly, illustrating the nature of the overlap.

Finally, we define a recursive relation:

$$T_2(\omega) = \frac{1}{2}T_1(\omega - 1) + \frac{1}{2}T_1(\omega + 1).$$

This relation shifts and averages $T_1(\omega)$, effectively smoothing it further. Letting $k_2 = F^{-1}[T_2]$, and using the property that the Fourier transform of a convolution is the product of Fourier transforms, we derive:

$$F[T_2](\omega) = F[T_1](\omega)\cos(\omega),$$

where the cosine factor arises from the Fourier transform of the shift operator, using $\exp(i\omega) = \cos(\omega) + i\sin(\omega)$.

Combining this, we arrive at:

$$k_2(x) = \left(\frac{\sin x}{\pi x}\right)^2 \cos(\omega),$$

which describes the kernel in the spatial domain. This form combines the smoothing effect of the sinc function squared with the oscillatory behavior of the cosine factor.

IMPORTANT: Missing here is the crossvalidation exercise we did when computing the empirical counterpart of the MISE-check notes or ask professor to understand what we are doing there exactly.

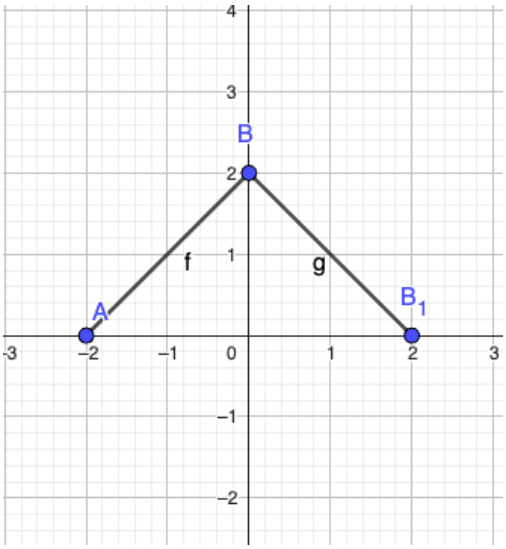

Figure 4: $T_1(w)$

Theorem 2.5. *Let $f_x(x) \leq f_{max} < \infty$; $\forall x; k \in L^2(\mathbb{R}^2) s.t F[k]$ is symmetric unimodal[19], and supp$[F[K]] \subset [-1,1]$ then $\exists C : \exists \gamma \in (0,1), \forall n > 1$, then:*

$$\mathbb{E}\left[||\hat{f}_x^{\hat{h}}(x) - f_x(x)||_2^2\right] \leq \left(1 + \frac{c}{n\gamma}\right) \min \mathbb{E}\left[||\hat{f}_x^{\hat{h}}(x) - f_x(x)||_2^2\right] + \frac{C(\log n)^{\frac{\gamma}{2}}}{n^{1-\gamma}}$$

And the last term of the sum is $\mathbb{O}_p$.

Corollary 2.6. *Let k be a sinc kernel, then* $\sup_{f_x \in \mathbb{P}(\beta,L)} \mathbb{E}\left[||\hat{f}_x^{\hat{h}}(x) - f_x(x)||_2^2\right] \leq C \cdot n^{\frac{-2\beta}{2\beta+d}}, \forall \beta > \frac{1}{2} and L > 0$[20]

2.3 Some notes on Lecture 8

2.3.1 Other Non-Parametric Estimators

Orthogonal Series Estimators

An orthonormal basis $\{e_j\}$ in $\mathbb{R}^d$ satisfies:

$$\forall j \neq l, \quad \langle e_j, e_l \rangle = 0, \quad \text{and} \quad \langle e_j, e_j \rangle = ||e_j||_2^2 = 1.$$

This means that the functions e_j are mutually orthogonal and have unit norm.

We equip $L^2(\mathbb{R}^d)$ with the inner product:

$$\langle f, g \rangle = \int_{\mathbb{R}^d} f(x)g(x)dx.$$

By a fundamental result in functional analysis, there exists an orthonormal basis $\{e_j\}$ of $L^2(\mathbb{R}^d)$ such that for any function $f \in L^2(\mathbb{R}^d)$:

$$\lim_{T \to \infty} \left\| f - \sum_{j=1}^{T} \langle f, e_j \rangle e_j \right\|_2 = 0.$$

[19] this means it has one maximum

[20] Here, d represents the dimension of the space where x is defined. As expected in nonparametric estimation, the convergence rate slows down as d increases due to the curse of dimensionality, which makes accurate estimation harder in higher-dimensional spaces. The exponent reflects the classical minimax rate for density estimation, demonstrating how smoothness (β) and dimensionality (d) jointly determine the speed at which the estimator improves with more data.

This means we can approximate f arbitrarily well by summing over a finite number of basis functions.

Now, if $f_x \in L^2(\mathbb{R}^d)$ is a probability density function, we compute its coefficients:

$$c_j = \langle f_x, e_j \rangle = \int_{\mathbb{R}^d} f_x(x) e_j(x) dx.$$

These coefficients can be estimated without bias using the empirical counterpart:

$$\hat{c_j} = \frac{1}{n} \sum_{i=1}^{n} e_j(x_i).$$

Thus, a natural estimator for $f_x(x)$ is:

$$\hat{f}_x(x) = \sum_{j=1}^{T} \hat{c_j} e_j(x).$$

Expanding this, we obtain:

$$\hat{f}_x(x) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{T} e_j(x_i) e_j(x).$$

Here, T determines the number of basis functions used. Choosing T correctly is crucial, as it controls the balance between bias (underfitting) and variance (overfitting).

Example 2.4. Assume $\text{supp}(x) \subset [0,1]^d$, then we can take the basis functions as:

$$e_j(x) = f_j(x_1) \dots f_j(x_d),$$

where $(f_j)_{j=1}^{\infty}$ is a basis of $L^2([0,1]^d)$. A natural choice is:

$$f_j(x) = \sqrt{2} \sin(\pi j x),$$

which forms an orthonormal basis of $L^2([0,1])$.

Class Exercise: Verify that for this choice of basis:

$$\forall l \neq j, \quad \int_0^1 f_j(x) f_l(x) dx = 0,$$

$$\forall j, \quad \int_0^1 f_j^2(x) dx = 1.$$

Now, consider the transformation relationship $f_y = A f_x$. If A is diagonal in the basis e_j, meaning $A e_j = \lambda_j e_j$, then we can express f_x as:

$$f_x = \sum_j \frac{\langle f_y, e_j \rangle}{\lambda_j} e_j.$$

This decomposition helps analyze how the transformation A affects the function f_x.

To study the **Mean Integrated Squared Error (MISE)**, we analyze the variance:

$$\sigma^2 = \mathbb{E}\left[||\hat{f}_x - \mathbb{E}[f_x]||_2^2 \right] = \sum_{j=1}^{T} \mathbb{E}\left[(\hat{c_j} - c_j)^2 \right].$$

From our lecture, the optimal choice of T depends on the smoothness β of f_x and the dimension d, and is given by:

$$T = \left\lfloor n^{\frac{1}{2\beta + d}} \right\rfloor.$$

However, this choice is not feasible in practice since β is unknown.

Remark 2.4. A practical approach is to use a data-driven estimator. We start with a large T (e.g., $T = n$), which results in low bias but high variance. Then, we shrink the coefficients $\hat{c_j}$ to reduce variance, using methods like:

- **LASSO** (L1 regularization)
- **BIC** (Bayesian Information Criterion)

These methods select an appropriate set of basis functions automatically.

In the end of the class we were also given some comments on MLE nonparametric estimation.

2.4 notes on lecture 9

2.4.1 Estimating a regression function

$y = f(x) + \epsilon$ where $\mathbb{E}[|\epsilon|] < \infty$ and $\mathbb{E}[\epsilon/x] = 0$. Then if the conditional density function of y on x, denoted $f_{y|x}$, is given by $f_{y|X=x} = \frac{f_{y,x}(y,x)}{f_x(x)}$ provided that $f_x(x) \neq 0$. An idea that was suggested during class is to use $\hat{f}_{y,x} = \frac{1}{nh^2}\sum_i^n k\left(\frac{y_i - y}{h}\right) k\left(\frac{x_i - y}{h}\right)$, this leads to the the Nadaray-Watson estimator:

$$f(x) = \frac{\sum_{i=1}^n y_i k\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}$$

.

A model is well-specified if the true $|_{\mathbb{P}_{\epsilon,x}} f^x$ is such that $y = f^*(x^*) + \epsilon^*$, and $\mathbb{E}\left[\epsilon^* | x^*\right] = 0$, $\mathbb{E}\left[|\epsilon^*|\right]$. Then $\int d\mathbb{P}_{\epsilon,x} d\epsilon = f^*(x)$ so we just need to distinguish x, x^*. Such that, for example, $0 = f^*(x) - f(x) + \epsilon^* - \epsilon \iff f^* = f(x)$.

We have that $\mathbb{E}[y|X = x] = \int_{\mathbb{R}} \frac{f_{y,x}(y,x)}{f_x(x)} dy = \ldots$ And from that we can derive the following expression: $\int_{\mathbb{R}} y f_{y,x} dy = \frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\left[\int_{\mathbb{R}} yk\left(\frac{y_i - y}{h}\right) dy\right]$, then by subsitution by parts we define $u = \frac{y_i - y}{h} \iff uh = y_i - y$ and $du = \frac{-dy}{h}$. If we substitute we get the following: $\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\left[\int_{\mathbb{R}} y_i k(u) du - h\int_{\mathbb{R}} uk(u) du\right] = \frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\left[\int_{\mathbb{R}} y_i k(u) du\right]$ if k is symmetric (we gave the proof in seminar 2)

Remark 2.5. A classical exam question can be of the type: what if we want to estimate $\mathbb{E}(f'(x))$? etc–¿ naive plug in $(\frac{1}{n}\sum_{i=1}^n f'(x))$ can give us the marginal effects (now TD2 makes some sort of sense).

2.4.2 Series Estimator

Assume the data-generating process (DGP) is given by:

$$y = f(x) + \epsilon, \quad x \sim U(0,1).$$

Let e_j be a basis of $L^2(0,1)$. Then, taking expectations,

$$\mathbb{E}[y \cdot e_j(x)] = \mathbb{E}[f(x) \cdot e_j(x)] + \mathbb{E}[\epsilon \cdot e_j(x)].$$

By the Law of Iterated Expectations (LIE) and the assumption that $\mathbb{E}[\epsilon|x] = 0$ (since ϵ is independent of x), we obtain:

$$\mathbb{E}[y \cdot e_j(x)] = \mathbb{E}[f(x) \cdot e_j(x)].$$

A natural estimator for this expectation is:

$$\hat{C}_j = \frac{1}{n}\sum_{i=1}^n y_i e_j(x_i),$$

which allows us to approximate $f(x)$ using a truncated series expansion:

$$\hat{f}^T(x) = \sum_{j=1}^T \hat{C}_j e_j(x).$$

Alternative Case: When $x \sim F(x)$ If instead we assume that $x \sim F(x)$ with known density function $f_x(x)$, we can rewrite the expectation as:

$$\mathbb{E}\left[y \cdot \frac{e_j(x)}{f_x(x)}\right] = \mathbb{E}\left[f(x) \cdot \frac{e_j(x)}{f_x(x)}\right].$$

Using the definition of expectation and canceling $f_x(x)$:

$$\int_{\mathbb{R}} f(x)\frac{e_j(x)}{f_x(x)} f_x(x)dx = \int_{\mathbb{R}} f(x)e_j(x)dx.$$

Thus, we can estimate the coefficients in this case using:

$$\hat{C}_j = \frac{1}{n}\sum_{i=1}^{n} y_i \frac{e_j(x_i)}{f_x(x_i)}.$$

Trimmed Estimation for Large T An alternative estimation approach for large T introduces a trimming value au to avoid numerical instability when $\hat{\beta}_x(x_i)$ is close to zero. The modified estimator is:

$$\hat{C}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i e_j(x_i)}{\max(\hat{\beta}_x(x_i), \tau)}.$$

This ensures numerical stability by preventing extreme weight values that could lead to high variance in estimation. The choice of τ depends on the application and is typically small.

2.4.3 Nonparametric least-squares

Let $(\varphi_h)_{h=1}^p$ be p functions and $f_\beta = \sum_{h=1}^p \beta_k \cdot \varphi_k$, then we define $f = f_\beta + r$ where r is the remainder. Let $\hat{\beta} \in \arg\min\left(\frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{p=1}^P \beta_h \cdot \varphi_k(x_i)\right)^2\right)$, and we take $\hat{f} = \hat{f}_\beta$. If $\text{rank}(X) = P$ then $\hat{\beta} = \left((\frac{1}{n}X'X)^{-1}\left(\frac{1}{n}X\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}\right)\right)$. If instead we have $T > n$ such that $\text{rank}(x) < p$ then we need to use penalized estimators such that, for example: $\hat{\beta} \in \arg\min\left(\frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{p=1}^P \beta_h \cdot \varphi_k(x_i)\right)^2 + \lambda\rho(\beta)\right)$ where $\rho(\beta)$ is the penalization coefficient. In LASSO, this coefficient is equal to $\sum_{h=1}^T |\beta_h|$ and another example can be the ridge criteria that uses the function $\sum_{k=1}^T \beta_k^2$