

# Topics in Statistical Theory — Summary

Lucas Riedstra

November 27, 2020

## Contents

<b>1</b>	<b>Basic concepts</b>	<b>2</b>
1.1	Parametric vs nonparametric models . . . . .	2
1.2	Estimating an arbitrary distribution function . . . . .	2
1.3	Order statistics and quantiles . . . . .	5
1.4	Concentration inequalities . . . . .	6
<b>2</b>	<b>Kernel density estimation</b>	<b>12</b>
2.1	The univariate kernel density estimator . . . . .	12
2.2	Bounds on variance and bias . . . . .	13
2.3	Bounds on the integrated variance and bias . . . . .	15
2.4	Bandwidth selection . . . . .	16
2.4.1	Least squares cross validation . . . . .	16
2.4.2	Lepski . . . . .	17
2.5	Choice of kernel . . . . .	17
2.6	Multivariate density estimation . . . . .	17
<b>3</b>	<b>Nonparametric regression</b>	<b>18</b>
3.1	Fixed and random design . . . . .	18
3.2	Local polynomial estimators . . . . .	18
3.3	Splines . . . . .	22
3.3.1	Cubic splines . . . . .	22
3.3.2	Natural cubic smoothing splines . . . . .	23
3.3.3	Choice of smoothing parameter . . . . .	23
<b>4</b>	<b>Minimax lower bounds</b>	<b>25</b>
4.1	Reduction to testing . . . . .	25
4.2	Divergences . . . . .	26
4.3	Le Cam's two point lemma . . . . .	28
4.4	Assouad's lemma . . . . .	28
4.5	The data processing inequality . . . . .	30

# 1 Basic concepts

## 1.1 Parametric vs nonparametric models

**Definition 1.1.** A *statistical model* is a family of possible data-generating mechanisms. If the parameter space  $\Theta$  is finite-dimensional, we speak of a *parametric model*.

A model is called *well-specified* if there is a  $\vartheta_0 \in \Theta$  for which the data was generated from the distribution with parameter  $\vartheta_0$ , and otherwise it is called *misspecified*.

**Recap 1.2.** Let  $(Y_n)$  be a sequence of random vectors and  $Y$  a random vector.

1. We say that  $(Y_n)$  *converges almost surely* to  $Y$ , notation  $Y_n \xrightarrow{\text{a.s.}} Y$ , if  $\mathbb{P}(Y_n \rightarrow Y) = 1$ .
2. We say that  $(Y_n)$  *converges in probability* to  $Y$ , notation  $Y_n \xrightarrow{\text{P}} Y$ , if for every  $\varepsilon > 0$  we have  $\mathbb{P}(\|Y_n - Y\| > \varepsilon) \rightarrow 0$ .
3. We say that  $(Y_n)$  *converges in distribution* to  $Y$ , notation  $Y_n \xrightarrow{\text{d}} Y$ , if  $\mathbb{P}(Y_n \leq y) \rightarrow \mathbb{P}(Y \leq y)$  for all  $y$  where the distribution function of  $Y$  is continuous.

This is equivalent to the condition that  $\mathbb{E}[f(Y_n)] \rightarrow \mathbb{E}[f(Y)]$  for all bounded Lipschitz functions  $f$ .

It is known that  $Y_n \xrightarrow{\text{a.s.}} Y \implies Y_n \xrightarrow{\text{P}} Y \implies Y_n \xrightarrow{\text{d}} Y$ .

If  $(Y_n)$  is a sequence of random vectors and  $(a_n)$  is a positive sequence, then we write  $Y_n = O_p(a_n)$  if, for all  $\varepsilon > 0$ , there exists  $C > 0$  such that for sufficiently large  $n$  we have

$$\mathbb{P}\left(\frac{\|Y_n\|}{a_n} > C\right) < \varepsilon.$$

We write  $Y_n = o_p(a_n)$  if  $Y_n/a_n \xrightarrow{\text{P}} 0$ .

In a well-specified parametric model, the maximum likelihood estimator (MLE)  $\hat{\vartheta}_n$  typically satisfies  $\hat{\vartheta}_n - \vartheta_0 \in O_p(n^{-1/2})$ . On the other hand, if the model is misspecified, any inference can give very misleading results. To circumvent this problem, we consider *nonparametric models*, which make much weaker assumptions. Such infinite-dimensional models are much less vulnerable to model misspecification, however we will typically pay a price in terms of a slower convergence rate than in well-specified parametric models.

**Example 1.3.** Examples of nonparametric models include:

1. Assume  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$  for some unknown distribution function  $F$ .
2. Assume  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f$  for some unknown density  $f$  belonging to a smoothness class.
3. Assume  $Y_i = m(x_i) + \varepsilon_i$  ( $i = 1, \dots, n$ ), where the  $x_i$  are known,  $m$  is unknown and belongs to some smoothness class, and the  $\varepsilon_i$  are i.i.d. with  $\mathbb{E}(\varepsilon_i) = 0$  and  $\text{Var}(\varepsilon_i) = \sigma^2$ .

## 1.2 Estimating an arbitrary distribution function

**Definition 1.4.** Let  $\mathcal{F}$  denote the class of all distribution functions on  $\mathbb{R}$  and suppose  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F \in \mathcal{F}$ . The *empirical distribution function*  $\hat{F}_n$  of  $X_1, \dots, X_n$  is defined as

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}.$$

**Recap 1.5.** The *strong law of large numbers* tells us that if  $(Y_n)$  are i.i.d. with finite mean  $\mu$ , then  $\bar{Y} := \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{\text{a.s.}} \mu$ .

Note that the strong law of large numbers immediately implies that  $\hat{F}_n(x)$  converges almost surely to  $F(x)$  as  $n \rightarrow \infty$ . However, the following stronger result states that this convergence holds uniformly in  $x$ :

**Theorem 1.6** (Glivenko-Cantelli). *Let  $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} F$ . Then we have*

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{\text{a.s.}} 0.$$

*Proof.* See lecture notes. The main idea of the proof is to “control”  $\hat{F}_n$  in a finite number of points  $x_1, \dots, x_k$ , and then deduce what happens between those points using the fact that distributions are increasing and right-continuous. On [Wikipedia](#), a simplified proof can be found assuming that  $F$  is continuous, which still encapsulates the main idea.  $\square$

**Theorem 1.7** (Dvoretzky-Kiefer-Wolfowitz). *Under the conditions of theorem 1.6, for every  $\varepsilon > 0$  it holds that*

$$\mathbb{P}_F \left( \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| > \varepsilon \right) \leq 2e^{-2n\varepsilon^2},$$

and this is a tight bound.

We will not prove this theorem, however, we will explore a few consequences. One of these consequences is the following:

**Corollary 1.8** (Uniform Glivenko-Cantelli theorem). *Under the conditions of theorem 1.6, for every  $\varepsilon > 0$ , it holds that*

$$\sup_{F \in \mathcal{F}} \mathbb{P}_F \left( \sup_{m \geq n} \sup_{x \in \mathbb{R}} \left| \hat{F}_m(x) - F(x) \right| > \varepsilon \right) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

*Proof.* By a union bound, the DKW inequality, and convergence of the geometric series we have

$$\begin{aligned} \sup_{F \in \mathcal{F}} \mathbb{P}_F \left( \sup_{m \geq n} \sup_{x \in \mathbb{R}} \left| \hat{F}_m(x) - F(x) \right| > \varepsilon \right) &\leq \sup_{F \in \mathcal{F}} \sum_{m=n}^{\infty} \mathbb{P}_F \left( \sup_{x \in \mathbb{R}} \left| \hat{F}_m(x) - F(x) \right| > \varepsilon \right) \\ &\leq 2 \sum_{m=n}^{\infty} e^{-2m\varepsilon^2}, \end{aligned}$$

which converges to 0 as it is the tail of a converging sum.  $\square$

For another consequence, we consider the problem of finding a confidence band for  $F$ . Given  $\alpha \in (0, 1)$ , set  $\varepsilon_n := \sqrt{-\frac{1}{2n} \log(\alpha/2)}$ . Then the DKW inequality tells us that

$$\mathbb{P}_F \left( \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| > \varepsilon_n \right) \leq \alpha,$$

or equivalently, that

$$\mathbb{P}_F \left( \hat{F}_n(x) - \varepsilon_n \leq F(x) \leq \hat{F}_n(x) + \varepsilon_n \text{ for all } x \in \mathbb{R} \right) \geq 1 - \alpha.$$

We can say even more.

**Recap 1.9.** For any distribution function  $F$ , its *quantile function* is defined as

$$F^{-1}: (0, 1] \rightarrow \mathbb{R} \cup \{\infty\}: p \mapsto \inf \{x \in \mathbb{R} \mid F(x) \geq p\}.$$

When necessary, we also define  $F^{-1}(0) := \sup \{x \in \mathbb{R} \mid F(x) = 0\}$ .

If  $U \sim U(0, 1)$  and  $X \sim F$ , then for any  $x \in \mathbb{R}$  we have

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x) = \mathbb{P}(X \leq x).$$

This can be written simply as  $F^{-1}(U) \stackrel{d}{=} X$ .

Let  $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0, 1)$  with empirical distribution function  $\hat{G}_n$ , and let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ . Then, we have

$$\hat{G}_n(F(x)) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{U_i \leq F(x)\}} \stackrel{d}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} = \hat{F}_n(x),$$

where  $\stackrel{d}{=}$  means equality in distribution. It follows that

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \stackrel{d}{=} \sup_{x \in \mathbb{R}} \left| \hat{G}_n(F(x)) - F(x) \right| \leq \sup_{t \in [0, 1]} \left| \hat{G}_n(t) - t \right|,$$

with equality if  $F$  is continuous. We conclude that if  $F$  is continuous, the distribution of  $\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right|$  does not depend on  $F$ .

Other generalisations of theorem 1.6 include Uniform Laws of Large Numbers. Let  $X, X_1, \dots, X_n$  be i.i.d. on a measurable space  $(\mathcal{X}, \mathcal{A})$ , and  $\mathcal{G}$  a class of measurable functions on  $\mathcal{X}$ . We say that  $\mathcal{G}$  satisfies a ULLN if

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(X_i) - \mathbb{E}[g(X)] \right| \xrightarrow{\text{a.s.}} 0.$$

In theorem 1.6, we showed that  $\mathcal{G} = \{\mathbb{1}_{\{\cdot \leq x\}} \mid x \in \mathbb{R}\}$  satisfies a ULLN.

**Recap 1.10.** We recall the central limit theorem: if  $X_1, \dots, X_n$  are i.i.d. random variables with mean  $\mu$  and variance  $\sigma^2 < \infty$ , then  $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ .

Dividing by  $\sigma$  yields

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1),$$

and multiplying both sides by  $n$  and writing  $V_i = \sum_{j=1}^i X_j$  we obtain

$$\frac{V_i - \mathbb{E}V_i}{\sqrt{\text{Var}(V_i)}} \xrightarrow{d} N(0, 1).$$

Another extension starts with the observation that  $\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, \sigma^2)$ , where

$$\sigma^2 = \text{Var}(\mathbb{1}_{\{X \leq x\}}) = \mathbb{E}[\mathbb{1}_{\{X \leq x\}}^2] - \mathbb{E}[\mathbb{1}_{\{X \leq x\}}]^2 = F(x) - F(x)^2 = F(x)(1 - F(x)).$$

This can be strengthened by considering  $(\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, \sigma^2) \mid x \in \mathbb{R})$  as a stochastic process.

### 1.3 Order statistics and quantiles

**Definition 1.11.** Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F \in \mathcal{F}$ . The *order statistics* are the ordered samples  $X_{(1)} \leq \dots \leq X_{(n)}$  (where the original order is preserved in case of a tie).

The order statistics of the uniform distribution can be computed explicitly:

**Proposition 1.12.** Let  $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0, 1)$ , let  $Y_1, \dots, Y_{n+1} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ , and write  $S_j := \sum_{i=1}^j Y_i$  ( $j = 1, \dots, n+1$ ). Then

$$U_{(j)} \stackrel{d}{=} \frac{S_j}{S_{n+1}} \sim \text{Beta}(j, n-j+1) \quad \text{for } j = 1, \dots, n.$$

*Proof.* See example sheet 1, question 1. □

**Definition 1.13.** Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ . Then the *sample quantile function* is defined as

$$\hat{F}_n^{-1}(p) = \inf \left\{ x \in \mathbb{R} \mid \hat{F}_n(x) \geq p \right\}.$$

**Proposition 1.14.** It holds that  $\hat{F}_n^{-1}(p) = X_{(\lceil np \rceil)}$ .

*Proof.* By definition,  $\hat{F}_n^{-1}(p)$  is the smallest value of  $x$  for which  $\hat{F}_n(x)$  is larger than  $p$ . Note that

$$\hat{F}_n(x) \geq p \iff \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \geq p \iff \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \geq np \iff \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \geq \lceil np \rceil.$$

The smallest value of  $x$  for which this occurs is the smallest value of  $x$  such that exactly  $\lceil np \rceil$  of the variables  $X_1, \dots, X_n$  satisfy  $X_i \leq x$ . We conclude that  $\hat{F}_n^{-1}(p) = X_{(\lceil np \rceil)}$ . □

For  $p = \frac{1}{2}$  for example, this proposition tells us that  $\hat{F}_n^{-1}(p) = X_{(\lceil n/2 \rceil)}$ , the median of the data. We now explore the distribution of  $X_{(\lceil np \rceil)}$ .

**Recap 1.15.** We recall two theorems. The first is *Slutsky's theorem*:

**Theorem 1.16.** Let  $(Y_n)$  and  $(Z_n)$  be sequences of random vectors with  $Y_n \xrightarrow{d} Y$  and  $Z_n \xrightarrow{p} c$  for some constant  $c$ . If  $g$  is a continuous real-valued function, then  $g(Y_n, Z_n) \xrightarrow{d} g(Y, c)$ .

The second is the *delta method*:

**Theorem 1.17.** Let  $(Y_n)$  be a sequence of random vectors such that  $\sqrt{n}(Y_n - \mu) \xrightarrow{d} Z$ . If  $g: \mathbb{R}^d \rightarrow \mathbb{R}$  is differentiable at  $\mu$ , then

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{d} g'(\mu)Z.$$

**Lemma 1.18.** If  $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0, 1)$  and  $p \in (0, 1)$ , then  $\sqrt{n}(U_{\lceil np \rceil} - p) \xrightarrow{d} N(0, p(1-p))$ .

*Proof.* Let  $Y_1, \dots, Y_{n+1} \stackrel{\text{iid}}{\sim} \text{Exp}(1)$ ,  $V_n := \sum_{i=1}^{\lceil np \rceil} Y_i$  and  $W_n := \sum_{i=\lceil np \rceil+1}^{n+1} Y_i$ . Then  $V_n$  and  $W_n$  are independent, and we have seen that  $U_{\lceil np \rceil} \sim \frac{V_n}{V_n + W_n}$ .

Noting that  $\mathbb{E}V_n = \text{Var}(V_n) = \lceil np \rceil$  we find

$$\begin{aligned} \sqrt{n} \left( \frac{V_n}{n} - p \right) &= \frac{\sqrt{\lceil np \rceil}}{\sqrt{n}} \left( \frac{V_n - \lceil np \rceil}{\sqrt{\lceil np \rceil}} \right) + \frac{\lceil np \rceil - np}{\sqrt{n}} \\ &= \frac{\sqrt{\lceil np \rceil}}{\sqrt{n}} \left( \frac{V_n - \mathbb{E}V_n}{\sqrt{\text{Var}(V_n)}} \right) + \frac{\lceil np \rceil - np}{\sqrt{n}}. \end{aligned}$$

Now, by the central limit theorem, the term between brackets converges to a standard  $N(0, 1)$  distribution. The term  $\sqrt{\lceil np \rceil} \sqrt{n}$  converges to  $\sqrt{p}$  and the term  $(\lceil np \rceil - np)/\sqrt{n}$  converges to 0, so by Slutsky's lemma, we find

$$\sqrt{n} \left( \frac{V_n}{n} - p \right) \xrightarrow{d} \sqrt{p} N(0, 1) = N(0, p).$$

An analogous calculation shows that  $\sqrt{n} \left( \frac{W_n}{n} - (1 - p) \right) \rightarrow N(0, 1 - p)$ .

Now we define  $g: (0, \infty)^2 \rightarrow (0, \infty)$  by  $g(x, y) := x/(x + y)$ , which is differentiable at  $(p, 1 - p)$ . Note that the distribution of  $(V_n, W_n)$  is an  $N(0, \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix})$  distribution. By the delta method we find

$$\begin{aligned} \sqrt{n}(U_{\lceil np \rceil} - p) &\stackrel{d}{=} \sqrt{n} \left( g \left( \frac{V_n}{n}, \frac{W_n}{n} \right) - g(p, q) \right) \\ &\stackrel{d}{\rightarrow} g'(p, 1 - p) N \left( 0, \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} \right) \\ &= N \left( 0, g'(p, 1 - p) \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix} g'(p, 1 - p)^\top \right) \\ &= N(0, p(1 - p)). \end{aligned}$$

□

We now relate what we know about the uniform distribution to the quantile function:

**Theorem 1.19.** *Let  $p \in (0, 1)$  and let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ . Suppose that  $F$  is differentiable at  $\xi_p := F^{-1}(p)$  with derivative  $f(\xi_p)$ . Then*

$$\sqrt{n}(X_{(\lceil np \rceil)} - \xi_p) \xrightarrow{d} N \left( 0, \frac{p(1 - p)}{f(\xi_p)^2} \right).$$

*Proof.* Let  $U_1, \dots, U_n \stackrel{\text{iid}}{\sim} U(0, 1)$ , then we know that  $F^{-1}(U_i) \stackrel{d}{=} X_i$  and thus  $F^{-1}(U_{(\lceil np \rceil)}) \stackrel{d}{=} X_{(\lceil np \rceil)}$ . Applying the delta method with  $g = F^{-1}$ , together with the previous theorem yields

$$\sqrt{n}(X_{(\lceil np \rceil)} - \xi_p) = \sqrt{n}(F^{-1}(U_{(\lceil np \rceil)}) - F^{-1}(p)) \xrightarrow{d} (F^{-1})'(p) \cdot N(0, p(1 - p)).$$

Noting that  $(F^{-1})'(p) = \frac{1}{f(\xi_p)}$  yields the result. □

## 1.4 Concentration inequalities

We turn our attention to concentration inequalities, with a focus on finite-sample results (instead of results that only hold for  $n \rightarrow \infty$ ).

**Definition 1.20.** A random variable  $X$  with mean 0 is called *sub-Gaussian* with parameter  $\sigma^2$  if

$$M_X(t) = \mathbb{E}(e^{tX}) \leq e^{t^2 \sigma^2 / 2}$$

for every  $t \in \mathbb{R}$ .

Note that equality holds when  $X \sim N(0, \sigma^2)$ , since the MGF of an  $N(\mu, \sigma^2)$  distribution is given by  $t \mapsto \exp(\mu t + \sigma^2 t^2 / 2)$ .

**Recap 1.21.** Recall the *tail bound formula* for the expectation: if  $X$  is a nonnegative random variable, then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x) dx.$$

Furthermore, recall that the *gamma function* is defined for  $z \in (0, \infty)$  by

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

and satisfies  $\Gamma(n) = (n-1)!$  for all  $n \in \mathbb{N}$ .

Finally, recall the following inequality: for all  $a, b \in \mathbb{R}$  and  $p \geq 1$

$$(a + b)^p \leq 2^{p-1}(a^p + b^p).$$

This follows from the convexity of the function  $x \mapsto x^p$ .

**Proposition 1.22.** *We consider some characterisations of sub-Gaussianity:*

(a) *Let  $X$  be sub-Gaussian with parameter  $\sigma^2$ . Then*

$$\max\{\mathbb{P}(X \geq x), \mathbb{P}(X \leq -x)\} \leq e^{-x^2/(2\sigma^2)} \quad \text{for every } x \geq 0. \quad (1)$$

(b) *Let  $X$  be a random variable which satisfies  $\mathbb{E}(X) = 0$  and eq. (1). Then for every  $q \in \mathbb{N}$  it holds that*

$$\mathbb{E}(X^{2q}) \leq 2 \cdot q!(2\sigma^2)^q \leq q!(2\sigma)^{2q}.$$

(c) *If  $X$  is a random variable with  $\mathbb{E}(X) = 0$  and  $\mathbb{E}(X^{2q}) \leq q!C^{2q}$  for all  $q \in \mathbb{N}$ , then  $X$  is sub-Gaussian with parameter  $4C^2$ .*

*Proof.* (a) We first consider  $\mathbb{P}(X \geq x)$ . By Markov's inequality, we have for all  $t \in \mathbb{R}$  that

$$\mathbb{P}(X \geq x) = \mathbb{P}(e^{tX} \geq e^{tx}) \leq e^{-tx} \mathbb{E}(e^{tX}) \leq e^{-tx+t^2\sigma^2/2}.$$

Since the LHS is independent of  $t$ , we can take the infimum over  $t$  on the RHS and obtain

$$\mathbb{P}(X \geq x) \leq \inf_{t \in \mathbb{R}} e^{-tx+t^2\sigma^2/2} = e^{-x^2/(2\sigma^2)},$$

since the infimum of  $t^2\sigma^2/2 - tx$  is attained at  $t = x/\sigma^2$  (this method is called *Chernoff bounding*).

For  $\mathbb{P}(X \leq -x) = \mathbb{P}(-X \geq x)$  we can use the fact that  $-X$  is also sub-Gaussian with parameter  $\sigma^2$ .

(b) By the previous part, we have  $\mathbb{P}(|X| \geq x) \leq 2e^{-x^2/(2\sigma^2)}$ . Some calculations give

$$\begin{aligned} \mathbb{E}(X^{2q}) &= \int_0^\infty \mathbb{P}(X^{2q} \geq x) dx = \int_0^\infty \mathbb{P}(|X| \geq x^{1/(2q)}) dx \\ &= 2q \int_0^\infty x^{2q-1} \mathbb{P}(|X| \geq x) dx \\ &\leq 4q \int_0^\infty x^{2q-1} e^{-x^2/(2\sigma^2)} dx. \end{aligned}$$

Now set  $t = x^2/2\sigma^2$ , so that  $x = \sigma(2t)^{1/2}$  and thus  $dx = \sigma(2t)^{-1/2} dt$ . Plugging that in we get

$$\begin{aligned} \mathbb{E}(X^{2q}) &\leq 4q \int_0^\infty (\sigma(2t)^{1/2})^{2q-1} e^{-t} \sigma(2t)^{-1/2} dt = 2^{q+1} q \sigma^{2q} \int_0^\infty t^{q-1} e^{-t} dt \\ &= 2^{q+1} q \sigma^{2q} \Gamma(q) = 2 \cdot q!(2\sigma)^q. \end{aligned}$$

- (c) Note that  $x \mapsto e^{-tx}$  is convex for every  $t \in \mathbb{R}$ , so  $\mathbb{E}(e^{-tX}) \geq e^{-t\mathbb{E}(X)} = e^0 = 1$  by Jensen's inequality. Let  $X'$  denote an independent copy of  $X$ : then  $X - X'$  has a symmetric distribution, so all its odd moments vanish. Therefore we find

$$\begin{aligned} \mathbb{E}[e^{tX}] &\leq \mathbb{E}[e^{-tX'}]\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(X-X')}] = \mathbb{E}\sum_{q=0}^{\infty} \left[ \frac{t^{2q}(X-X')^{2q}}{(2q)!} \right] \\ &= \sum_{q=0}^{\infty} \frac{t^{2q}\mathbb{E}[(X-X')^{2q}]}{(2q)!} \leq \sum_{q=0}^{\infty} \frac{2^{2q-1}t^{2q}(\mathbb{E}[X^{2q}] + \mathbb{E}[(X')^{2q}])}{(2q)!} \\ &\leq \sum_{q=0}^{\infty} \frac{2^{2q-1}t^{2q}2q!C^{2q}}{(2q)!} = \sum_{q=0}^{\infty} \frac{(2tC)^{2q}q!}{(2q)!} = \sum_{q=0}^{\infty} \frac{(2tC)^{2q}}{\prod_{j=1}^q (q+j)} \\ &\leq \sum_{q=0}^{\infty} \frac{(2tC)^{2q}}{\prod_{j=1}^q (2j)} = \sum_{q=1}^{\infty} \frac{(2t^2C^2)^q}{q!} = e^{2t^2C^2}. \end{aligned}$$

This shows that  $X$  is sub-Gaussian with parameter  $4C^2$ . □

Note that the proposition is not an “if and only if”-type theorem: suppose we start with a sub-Gaussian variable  $X$  with parameter  $\sigma^2$ . Then by (b), we have  $\mathbb{E}[X^{2q}] \leq q!(2\sigma)^{2q}$ , and (c) then implies that  $X$  is sub-Gaussian with parameter  $16\sigma^2$ .

**Theorem 1.23** (Hoeffding's inequality). *Let  $X_1, \dots, X_n$  be independent sub-Gaussian random variables, with  $X_i$  having parameter  $\sigma_i^2$ . Then  $\bar{X}$  is sub-Gaussian with parameter  $\bar{\sigma}^2$ . In particular, we have*

$$\max\{\mathbb{P}(\bar{X} \geq x), \mathbb{P}(\bar{X} \leq -x)\} \leq e^{-nx^2/(2\bar{\sigma}^2)}.$$

*Proof.* For  $t \in \mathbb{R}$ , we have

$$\mathbb{E}[e^{t\bar{X}}] = \mathbb{E}[e^{(t/n)\sum_i X_i}] = \prod_{i=1}^n \mathbb{E}[e^{(t/n)X_i}] \leq \prod_{i=1}^n e^{t^2\sigma_i^2/(2n^2)} = e^{t^2\bar{\sigma}^2/(2n)},$$

which shows  $\bar{X}$  is sub-Gaussian with parameter  $\bar{\sigma}^2/n$ . Applying part (a) of the previous proposition shows the second result. □

*Remark.* A direct consequence of Hoeffding's inequality is that

$$\mathbb{P}(|\bar{X}| \geq x) \leq 2e^{-nx^2/(2\bar{\sigma}^2)}.$$

The inequality is often stated in this weaker way.

**Lemma 1.24** (Hoeffding's lemma). *Let  $X$  be a random variable with  $\mathbb{E}X = 0$  that satisfies  $a \leq X \leq b$ . Then  $X$  is sub-Gaussian with parameter  $(b-a)^2/4$ .*

*Proof.* See Example Sheet 1, question 2. □

**Corollary 1.25.** *Let  $X_1, \dots, X_n$  be independent random variables where  $\mathbb{E}[X_i] = \mu_i$  and  $a_i \leq X_i \leq b_i$ . Then we have*

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq x) \leq \exp\left(-\frac{2n^2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

*Proof.* By Hoeffding's lemma,  $X_i - \mu_i$  is sub-Gaussian with parameter  $(b_i - a_i)^2/4$  for each  $i$ . The result now follows from theorem 1.23. □



Note that when  $X$  takes values in  $[a, b]$ , its variance is at most  $(b - a)^2$ . However, when  $\text{Var}(X_i) \ll (b_i - a_i)^2$ , Hoeffding's inequality can be loose (for example, when  $X_i \sim \text{Bern}(p_i)$  with  $p_i$  small). In such circumstances, Bennett's or Bernstein's inequality may give better results.

**Theorem 1.26** (Bennett's inequality). *Let  $X_1, \dots, X_n$  be independent random variables with  $\mathbb{E}[X_i] = 0$ ,  $\sigma_i^2 := \text{Var}(X_i) < \infty$ , and  $X_i \leq b$  for some  $b > 0$ . Define  $S := \sum_{i=1}^n X_i$ ,  $\nu := \overline{\text{Var}(X_i)}$  and  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  by  $\varphi(u) := e^u - 1 - u = \sum_{k=2}^{\infty} \frac{u^k}{k!}$ , then for every  $t > 0$  we have*

$$\log \mathbb{E}[e^{tS}] \leq \frac{n\nu}{b^2} \varphi(bt).$$

Defining  $h: (0, \infty) \rightarrow [0, \infty)$  by  $h(u) := (1 + u) \log(1 + u) - u$ , we have for every  $x > 0$  that

$$\mathbb{P}(\bar{X} \geq x) \leq \exp \left( -\frac{n\nu}{b^2} h\left(\frac{bx}{\nu}\right) \right).$$

*Proof.* Define  $g: \mathbb{R} \rightarrow \mathbb{R}$  by

$$g(u) := \sum_{k=0}^{\infty} \frac{u^k}{(k+2)!} = \begin{cases} \frac{\varphi(u)}{u^2} & \text{if } u \neq 0, \\ \frac{1}{2} & \text{if } u = 0. \end{cases}$$

Then one can check that  $g$  is increasing on  $\mathbb{R}$ , so

$$e^{tX_i} - 1 - tX_i = t^2 X_i^2 g(tX_i) \leq t^2 X_i^2 g(tb) = X_i^2 \frac{\varphi(bt)}{b^2},$$

and therefore

$$e^{tX_i} \leq 1 + tX_i + X_i^2 \frac{\varphi(bt)}{b^2} \implies \mathbb{E}[e^{tX_i}] \leq 1 + \mathbb{E}[X_i^2] \frac{\varphi(bt)}{b^2} = 1 + \text{Var}(X_i) \frac{\varphi(bt)}{b^2}.$$

Hence for  $t > 0$  we have

$$\begin{aligned} \log \mathbb{E}[e^{tS}] &= \sum_{i=1}^n \log \mathbb{E}[e^{tX_i}] \leq n \cdot \frac{1}{n} \sum_{i=1}^n \log \left( 1 + \text{Var}(X_i) \frac{\varphi(bt)}{b^2} \right) \\ &\stackrel{*}{\leq} n \log \left( 1 + \frac{\nu \varphi(bt)}{b^2} \right) \stackrel{**}{\leq} \frac{n\nu}{b^2} \varphi(bt). \end{aligned}$$

Here, (\*) follows from the fact that  $\log$  is a concave function while (\*\*) follows from the fact that  $\log(1 + u) \leq u$  for all  $u \geq 0$ . This concludes the proof for the first part of the theorem.

Now, we apply the method of Chernoff bounding and find

$$\mathbb{P}(\bar{X} \geq x) = \mathbb{P}(S \geq nx) \leq \inf_{t>0} e^{-ntx} \mathbb{E}[e^{tS}] \leq \inf_{t>0} e^{-ntx + n\nu \varphi(bt)/b^2} = \exp \left( -\frac{n\nu}{b^2} h\left(\frac{bx}{\nu}\right) \right),$$

since one can check that the infimum is attained at  $t = b^{-1} \log(1 + bx/\nu)$ .  $\square$

**Definition 1.27.** A random variable  $X$  with  $\mathbb{E}X = 0$  is called *sub-Gamma in the right tail* with variance factor  $\sigma^2 > 0$  and scale  $c > 0$  if

$$\mathbb{E}[e^{tX}] \leq \exp \left( \frac{\sigma^2 t^2}{2(1 - ct)} \right)$$

for all  $t \in [0, 1/c)$ .

Note that this definition looks like that of sub-Gaussianity, except that  $e^{tX}$  can explode as  $t$  approaches  $1/c$ . We give some characteristics of sub-Gamma distributions:

**Definition 1.28.** For any  $x \in \mathbb{R}$  we define  $x_+ := \max(x, 0)$ .

**Proposition 1.29.** (a) Let  $X$  be sub-Gamma in the right tail with variance factor  $\sigma^2$  and scale  $c$ . Then

$$\mathbb{P}(X \geq x) \leq \exp\left(-\frac{x^2}{2(\sigma^2 + cx)}\right)$$

for all  $x \geq 0$ .

(b) Let  $X$  be a random variable with  $\mathbb{E}X = 0$ ,  $\mathbb{E}[X^2] \leq \sigma^2$  and  $\mathbb{E}[(X_+)^q] \leq q!\sigma^2 c^{q-2}/2$  for all  $q \geq 3$ . Then  $X$  is sub-Gamma in the right tail with variance factor  $\sigma^2$  and scale parameter  $c$ .

*Proof.* (a) Again, we apply a Chernoff bound: we have

$$\begin{aligned} \mathbb{P}(X \geq x) &\leq \inf_{t \in [0, 1/c)} e^{-tx} \mathbb{E}[e^{tX}] \leq \inf_{t \in [0, 1/c)} \exp\left(-tx + \frac{\sigma^2 t^2}{2(1-ct)}\right) \\ &\leq \exp\left(-\frac{x^2}{2(\sigma^2 + cx)}\right), \end{aligned}$$

where we have set  $t = x/(\sigma^2 + cx) \in [0, 1/c)$  in the final step.

(b) Recall from the proof of Bennett's inequality that  $g$  is increasing and therefore for  $u \leq 0$  we have  $\varphi(u) = u^2 g(u) \leq u^2 g(0) = \frac{u^2}{2}$ . Therefore, for every  $u \in \mathbb{R}$  we have

$$\varphi(u) \leq \frac{u^2}{2} + \sum_{q=3}^{\infty} \frac{(u_+)^q}{q!}.$$

We deduce that for  $t \in [0, 1/c)$  we have (note  $\log(x) \leq x - 1$  for all  $x$ ):

$$\log \mathbb{E}[e^{tX}] \leq \mathbb{E}(e^{tX}) - 1 = \mathbb{E}[\varphi(tX)] \leq \mathbb{E}\left[\frac{t^2 X^2}{2} + \sum_{q=3}^{\infty} \frac{t^q X_+^q}{q!}\right].$$

By Fubini's theorem, since the infinite sum has only positive terms we may interchange sum and expectation to obtain

$$\mathbb{E}\left[\frac{t^2 X^2}{2} + \sum_{q=3}^{\infty} \frac{t^q \mathbb{E}[X_+^q]}{q!}\right] = \frac{t^2 \text{Var}[X]}{2} + \sum_{q=3}^{\infty} \frac{t^q \mathbb{E}[X_+^q]}{q!} \leq \frac{\sigma^2 t^2}{2} \sum_{q=2}^{\infty} t^{q-2} c^{q-2} = \frac{\sigma^2 t^2}{2(1-ct)}.$$

□

Following this proposition, we can prove Bernstein's inequality:

**Theorem 1.30** (Bernstein's inequality). Let  $X_1, \dots, X_n$  be independent random variables with  $\mathbb{E}[X] = 0$ ,  $\frac{1}{n} \sum_{i=1}^n \text{Var}(X_i) \leq \sigma^2$  and  $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \leq q!\sigma^2 c^{q-2}/2$  some  $\sigma, c > 0$  and for all  $q \geq 3$ . Then  $S := \sum_{i=1}^n X_i$  is sub-Gamma in the right tail with variance factor  $n\sigma^2$  and scale parameter  $c$ . In particular we have

$$\mathbb{P}(\bar{X} \geq x) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + cx)}\right),$$

for all  $x \geq 0$ .

*Proof.* We have by part (b) of the previous proposition

$$\log \mathbb{E}[e^{tS}] = \sum_{i=1}^n \log \mathbb{E}[e^{tX_i}] \leq n \frac{\sigma^2 t^2}{2(1-ct)},$$

and the second claim follows from part (a) of the previous proposition:

$$\mathbb{P}(\bar{X} \geq x) = \mathbb{P}(S \geq nx) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + cx)}\right).$$

□

## 2 Kernel density estimation

Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f$ , and suppose we wish to estimate the density function  $f$ . The oldest way to do this is with a histogram: we divide  $\mathbb{R}$  into equally sized intervals or *bins*, and let  $I_x$  denote the bin containing  $x \in \mathbb{R}$ .

**Definition 2.1.** The *histogram density estimator*  $\hat{f}_n^H$  with bin width  $b > 0$  is given by

$$\hat{f}_n^H(x) := \frac{1}{nb} \sum_{i=1}^n \mathbb{1}_{X_i \in I_x}.$$

There are a few major drawbacks to using histograms: it is difficult to choose  $b$  and the positioning of bin edges, the theoretical performance is suboptimal (mostly due to their discontinuity) and graphical display in the multivariate case is difficult.

### 2.1 The univariate kernel density estimator

**Definition 2.2.** A Borel measurable function  $K: \mathbb{R} \rightarrow \mathbb{R}$  is called a *kernel* if it satisfies  $\int_{\mathbb{R}} K(x) dx = 1$ . A *univariate kernel density estimator* of  $f$  with kernel  $K$  and *bandwidth*  $h > 0$  is defined as

$$\hat{f}_n(x) := \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

Defining  $K_h(x) := \frac{1}{h} K\left(\frac{x}{h}\right)$ , we can rewrite this as

$$\hat{f}_n(x) := \frac{1}{n} \sum_{i=1}^n K_h(x - X_i).$$

Usually  $K$  is chosen to be non-negative (which ensures that  $K$  itself and  $\hat{f}_n$  are themselves density functions), and  $K$  is often chosen to be symmetric about 0. Generally, the choice of kernel  $K$  is much less important than the choice of bandwidth  $h$ .

If we consider  $\hat{f}_n(x)$  as a point estimator of  $f(x)$ , we typically wish to minimise the *mean squared error*

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

Other possibilities include the mean absolute error which (unlike the MSE) is scale-invariant. However, the MSE has an appealing decomposition into variance and bias terms:

$$\text{MSE}(\hat{f}_n(x)) = \text{Var}(\hat{f}_n(x)) + \text{Bias}^2(\hat{f}_n(x)).$$

We can express the MSE in terms of convolutions:

**Definition 2.3.** Let  $g_1, g_2: \mathbb{R} \rightarrow \mathbb{R}$  be measurable. Then the *convolution* of  $g_1$  and  $g_2$ , denoted  $g_1 * g_2$ , is defined by

$$(g_1 * g_2)(x) := \int_{\mathbb{R}} g_1(x - z)g_2(z) dz.$$

We can compute

$$\begin{aligned} \text{Bias } \hat{f}_n(x) &= \mathbb{E}[\hat{f}_n(x)] - f(x) = \mathbb{E}[K_h(x - X_1)] - f(x) = \int_{\mathbb{R}} K_h(x - z)f(z) dz \\ &= (K_h * f)(x) - f(x). \end{aligned} \tag{2}$$

Analogously, letting  $\xi_i := K_h(x - X_i)$  (note that these are i.i.d. random variables), we have

$$\text{Var } \hat{f}_n(x) = \frac{1}{n} \text{Var}(\xi_1) = \frac{1}{n} (\mathbb{E}[\xi_1^2] - \mathbb{E}^2[\xi_1]) = \frac{1}{n} [(K_h^2 * f)(x) - (K_h * f)^2(x)]. \quad (3)$$

To assess performance of  $h$  and  $K$ , we want to assess the performance of  $\hat{f}_n$  as an estimation of  $f$  as a function. This gives the following definition:

**Definition 2.4.** We define the *mean integrated squared error* or MISE as

$$\text{MISE}(\hat{f}_n) := \mathbb{E} \left( \int_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 dx \right) \stackrel{*}{=} \int_{\mathbb{R}} \text{MSE}(\hat{f}_n(x)) dx,$$

where  $\star$  follows from Fubini's theorem since the integrand is nonnegative.

We now aim to find bounds on the bias and the variance of  $\hat{f}_n$  in order to choose  $h$  and  $K$  appropriately.

## 2.2 Bounds on variance and bias

**Definition 2.5.** For a kernel  $K$ , define  $R(K) := \int_{\mathbb{R}} K^2(u) du$ .

**Proposition 2.6.** Let  $\hat{f}_n$  be the kernel density estimator with kernel  $K$  and bandwidth  $h > 0$  constructed from  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f$ . Then for any  $x \in \mathbb{R}, h > 0, n \in \mathbb{N}$  we have

$$\text{Var } \hat{f}_n(x) \leq \frac{1}{nh} \|f\|_{\infty} R(K).$$

*Proof.* By eq. (3) we have

$$\begin{aligned} \text{Var } \hat{f}_n(x) &\leq \frac{1}{n} (K_h^2 * f)(x) = \frac{1}{nh^2} \int_{\mathbb{R}} K^2\left(\frac{x-z}{h}\right) f(z) dz = \frac{1}{nh} \int_{\mathbb{R}} K^2(u) f(x-uh) du \\ &\leq \frac{1}{nh} \|f\|_{\infty} \int_{\mathbb{R}} K^2(u) du = \frac{1}{nh} \|f\|_{\infty} R(K). \end{aligned} \quad (4)$$

□

Obtaining a bound on the bias is not at all straightforward: we will need to introduce conditions on both the density  $f$  and the kernel  $K$ .

**Definition 2.7.** Let  $I \subseteq \mathbb{R}$  be an interval, fix  $\beta, L > 0$ , and let  $m := \lceil \beta \rceil - 1$ . A function  $f: \mathbb{R} \rightarrow \mathbb{R}$  is said to belong to the *Hölder class*  $\mathcal{H}(\beta, L)$  if  $f$  is  $m$  times differentiable on  $I$  and

$$\left| f^{(m)}(x) - f^{(m)}(y) \right| \leq L|x-y|^{\beta-m} \quad \text{for all } x, y \in I.$$

If  $I$  is unspecified, we let  $I = \mathbb{R}$ .

The densities in  $\mathcal{H}(\beta, L)$  are denoted by

$$\mathcal{F}(\beta, L) := \left\{ f \in \mathcal{H}(\beta, L) \mid f \geq 0 \text{ and } \int_{\mathbb{R}} f dx = 1 \right\}.$$

**Definition 2.8.** Fix  $\ell \in \mathbb{N}$ . We say a kernel  $K$  is of *order*  $\ell$  if  $\int_{\mathbb{R}} x^j k(x) dx = 0$  for  $j = 1, \dots, \ell - 1$ .

*Remark.* Most kernels used in practice are of order 2. Note that a kernel of order  $\geq 3$  cannot be nonnegative, since we have  $\int_{\mathbb{R}} x^2 K(x) dx = 0$ . Therefore, the kernels are not themselves densities and the corresponding kernel density estimate is not guaranteed to be a density.

**Proposition 2.9.** Assume that  $f \in \mathcal{F}(\beta, L)$  and that  $K$  is a kernel of order  $\ell := \lceil \beta \rceil$ , and furthermore assume that

$$\mu_\beta(K) := \int_{\mathbb{R}} |u|^\beta |K(u)| \, du < \infty.$$

Then the kernel density estimate with bandwidth  $h$  and kernel  $K$  based on  $X_1, \dots, X_n \sim f$  satisfies

$$\left| \text{Bias } \hat{f}_n(x) \right| \leq \frac{L}{(\ell-1)!} \mu_\beta(K) h^\beta \quad \text{for all } x \in \mathbb{R}, h > 0, n \in \mathbb{N}.$$

*Proof.* By eq. (2), we have

$$\text{Bias } \hat{f}_n(x) = \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-z}{h}\right) f(z) \, dz - f(x) = \int_{\mathbb{R}} K(u) (f(x-uh) - f(x)) \, du.$$

By applying Taylor's theorem with the Lagrange remainder we obtain, with  $m = \lceil \beta \rceil - 1$ , that

$$f(x-uh) - f(x) = \sum_{j=1}^{m-1} \frac{(-uh)^j}{j!} f^{(j)}(x) + \frac{(-uh)^m}{m!} f^{(m)}(x - \tau uh) \quad \text{for some } \tau \in [0, 1].$$

Since  $\int_{\mathbb{R}} u^j K(u) \, du = 0$  for all  $j \leq m$ , plugging the sum into the integral will give 0. Therefore, we find

$$\text{Bias } \hat{f}_n(x) = \frac{(-h)^m}{m!} \int_{\mathbb{R}} u^m K(u) f^{(m)}(x - \tau uh) \, du = \frac{(-h)^m}{m!} \int_{\mathbb{R}} u^m K(u) \left[ f^{(m)}(x - \tau uh) - f^{(m)}(x) \right] \, du,$$

where the last inequality follows again from the fact that  $K$  is of order  $m+1$ .

Now we use that  $f \in \mathcal{F}(\beta, L)$ , and conclude

$$\left| \text{Bias } \hat{f}_n(x) \right| \leq \frac{Lh^m}{m!} \int_{\mathbb{R}} |u|^m |K(u)| |\tau uh|^{\beta-m} \, du \leq \frac{Lh^\beta}{m!} \int_{\mathbb{R}} |u|^\beta |K(u)| \, du = \frac{L}{(\ell-1)!} \mu_\beta(K) h^\beta,$$

which concludes the proof.  $\square$

Combining propositions 2.6 and 2.9, we find that

$$\text{MSE } \hat{f}_n(x) \leq \frac{1}{nh} R(K) \|f\|_\infty + \frac{L^2}{((\ell-1)!)^2} \mu_\beta^2(K) h^{2\beta}.$$

By minimising this w.r.t.  $h$ , we find that the optimal  $h$  is given by

$$h_n^* = \left( \frac{((\ell-1)!)^2 \|f\|_\infty R(K)}{2\beta L^2 \mu_\beta^2(K)} \right)^{1/(2\beta+1)} n^{-1/(2\beta+1)},$$

and the corresponding MSE is given by

$$\text{MSE } \hat{f}_n(x) \leq \left( \frac{\|f\|_\infty^{2\beta} R(K)^{2\beta} L^2 \mu_\beta^2(K) [(2\beta)^{2\beta+1} + 1]}{((\ell-1)!)^2 (2\beta)^{2\beta}} \right) n^{-2\beta/(2\beta+1)},$$

This  $O(n^{-2\beta/(2\beta+1)})$  bound on the rate is slower than the  $O(1/n)$  rate found in parametric problems, but such a rate is only obtained when the assumed model is correct.

We can strengthen this as follows:

**Theorem 2.10.** Assume that  $K$  is a kernel of order  $\ell := \lceil \beta \rceil$  and that  $\mu_\beta(K)$  and  $R(K)$  are both finite. Fix  $\alpha > 0$  and let  $h = \alpha n^{-1/(2\beta+1)}$ . Then there exists  $C > 0$ , independent of  $n$ , such that

$$\sup_{x \in \mathbb{R}} \sup_{f \in \mathcal{F}(\beta, L)} \text{MSE } \hat{f}_n(x) \leq C n^{-2\beta/(2\beta+1)}.$$

*Proof.* We will show that the class  $\mathcal{F}(\beta, L)$  is uniformly bounded in supremum norm. Let  $K^*$  be a bounded kernel of order  $\ell$  (see example sheet **TODO:** ), then by the previous proposition with  $h = 1$  we have by nonnegativity of  $f$  that

$$\begin{aligned} f(x) &\leq \left| f(x) - \int_{-\infty}^{\infty} K^*(x-z)f(z) dz \right| + \left| \int_{-\infty}^{\infty} K^*(x-z)f(z) dz \right| \\ &\leq \left| \text{Bias } \hat{f}_{n,K^*}(x) \right| + \|K^*\|_{\infty} \int_{-\infty}^{\infty} f(z) dz \\ &\leq \frac{L}{(\ell-1)!} \mu_{\beta}(K^*) + \|K^*\|_{\infty}, \end{aligned}$$

and this bound is independent of  $f$  and  $x$ .

Now we have

$$\text{MSE } \hat{f}_n(x) \leq \frac{R(K)\|f\|_{\infty}}{nh} + \frac{L^2}{((\ell-1)!)^2} \mu_{\beta}^2(K) h^{2\beta} \leq C n^{-2\beta/(2\beta+1)}.$$

□

### 2.3 Bounds on the integrated variance and bias

To bound the MISE, we will give bounds on the integrated variance and bias.

**Proposition 2.11.** Let  $\hat{f}_n$  denote the kernel density estimate with bandwidth  $h$  and kernel  $K$  based on  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$  (where  $P$  is a distribution on  $\mathbb{R}$ ). Then

$$\int_{-\infty}^{\infty} \text{Var } \hat{f}_n(x) dx = \frac{1}{nh} R(K).$$

*Proof.* We have by Fubini and eq. (4) that

$$\begin{aligned} \int_{-\infty}^{\infty} \text{Var } \hat{f}_n(x) dx &\leq \frac{1}{nh^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K^2\left(\frac{x-z}{h}\right) f(z) dz dx = \frac{1}{nh^2} \int_{-\infty}^{\infty} f(z) \int_{-\infty}^{\infty} K^2\left(\frac{x-z}{h}\right) dx dz \\ &= \frac{1}{nh} R(K) \int_{-\infty}^{\infty} f(z) dz = \frac{1}{nh} R(K). \end{aligned}$$

□

**Recap 2.12.** Let  $[a, b] = I \subseteq \mathbb{R}$  be an interval, then  $f: I \rightarrow \mathbb{R}$  is called *absolutely continuous* if, for every  $\varepsilon > 0$ , there exists  $\delta > 0$  such that, whenever  $(x_1, y_1), \dots, (x_m, y_m)$  are disjoint subintervals of  $I$  with  $\sum_{i=1}^m (y_i - x_i) < \delta$ , we have  $\sum_{i=1}^m |f(y_i) - f(x_i)| < \varepsilon$ .

It is known that absolute continuity is equivalent to being differentiable Lebesgue almost everywhere with a so-called *weak derivative*  $f'$  that satisfies  $f(x) = f(a) + \int_a^x f'(t) dt$  for all  $x \in [a, b]$ .

**Recap 2.13.** The *generalised Minkowski inequality* states that any Borel measurable function  $g: \mathbb{R}^2 \rightarrow \mathbb{R}$  we have that

$$\int_{\mathbb{R}} \left( \int_{\mathbb{R}} g(u, x) du \right)^2 dx \leq \left( \int_{\mathbb{R}} \left( \int_{\mathbb{R}} g^2(u, x) dx \right)^{1/2} du \right)^2.$$

To obtain bounds on the integrated squared bias, we will require smoothness conditions w.r.t. the  $L^2(\mathbb{R})$  norm.

**Definition 2.14.** Fix  $\beta, L > 0$  and let  $m := \lceil \beta \rceil - 1$ . The *Nikolski class*  $\mathcal{N}(\beta, L)$  consists of functions  $f: \mathbb{R} \rightarrow \mathbb{R}$  that are  $(m - 1)$  times differentiable and for which  $f^{(m-1)}$  is absolutely continuous with weak derivative  $f^{(m)}$  satisfying

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

The densities in  $\mathcal{N}(\beta, L)$  are denoted by  $\mathcal{F}_{\mathcal{N}}(\beta, L)$ .

**Proposition 2.15.** Fix  $\beta, L > 0$  and let  $K$  be a kernel of order  $\ell := \lceil \beta \rceil$ . Let  $\hat{f}_n$  denote the KDE with kernel  $K$  and bandwidth  $h$  based on  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f \in \mathcal{F}_{\mathcal{N}}(\beta, L)$ . Then we have

$$\int_{-\infty}^{\infty} \text{Bias}^2 \hat{f}_n(x) dx \leq \frac{L^2}{((\ell - 1)!)^2} \mu_{\beta}^2(K) h^{2\beta}.$$

*Proof.* **TODO:** write this out (integration + taylor expansion + 2x minkowski). □

Putting everything together, we obtain the following:

**Theorem 2.16.** Fix  $\beta, L > 0$ , and let  $K$  be a kernel of order  $\ell = \lceil \beta \rceil$  with  $R(K)$  and  $\mu_{\beta}(K)$  finite. Let  $\hat{f}_n$  be the KDE with kernel  $K$  and bandwidth  $h$  based on  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f \in \mathcal{F}_{\mathcal{N}}(\beta, L)$ . Then we have

$$\text{MISE } \hat{f}_n \leq \frac{R(K)}{nh} + \frac{L^2}{((\ell - 1)!)^2} \mu_{\beta}^2(K) h^{2\beta}.$$

In particular, fixing  $\alpha > 0$  and taking  $h = \alpha n^{-1/(2\beta+1)}$ , there exists  $C > 0$  independent of  $n$  such that

$$\sup_{f \in \mathcal{F}_{\mathcal{N}}(\beta, L)} \text{MISE } \hat{f}_{n,h,K} \leq C n^{-2\beta/(2\beta+1)}.$$

## 2.4 Bandwidth selection

The choice of bandwidth in the previous theorem is not practical since we have not specified  $\alpha$  and  $\beta$  is typically unknown.

### 2.4.1 Least squares cross validation

One possible approach is *least squares cross validation*. For this, note that minimising the MISE is equivalent to minimising

$$\text{MISE}(\hat{f}_n) - \int_{\mathbb{R}} f^2(x) dx = \mathbb{E} \left[ \int_{\mathbb{R}} \hat{f}_n^2(x) dx \right] - 2\mathbb{E} \left[ \int_{\mathbb{R}} \hat{f}_n(x) f(x) dx \right],$$

and it can be checked that an unbiased estimator for the above is given by

$$\text{LSCV}(h) := \int_{\mathbb{R}} \hat{f}_n^2(x) dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{n,-i}(X_i),$$

where  $\hat{f}_{n,-i}$  is the KDE based on all observations except  $X_i$ . We now choose  $h$  such that LSCV is minimised.



### 2.4.2 Lepski

**TODO:** write this subsection (**TODO:** understand this first)

## 2.5 Choice of kernel

To choose a kernel, we first fix the scale of the kernel by setting  $\mu_2(K) = 1$ . Now, by our bound on the MISE (theorem 2.16) it is reasonable to minimise  $R(K)$ , where for simplicity we assume that  $K$  is a nonnegative second-order kernel. The solution is the Epanechnikov kernel

$$K_E(x) := \frac{3}{4\sqrt{5}} \left(1 - \frac{x^2}{5}\right) \mathbb{1}_{|x| \leq \sqrt{5}},$$

and the ratio  $R(K_E)/R(K)$  is called the *efficiency* of a kernel  $K$ . We find that for different kernels, the efficiency is greater than 0.9, which shows that kernel shape does not affect efficiency greatly.

## 2.6 Multivariate density estimation

The general  $d$ -dimensional KDE is

$$\hat{f}_n(x) := \frac{1}{n\sqrt{\det H}} \sum_{i=1}^n K(H^{-1/2}(x - X_i)),$$

where  $H$  is a symmetric positive-definite bandwidth matrix (often chosen to be diagonal or a multiple of  $I$ ). If  $H = h^2 I$ , it can be shown that, under an appropriate definition of a  $\beta$  smoothness class, we have an optimal bandwidth of order  $n^{-1/(d+2\beta)}$ , and with this choice, a MISE of order  $n^{-2\beta/(d+2\beta)}$ . This is called the “curse of dimensionality”: the higher the dimension becomes, the harder nonparametric estimation gets.

### 3 Nonparametric regression

#### 3.1 Fixed and random design

In *fixed design*, we assume we have data  $x_1 \leq \dots \leq x_n$  and the response variables satisfy

$$Y_i := m(x_i) + \sqrt{v(x_i)}\varepsilon_i,$$

where the  $\varepsilon_i$  are independent, mean-zero random variables with  $\text{Var}(\varepsilon_i) = 1$ . The function  $m$  is called the *regression function*, and the function  $v$  is the *variance function*. If  $v$  is constant, the model is called *homoscedastic*, else it is called *heteroscedastic*.

In *random design*, we assume we have i.i.d. data pairs  $(X_i, Y_i)$  with

$$Y_i = m(X_i) + \sqrt{v(X_i)}\varepsilon_i,$$

where the  $\varepsilon_i$  are again independent with  $\mathbb{E}[\varepsilon_1|X_1] = 0$  and  $\text{Var}(\varepsilon_1|X_1) = 1$ . The regression function is given by  $m(x) = \mathbb{E}(Y_1|X_1 = x)$  and the variance function by  $v(x) = \text{Var}(Y_1|X_1 = x)$ .

#### 3.2 Local polynomial estimators

We will assume the fixed design setting.

**Definition 3.1.** Let  $K$  be a kernel,  $h > 0$  a bandwidth and  $p \in \mathbb{N}$ . Then the *local polynomial estimator of degree  $p$  with bandwidth  $h$  and kernel  $K$* , denoted  $\hat{m}_n(\cdot; p) \equiv \hat{m}_n(\cdot; p, h, K)$ , is constructed at  $x \in \mathbb{R}$  by fitting a polynomial  $p$  to the data using weighted least squares, where the pair  $(x_i, Y_i)$  is assigned weight  $K_h(x_i - x)$ .

To write this in formulas, define  $Q(u) := (1, u, \frac{u^2}{2}, \dots, \frac{u^p}{p!}) \in \mathbb{R}^{p+1}$  and  $Q_h(\cdot) = Q(\cdot/h)$ , we then have

$$\hat{m}_n(x; p) = \hat{\beta}_0, \quad \text{where } \hat{\beta} := \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i - \beta^\top Q_h(x_i - x))^2 K_h(x_i - x).$$

In matrix-vector notation, writing

$$X \equiv X(x; p, h) := \begin{pmatrix} Q_h(x_1 - x)^\top \\ \vdots \\ Q_h(x_n - x)^\top \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}, \quad Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n,$$

$$W \equiv W(x; h, K) := \text{diag}(K_h(x_1 - x), \dots, K_h(x_n - x)) \in \mathbb{R}^{n \times n},$$

we have

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} (Y - X\beta)^\top W (Y - X\beta).$$

By standard weighted least squares theory, we know that  $\hat{\beta}$  must satisfy  $X^\top W X \hat{\beta} = X^\top W Y$ .

**Proposition 3.2.** Suppose  $X^\top W X$  is positive definite. Then we have

$$\hat{\beta} = (X^\top W X)^{-1} X^\top W Y.$$

We will assume from here on that  $X^\top W X$  is indeed positive definite. In this case, since the entries of  $W$  and  $X$  are functions of  $x_i - x$ , we can write the local polynomial estimator in the form

$$\hat{m}_n(x) = n^{-1} \sum_{i=1}^n w(x, x_i) Y_i.$$

(?) The set of weights  $\{w(x, x_i)\}$  is called the *effective kernel* at  $x$ .

For  $p = 0$  and  $p = 1$ , there exist explicit formulas for the local polynomial estimator of degree  $p$ .

**Proposition 3.3** (Reproducing property). *Let  $\{w_p(x, x_i)\}$  denote the effective kernel of a local polynomial estimator of degree  $p$  based on data  $(x_1, Y_1), \dots, (x_n, Y_n)$ , and let  $R$  denote a polynomial of degree at most  $p$ . If  $X^\top W X$  is positive definite, then*

$$\frac{1}{n} \sum_{i=1}^n w_p(x, x_i) R(x_i) = R(x).$$

*Proof.* See example sheet 2 question 3. □

Before we can study the bias and variance of local polynomial estimators, we require the following lemma:

**Lemma 3.4.** *Let  $K$  be a kernel that vanishes outside  $[-1, 1]$ , and suppose that  $n^{-1} X^\top W X$  is positive definite with minimal eigenvalue  $\lambda_0 \equiv \lambda_{0,n,x} > 0$ . Then*

- (i)  $\sup_{x \in [0,1]} \max_{i=1, \dots, n} \frac{1}{n} |w(x, x_i)| \leq \frac{2\|K\|_\infty}{\lambda_0 n h};$
- (ii)  $n^{-1} \sum_{i=1}^n |w(x, x_i)| \leq \frac{2\|K\|_\infty}{\lambda_0 n h} \cdot \sum_{i=1}^n \mathbb{1}_{|x_i - x| \leq h};$
- (iii)  $w(x, x_i) = 0$  when  $|x_i - x| > h$ .

*Proof.* (i) Note that  $n^{-1} w(x, x_i)$  is the  $(0, i)$  entry of the matrix  $(X^\top W X)^{-1} X^\top W$ , and it is therefore less than the norm of the  $i$ -th column of  $(X^\top W X)^{-1} X^\top W$ . For  $x \in [0, 1]$  and  $i = 1, \dots, n$ , we find

$$\begin{aligned} \frac{1}{n} |w(x, x_i)| &\leq \|(X^\top W X)^{-1} Q_h(x_i - x) K_h(x_i - x)\| \stackrel{*}{\leq} \|K_h\|_\infty \|(X^\top W X)^{-1}\| \|Q_h(x_i - x)\| \mathbb{1}_{|x_i - x| \leq h} \\ &= \frac{\|K\|_\infty}{h} \frac{1}{\lambda_0 n} \|Q_h(x_i - x)\| \mathbb{1}_{|x_i - x| \leq h} \leq \frac{\|K\|_\infty}{\lambda_0 n h} \|Q(1)\| = \frac{\|K\|_\infty}{\lambda_0 n h} \left( \sum_{j=0}^p \frac{1}{(j!)^2} \right)^{1/2} \\ &\leq \frac{\|K\|_\infty}{\lambda_0 n h} e^{1/2} \leq \frac{2\|K\|_\infty}{\lambda_0 n h}. \end{aligned}$$

Here, the indicator function in  $\star$  appears because  $K$  vanishes outside  $[-1, 1]$ . Since the upper bound is independent of both  $x$  and  $i$ , the claim is proven.

(ii) Similarly as before, we find

$$\frac{1}{n} \sum_{i=1}^n |w(x, x_i)| \leq \frac{\|K\|_\infty}{\lambda_0 n h} \sum_{i=1}^n \|Q_h(x_i - x)\| \mathbb{1}_{|x_i - x| \leq h} \leq \frac{2\|K\|_\infty}{\lambda_0 n h} \sum_{i=1}^n \mathbb{1}_{|x_i - x| \leq h}.$$

(iii) This follows immediately from inequality  $\star$  in the proof of (i). □

Now, we can compute bounds on the variance and bias of our local polynomial estimator. For simplicity, we will assume  $x_i = i/n$ .

**Proposition 3.5.** *Assume the model  $Y_i = m(x_i) + v^{1/2} \varepsilon_i$  with  $m \in \mathcal{H}(\beta, L)$  on  $[0, 1]$  and  $\max_i v(x_i) \leq \sigma_{\max}^2$ . Let  $K$  be a kernel that vanishes outside  $[-1, 1]$ , and suppose that  $\lambda_0$ , the minimal eigenvalue of  $n^{-1} X^\top W X$ , is strictly positive. Then, for  $p \geq \lceil \beta \rceil - 1 =: \beta_0$  and for each  $x_0 \in [0, 1]$ ,  $n \in \mathbb{N}$  and  $h \geq 1/(2n)$ , we have*

$$\text{Var } \hat{m}_n(x_0; p) \leq \frac{16\|K\|_\infty^2 \sigma_{\max}^2}{\lambda_0^2 n h}, \quad |\text{Bias } \hat{m}_n(x_0; p)| \leq \frac{8L\|K\|_\infty}{\lambda_0 \beta_0!} h^\beta.$$

*Proof.* Using the previous lemma, we obtain

$$\begin{aligned}\text{Var } \hat{m}_n(x_0; p) &= \text{Var} \left( \frac{1}{n} \sum_{i=1}^n w(x_0, x_i) Y_i \right) = \frac{1}{n^2} \sum_{i=1}^n w(x_0, x_i)^2 \text{Var}(Y_i) \\ &\leq \frac{\sigma_{\max}^2}{n^2} \sum_{i=1}^n w(x_0, x_i)^2 \leq \sigma_{\max}^2 \left( \sup_{x \in [0,1]} \max_i \frac{1}{n} |w(x, x_i)| \right) \frac{1}{n} \sum_{i=1}^n |w(x_0, x_i)| \\ &\leq \frac{4\|K\|_{\infty}^2 \sigma_{\max}^2}{\lambda_0^2 n^2 h^2} \sum_{i=1}^n \mathbb{1}_{|i/n - x_0| \leq h}.\end{aligned}$$

Now, we have  $|i/n - x_0| \leq h \iff i \in [nx_0 - nh, nx_0 + nh]$ , and there are at most  $2nh + 1$  integers in this interval. Recalling that  $1 \leq 2nh$  we obtain

$$\text{Var } \hat{m}_n(x_0; p) \leq \frac{4(2nh + 1)\|K\|_{\infty}^2 \sigma_{\max}^2}{\lambda_0^2 n^2 h^2} \leq \frac{16\|K\|_{\infty}^2 \sigma_{\max}^2}{\lambda_0^2 nh}.$$

For the bias, we will first use a Taylor expansion combined with the reproducing property proposition 3.3. Firstly, since  $\frac{1}{n} \sum_{i=1}^n w(x_0, x_i) = 1$  by this property, we have

$$\text{Bias } \hat{m}_n(x_0; p) = \left( \frac{1}{n} \sum_{i=1}^n w(x_0, x_i) m(x_i) \right) - m(x_0) = \frac{1}{n} \sum_{i=1}^n w(x_0, x_i) \{m(x_i) - m(x_0)\}.$$

Now, we apply a Taylor expansion to write

$$m(x_i) - m(x_0) = P(x_i - x_0) + \frac{1}{\beta_0!} m^{(\beta_0)}(x_0 + \tau_i(x_i - x_0))(x_i - x_0)^{\beta_0},$$

where  $P$  has degree at most  $\beta_0 - 1 < p$  and a constant coefficient equal to 0, and  $\tau_i \in [0, 1]$ . By the reproducing property we have  $n^{-1} \sum_{i=1}^n w(x_0, x_i) P(x_i - x_0) = P(0) = 0$ , so we obtain

$$\begin{aligned}\text{Bias } \hat{m}_n(x; p) &= \frac{1}{n} w(x_0, x_i) \sum_{i=1}^n w(x_0, x_i) \frac{m^{(\beta_0)}(x_0 + \tau_i(x_i - x_0))}{\beta_0!} (x_i - x_0)^{\beta_0} \\ &= \frac{1}{n} w(x_0, x_i) \sum_{i=1}^n w(x_0, x_i) \frac{m^{(\beta_0)}(x_0 + \tau_i(x_i - x_0)) - m^{(\beta_0)}(x_0)}{\beta_0!} (x_i - x_0)^{\beta_0},\end{aligned}$$

where the last line follows again from the reproducing property. Now we apply the fact that  $m \in \mathcal{H}(\beta, L)$  with the previous lemma and find

$$\begin{aligned}|\text{Bias } \hat{m}_n(x; p)| &\leq \frac{L}{n} \sum_{i=1}^n |w(x_0, x_i)| \frac{|x_i - x_0|^{\beta}}{\beta_0!} = \frac{L}{n} \sum_{i=1}^n |w(x_0, x_i)| \frac{|x_i - x_0|^{\beta}}{\beta_0!} \mathbb{1}_{|x_i - x_0| \leq h} \\ &\leq \frac{Lh^{\beta}}{n\beta_0!} \sum_{i=1}^n |w(x_0, x_i)| \leq \frac{8L\|K\|_{\infty}}{\lambda_0\beta_0!} h^{\beta}.\end{aligned}$$

□

So again, we have a variance term of order  $1/nh$  and a bias term of order  $h^{\beta}$ . However, both our bounds depend on  $\lambda_0$ , which depends on both  $n$  and  $x$ .

**Proposition 3.6.** *Suppose  $x_i = i/n$  and that  $K(u) \geq K_0 \mathbb{1}_{|u| \leq \Delta}$  for some  $K_0, \Delta > 0$ . Then, for  $n \geq 2$  and  $h \leq \frac{1}{4\Delta}$ ,*

$$\inf_{x \in [0,1]} \lambda_{0,n,x} \geq K_0 \inf_{v \in S^p} \min \left\{ \int_0^{\Delta} (v^{\top} Q(u))^2 du, \int_{-\Delta}^0 (v^{\top} Q(u))^2 du \right\} - \frac{(4\Delta + 2)K_0 e^{\Delta^2}}{nh}.$$

*Proof.* Since  $\lambda_{0,n,x} = \inf_{v \in S^p} v^\top \left( \frac{1}{n} X^\top W X \right) v$ , we will try to bound this quantity from below.

Let  $u_i := \frac{x_i - x}{h}$  for  $i = 1, \dots, n$ , so  $u_1 \leq \frac{x_1}{h} = \frac{1}{nh}$ , and let  $u_0 := 0$ .

First, we assume  $x < 1 - h\Delta$ , so that  $u_n > \frac{1 - (1 - h\Delta)}{h} = \Delta$ . Then, for any  $v \in S^p$ , we have

$$\begin{aligned} v^\top \left( \frac{1}{n} X^\top W X \right) &= \frac{1}{n} (Xv)^\top W (Xv) = \frac{1}{n} \sum_{i=1}^n (Xv)_i^2 W_{ii} \\ &= \frac{1}{nh} \sum_{i=1}^n (v^\top Q(u_i))^2 K(u_i) \geq \frac{K_0}{nh} \sum_{i=1}^n (v^\top Q(u_i))^2 \mathbb{1}_{u_i \in [0, \Delta]}. \end{aligned}$$

This is a Riemann sum (since  $u_{i+1} - u_i = 1/(nh)$ ) and we will try to approximate it by the corresponding integral  $K_0 \int_0^\Delta (v^\top Q(u))^2 du$ .

To do this, note that we have for  $u \in [0, \Delta]$  that

$$\|Q(u)\| \leq \left\{ \sum_{\ell=0}^p \frac{u^{2\ell}}{(\ell!)} \right\}^{1/2} \leq \left( \sum_{\ell=0}^\infty \frac{u^{2\ell}}{\ell!} \right)^{1/2} \leq e^{\Delta^2/2}.$$

Furthermore, for  $a, b \geq 0$ , it can be shown that

$$|b^\ell - a^\ell| \leq \max(1, \ell/2)(b^{\ell-1} + a^{\ell-1})|b - a|,$$

and using this we obtain for  $u, v \in [0, \Delta]$  that

$$\begin{aligned} \|Q(u) - Q(v)\| &\leq \left\{ \sum_{\ell=1}^p \frac{\max(1, \ell^2/4)(u^{\ell-1} + v^{\ell-1})^2(u-v)^2}{(\ell!)^2} \right\}^{1/2} \\ &\leq |u - v| \left\{ \sum_{\ell=1}^p \frac{\max(1, \ell^2/4)(2\Delta^{\ell-1})^2}{(\ell!)^2} \right\}^{1/2} \leq |u - v| \left\{ \sum_{\ell=1}^p \frac{\max(4, \ell^2)\Delta^{2\ell-2}}{(\ell!)^2} \right\}^{1/2} \\ &\stackrel{*}{\leq} 2|u - v| \left\{ \sum_{\ell=0}^{p-1} \frac{\Delta^{2\ell}}{(\ell!)^2} \right\}^{1/2} \leq 2|u - v| \left\{ \sum_{\ell=0}^\infty \frac{\Delta^{2\ell}}{\ell!} \right\}^{1/2} = 2e^{\Delta^2/2}|u - v|, \end{aligned}$$

where in  $\star$  we used that  $\max(4, \ell^2) \leq 4\ell^2$ .

Now we estimate the difference between the Riemann sum and the corresponding integral. Write  $i_- := \min\{i : u_i > 0\}$  and  $i_+ := \min\{i : u_i > \Delta\}$ . Then

$$\begin{aligned} \frac{1}{nh} \sum_{i=1}^n (v^\top Q(u_i))^2 \mathbb{1}_{u_i \in [0, \Delta]} &= \sum_{i=i_-}^{i_+-1} \int_{u_{i-1}}^{u_i} (v^\top Q(u_i))^2 du \\ &\leq \sum_{i=i_-}^{i_+} \int_{\max(u_{i-1}, 0)}^{\min(u_i, \Delta)} (v^\top Q(u_i))^2 du + \frac{1}{nh} \sup_{u \in [0, \Delta]} (v^\top Q(u))^2. \end{aligned}$$

(Note that the last integral, from  $u_{i-1}$  to  $\Delta$ , is 0 since ) Therefore we obtain

$$\begin{aligned} &\left| \frac{1}{nh} \sum_{i=1}^n (v^\top Q(u_i))^2 \mathbb{1}_{u_i \in [0, \Delta]} - \int_0^\Delta (v^\top Q(u))^2 du \right| \\ &\leq \left| \sum_{i=i_-}^{i_+} \int_{\max(u_{i-1}, 0)}^{\min(u_i, \Delta)} (v^\top Q(u_i))^2 - (v^\top Q(u))^2 du \right| + \frac{1}{nh} \sup_{u \in [0, \Delta]} (v^\top Q(u))^2. \end{aligned} \quad (5)$$

Now note that by Cauchy-Schwarz since  $\|v\| = 1$  we have

$$\begin{aligned} |(v^\top Q(u_i))^2 - (v^\top Q(u))^2| &= |(v^\top Q(u_i) + v^\top Q(u))|(v^\top Q(u_i) - v^\top Q(u))| \\ &\leq (\|Q(u_i)\| + \|Q(u)\|)\|Q(u_i) - Q(u)\| \\ &\leq 2e^{\Delta^2/2} \cdot 2e^{\Delta^2/2}|u_i - u| = 4e^{\Delta^2}|u_i - u|, \end{aligned}$$

and therefore

$$(5) \leq 4e^{\Delta^2} \sum_{i=i_-}^{i_+} \int_{\max(u_{i-1}, 0)}^{\min(u_i, \Delta)} (u_i - u) du + \frac{e^{\Delta^2}}{nh} \leq \frac{(4\Delta + 1)e^{\Delta^2}}{nh},$$

which concludes the case  $x < 1 - h\Delta$ .

Suppose  $x \geq 1 - h\Delta$ , then we have  $u_1 \leq -\Delta$  and  $u_n \geq 0$ , and we can apply very similar arguments to reach the desired conclusion.  $\square$

Now that we have bounds on the variance and bias, we can prove our uniform bound:

**Theorem 3.7.** *Under the conditions of the previous two propositions, if we choose  $h = \alpha n^{-1/(2\beta+1)}$  for some  $\alpha > 0$ , there exists  $n_0 \in \mathbb{N}, C > 0$ , depending only on  $\beta, L, \alpha, K, \sigma_{\max}^2$ , such that*

$$\sup_{m \in \mathcal{H}(\beta, L)} \sup_{x_0 \in [0, 1]} \mathbb{E} \left[ \{\hat{m}_n(x_0; p) - m(x_0)\}^2 \right] \leq C n^{-2\beta/(2\beta+1)}.$$

### 3.3 Splines

#### 3.3.1 Cubic splines

Let  $n \geq 3$  and  $a \leq x_1 < \dots < x_n \leq b$ .

**Definition 3.8.** A function  $g: [a, b] \rightarrow \mathbb{R}$  is called a *cubic spline* with *knots*  $x_1, \dots, x_n$  if

1.  $g$  is a cubic polynomial on each interval  $(a, x_1), (x_1, x_2), \dots, (x_n, b)$ ;
2.  $g \in C^2[a, b]$ .

Furthermore,  $g$  is called *natural* if it is linear on  $[a, x_1]$  and  $[x_n, b]$ , (i.e.,  $g''(a) = g'''(a) = g''(b) = g'''(b) = 0$ ).

We often represent a natural cubic spline by the vectors  $\mathbf{g} \in \mathbb{R}^n$  with  $g_i = g(x_i)$ , and  $\boldsymbol{\gamma} \in \mathbb{R}^{n-2}$  with  $\gamma_i = g''(x_i)$  (excluding  $\gamma_1$  and  $\gamma_n$ ). Writing  $h_i := x_{i+1} - x_i$  we have for  $x \in [x_i, x_{i+1}]$  that

$$g(x) = \frac{(x - x_i)g_{i+1} - (x_{i+1} - x)g_i}{h_i} - \frac{1}{6}(x - x_i)(x_{i+1} - x) \left\{ \left( 1 + \frac{x - x_i}{h_i} \right) \gamma_{i+1} + \left( 1 + \frac{x_{i+1} - x}{h_i} \right) \gamma_i \right\}.$$

**Proposition 3.9.** *Given  $\mathbf{g} \in \mathbb{R}^n$ , there exists a unique natural cubic spline  $g$  with knots at  $x_1, \dots, x_n$  satisfying  $g(x_i) = g_i$  for all  $i$ , and there exists  $K \succeq 0$  (depending on  $x_1, \dots, x_n$ ) such that*

$$\int_a^b g''(x)^2 dx = \mathbf{g}^\top K \mathbf{g}.$$

We call  $g$  the *natural cubic spline interpolant to  $\mathbf{g}$  at  $x_1, \dots, x_n$* .

**Definition 3.10.** We define  $\mathcal{S}_2[a, b]$  as the set of real-valued functions on  $[a, b]$  with an absolutely continuous first derivative. For  $f \in \mathcal{S}_2[a, b]$ , we define the *roughness* of  $f$  by  $R(f) := \int_a^b f''(x)^2 dx$ .

**Proposition 3.11.** *For any  $\mathbf{g} \in \mathbb{R}^n$ , the natural cubic spline interpolant to  $\mathbf{g}$  at  $x_1, \dots, x_n$  is the unique minimiser of  $R$  over all  $g \in \mathcal{S}_2[a, b]$  that satisfy  $g(x_i) = g_i$  for all  $i$ .*

### 3.3.2 Natural cubic smoothing splines

Consider the nonparametric regression model  $Y_i = g(x_i) + \sigma\varepsilon_i$ , where  $\mathbb{E}[\varepsilon_i] = 0$  and  $\text{Var}[\varepsilon_i] = 1$ . A way to estimate a nonparametric regression function is to balance data fidelity against roughness of the curve, which can be done by minimising

$$S_\lambda(g) := \sum_{i=1}^n (Y_i - g(x_i))^2 + \lambda R(g),$$

where  $\lambda > 0$  is a regularisation parameter. For small  $\lambda$ , this is almost an exact fit to the data. For large  $\lambda$ , we are basically minimising  $|g''|$ , which means we will approximate the linear regression fit.

**Theorem 3.12.** *For each  $\lambda \in (0, \infty)$ , there exists a unique minimiser  $\hat{g}_\lambda$  of  $S_\lambda$  over  $\mathcal{S}_2[a, b]$ . It is the natural cubic spline with knots at  $x_1, \dots, x_n$  and  $\mathbf{g} = (I + \lambda K)^{-1} \mathbf{Y}$ .*

*Proof.* If  $g$  is not a natural cubic spline, we know that the natural cubic spline  $g^*$  which interpolates  $g(x_1), \dots, g(x_n)$  at  $x_1, \dots, x_n$  has a strictly lower value of  $S_\lambda$ , so we know the minimiser must be a natural cubic spline.

If  $g$  is a natural cubic spline, then there exists  $K \succeq 0$  such that

$$\begin{aligned} S_\lambda(g) &= (\mathbf{Y} - \mathbf{g})^\top (\mathbf{Y} - \mathbf{g}) + \lambda \mathbf{g}^\top K \mathbf{g} \\ &= \mathbf{g}^\top (I + \lambda K) \mathbf{g} - 2\mathbf{Y}^\top \mathbf{g} + \mathbf{Y}^\top \mathbf{Y}, \end{aligned}$$

and by “completing the square” we write, for some  $Z$  independent of  $\mathbf{g}$ ,

$$S_\lambda(g) = (\mathbf{g} - (I + \lambda K)^{-1} \mathbf{Y})^\top (I + \lambda K) (\mathbf{g} - (I + \lambda K)^{-1} \mathbf{Y}) + Z$$

Since  $I + \lambda K$  is positive definite it follows that  $\mathbf{g} = (I + \lambda K)^{-1} \mathbf{Y}$  gives the minimiser.  $\square$

The function  $\hat{g}_\lambda$  is called the *natural cubic smoothing spline* for data  $(x_1, Y_1), \dots, (x_n, Y_n)$  and smoothing parameter  $\lambda$ .

### 3.3.3 Choice of smoothing parameter

We are left with the question of how to choose the smoothing parameter  $\lambda$ . A standard method is to minimise the cross-validation score

$$\text{CV}(\lambda) := \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{g}_{-i, \lambda}(x_i))^2,$$

where  $\hat{g}_{-i, \lambda}$  is the natural cubic smoothing spline for all data points except  $(x_i, Y_i)$ . It seems like computing  $\text{CV}(\lambda)$  requires the computation of  $n$  natural cubic smoothing splines, but it turns out that this is not the case:

**Proposition 3.13.** *Write  $A(\lambda) = (I + \lambda K)^{-1}$ , then we have*

$$\text{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \hat{g}_\lambda(x_i)}{1 - A_{ii}(\lambda)} \right)^2.$$

*Proof.* Example sheet 3.  $\square$

In the above formula, we can consider the quantity  $A_{ii}(\lambda)$  as the “leverage” of the  $i$ -th observation. In the *generalised cross-validation* score, we give every observation equal leverage: it is defined as

$$\text{GCV}(\lambda) := \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \hat{g}_\lambda(x_i)}{1 - n^{-1} \text{tr } A(\lambda)} \right)^2.$$

**Regression splines** There are many different types of splines and different directions to go in. For example, one disadvantage of natural cubic smoothing splines is that we have a “parameter” of dimension  $n$  to estimate (namely, the vector  $\mathbf{g}$ ). We can also reduce the number of knots to  $\xi_1, \dots, \xi_K$  and locate  $\xi_k$  at the  $\left(\frac{k+1}{K+2}\right)$ -th sample quantile of  $x_1, \dots, x_n$ . Splines of order  $p$  can then be expanded in the *truncated power series basis*

$$1, x, x^2, \dots, x^p, (x - \xi_1)_+^p, \dots, (x - \xi_K)_+^p.$$

Therefore we can minimise the residual error over all polynomials in the span of this basis, which gives a parameter in  $\mathbb{R}^{p+1+K}$  to estimate using least squares. The solution is called a *regression spline*. Here,  $K$  plays the role of the smoothing parameter.



## 4 Minimax lower bounds

We seek lower bounds on the worst-case risk of any procedure, which provide a ‘benchmark’ against which we can measure the performance of a proposed method.

### 4.1 Reduction to testing

We will assume our parameter space  $(\Theta, d)$  is a metric space or a semi-metric space (where we don’t require that  $d(\vartheta, \vartheta') = 0 \implies \vartheta = \vartheta'$ ). We denote our collections of distributions depending on our parameters by  $\{P_\vartheta \mid \vartheta \in \Theta\}$ , which are probability measures on some measurable space  $(\mathcal{X}, \mathcal{A})$ .

Now, we let  $(\Omega, \mathcal{F})$  be any measurable space with a collection of probability measures  $\{\mathbb{P}_\vartheta \mid \vartheta \in \Theta\}$  and a measurable function  $X: \Omega \rightarrow \mathcal{X}$  so that  $X \sim P_\vartheta$  on  $(\Omega, \mathcal{F}, \mathbb{P}_\vartheta)$ . Let  $\hat{\Theta}$  denote the set of possible estimators for  $\vartheta$ , i.e., all measurable functions  $\mathcal{X} \rightarrow \Theta$ .

Now, suppose we wish to estimate  $\vartheta$  with a loss function of the form

$$L(\vartheta', \vartheta) = g(d(\vartheta', \vartheta)) \quad g \text{ increasing, } \vartheta', \vartheta \in \Theta.$$

We then define the *minimax risk* as

$$\mathcal{M} := \inf_{\hat{\vartheta} \in \hat{\Theta}} \sup_{\vartheta \in \Theta} \mathbb{E}_\vartheta L(\hat{\vartheta}(X), \vartheta),$$

i.e., the lowest worst-case estimated loss of any possible estimator  $\hat{\vartheta}$ .

**Example 4.1.** Suppose we are trying to estimate the mean of a normally distributed random variable with variance 1, and we have a sample  $(X_1, \dots, X_n)$ . Then  $\Theta = \mathbb{R}$  with the Euclidian distance,  $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^n, \mathcal{B}\mathbb{R}^n)$ , and  $P_\vartheta(F) = \int_F f_\vartheta(x_1) \cdots f_\vartheta(x_n) d\lambda^n(x)$ , where  $f_\vartheta$  is the density function of a  $N(\vartheta, 1)$  distribution.

Now, let  $X_\vartheta: (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^n, \mathcal{B}\mathbb{R}^n)$  be a  $N_n(\vartheta \mathbf{1}, I_n)$  distributed random variable. In this case,  $\mathbb{P}_\vartheta$  is a probability measure on  $\Omega$  such that  $P_\vartheta(F) = \mathbb{P}_\vartheta(X \in F)$ .

Let  $\hat{\Theta}$  denote the set of all estimators of  $\vartheta$ , which are functions  $\mathbb{R}^n \rightarrow \mathbb{R}$ . Our loss function could simply be  $L(\vartheta', \vartheta) = |\vartheta' - \vartheta|$  or  $L(\vartheta', \vartheta) = (\vartheta' - \vartheta)^2$ .

For  $M \in \mathbb{N}$ , let  $[M] := \{1, \dots, M\}$ , and let  $\hat{\mathcal{T}}$  denote the set of measurable functions  $\mathcal{X} \rightarrow [M]$ . Given any  $\vartheta_1, \dots, \vartheta_M \in \Theta$  and  $\hat{\vartheta} \in \hat{\Theta}$ , we can define  $T_{\hat{\vartheta}} \in \hat{\mathcal{T}}$  by

$$T_{\hat{\vartheta}}(x) := \arg \min_{j \in [M]} d(\hat{\vartheta}(x), \vartheta_j),$$

where we pick the smallest  $j$  in case of a tie. Intuitively, we are simply approximating  $\hat{\vartheta}$  by the closest  $\vartheta_j$ . Now, we will lower-bound the minimax risk by an expression that only depends on estimators in  $\hat{\mathcal{T}}$ .

Writing  $\eta = \frac{1}{2} \min_{j,k} d(\vartheta_j, \vartheta_k)$ , we can lower-bound the worst-case loss of any fixed estimator  $\hat{\vartheta}$  by

$$\begin{aligned} \sup_{\vartheta \in \Theta} \mathbb{E}_\vartheta L(\hat{\vartheta}(X), \vartheta) &\geq \max_{j \in [M]} \mathbb{E}_{\vartheta_j} g(d(\hat{\vartheta}, \vartheta_j)) \\ &= \max_{j \in [M]} \mathbb{E}_{\vartheta_j} \left\{ g(d(\hat{\vartheta}, \vartheta_j)) \mathbb{1}_{T_{\hat{\vartheta}} \neq j} \right\} \\ &\stackrel{\star}{\geq} g(\eta) \max_{j \in [M]} \mathbb{E}_{\vartheta_j} \mathbb{1}_{T_{\hat{\vartheta}} \neq j} \\ &= g(\eta) \max_{j \in [M]} P_{\vartheta_j}(T_{\hat{\vartheta}} \neq j), \end{aligned}$$

where  $\star$  holds because if  $T_{\hat{\vartheta}} \neq j$ , then  $d(\hat{\vartheta}(x), \vartheta_j) \geq \eta$ .

We therefore have

$$\begin{aligned}\mathcal{M} &\geq g(\eta) \inf_{\vartheta \in \hat{\Theta}} \max_{j \in [M]} P_{\vartheta_j}(T_{\hat{\vartheta}} \neq j) \geq g(\eta) \inf_{T \in \hat{\mathcal{T}}} \max_{j \in [M]} P_{\vartheta_j}(T \neq j) \\ &= g(\eta) \left\{ 1 - \sup_{T \in \hat{\mathcal{T}}} \min_{j \in [M]} P_{\vartheta_j}(T = j) \right\} \geq g(\eta) \left\{ 1 - \sup_{T \in \hat{\mathcal{T}}} \frac{1}{M} \sum_{j=1}^M P_{\vartheta_j}(T = j) \right\}.\end{aligned}$$

Therefore, we have now reduced the problem of lower-bounding  $\mathcal{M}$  to the problem of upper-bounding  $\sup_{T \in \hat{\mathcal{T}}} \frac{1}{M} \sum_{j=1}^M P_{\vartheta_j}(T = j)$ , which is a testing problem. We repeat the main result:

$$\mathcal{M} \geq g(\eta) \left\{ 1 - \sup_{T \in \hat{\mathcal{T}}} \frac{1}{M} \sum_{j=1}^M P_{\vartheta_j}(T = j) \right\}, \quad (6)$$

where  $\eta = \frac{1}{2} \min_{j,k} d(\vartheta_j, \vartheta_k)$ .

## 4.2 Divergences

**Definition 4.2.** Let  $\mu, \nu$  be measures on  $(\mathcal{X}, \mathcal{A})$ . We say that  $\mu$  is *absolutely continuous* w.r.t.  $\nu$ , notation  $\mu \ll \nu$ , if

$$\nu(A) = 0 \implies \mu(A) = 0.$$

We say that  $\mu, \nu$  are *mutually singular*, notation  $\mu \perp \nu$ , if there exists  $A \in \mathcal{A}$  such that  $\mu(A) = 0$  and  $\nu(A^c) = 0$ .

Note that mutual singularity means that  $\mu$  “lives on”  $A^c$ , while  $\nu$  “lives on”  $A$ .

**Theorem 4.3** (Lebesgue). *If  $\mu, \nu$  are  $\sigma$ -finite measures on  $(\mathcal{X}, \mathcal{A})$ , then there exists measures  $\mu_{\text{ac}}$  and  $\mu_{\text{sing}}$  on  $(\mathcal{X}, \mathcal{A})$  such that  $\mu$  can be decomposed as  $\mu = \mu_{\text{ac}} + \mu_{\text{sing}}$ , where  $\mu_{\text{ac}} \ll \nu$  and  $\mu_{\text{sing}} \perp \nu$ . Furthermore, this decomposition is unique.*

Let  $f: (0, \infty) \rightarrow \mathbb{R}$  be convex. Then for any  $y > 0$ , the function  $x \mapsto \frac{f(x) - f(y)}{x - y}$  is increasing on  $(y, \infty)$  (this is easy to check). Furthermore, we have

$$\lim_{x \rightarrow \infty} \frac{f(x) - f(y)}{x - y} = \lim_{x \rightarrow \infty} \frac{f(x)}{x - y} - \lim_{x \rightarrow \infty} \frac{f(y)}{x - y} = \lim_{x \rightarrow \infty} \frac{f(x)}{x},$$

so in particular the limit is independent of  $y$  and we can define the *maximal slope* of  $f$  by

$$M_f := \lim_{x \rightarrow \infty} \frac{f(x)}{x} \in (-\infty, \infty].$$

We define  $f(0) := \lim_{x \downarrow 0} f(x) \in (-\infty, \infty]$  (a convex function is continuous on an open interval, so this limit exists). In this case, we have

$$f(x + y) = f(x) + y \frac{f(x + y) - f(x)}{y} \leq f(x) + y M_f \quad \forall x, y \geq 0.$$

**Definition 4.4.** Given a convex function  $f: (0, \infty) \rightarrow \mathbb{R}$  with  $f(1) = 0$ , and probability measures  $P, Q$  on a measurable space  $(\mathcal{X}, \mathcal{A})$ , we define the  $f$ -divergence

$$\text{Div}_f(P, Q) := \int_{\mathcal{X}} f\left(\frac{dP_{\text{ac}}}{dQ}\right) dQ + P_{\text{sing}}(\mathcal{X}) \cdot M_f.$$

By Jensen's inequality we have

$$\begin{aligned}\text{Div}_f(P, Q) &= \int_{\mathcal{X}} f\left(\frac{dP_{\text{ac}}}{dQ}\right) dQ + P_{\text{sing}}(\mathcal{X}) \cdot M_f \geq f\left(\int_{\mathcal{X}} \frac{dP_{\text{ac}}}{dQ} dQ\right) + P_{\text{sing}}(\mathcal{X}) \cdot M_f \\ &= f(P_{\text{ac}}(\mathcal{X})) + P_{\text{sing}}(\mathcal{X}) \cdot M_f = f(P_{\text{ac}}(\mathcal{X}) + P_{\text{sing}}(\mathcal{X})) = f(P(\mathcal{X})) = f(1) = 0,\end{aligned}$$

so  $f$ -divergences are nonnegative.

**Example 4.5.** 1. If  $f(x) = x \log x$ , then  $M_f = \infty$ . If  $P \ll Q$  (i.e.,  $P_{\text{sing}} = 0$ ), then we have

$$\text{Div}_f(P, Q) = \int_{\mathcal{X}} \frac{dP}{dQ} \log\left(\frac{dP}{dQ}\right) dQ = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP,$$

and otherwise  $\text{Div}_f(P, Q) = 0$ .

This divergence is known as the *Kullback-Leibler* divergence from  $Q$  to  $P$ , denoted  $\text{KL}(P, Q)$ .

If  $P \ll Q$  and  $P$  and  $Q$  have densities  $p$  and  $q$  w.r.t. a measure  $\mu$ , we have  $\text{KL}(P, Q) = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu$ .

2. If  $f(x) = x^2 - 1$ , then  $M_f = \infty$ . If  $P \ll Q$  we have

$$\text{Div}_f(P, Q) = \int_{\mathcal{X}} \left(\frac{dP}{dQ}\right)^2 dQ - \int_{\mathcal{X}} dQ = \int_{\mathcal{X}} \left(\frac{dP}{dQ}\right)^2 dQ - 1,$$

and otherwise  $\text{Div}_f(P, Q) = \infty$ .

This divergence is known as the  $\chi^2$  divergence from  $Q$  to  $P$ , denoted  $\chi^2(P, Q)$ .

If  $P \ll Q$  and  $P$  and  $Q$  have densities  $p$  and  $q$  w.r.t. a measure  $\mu$ , we have  $\chi^2(P, Q) = \int_{\mathcal{X}} \frac{p^2}{q} d\mu - 1$ .

3. If  $f(x) = (\sqrt{x} - 1)^2 = x + 1 - 2\sqrt{x}$  (note that this is convex since  $\sqrt{x}$  is concave), then  $M_f = 1$  and therefore

$$\text{Div}_f(P, Q) = \int_{\mathcal{X}} \left(\sqrt{\frac{dP_{\text{ac}}}{dQ}} - 1\right)^2 dQ + P_{\text{sing}}(\mathcal{X}) =: H^2(P, Q),$$

the *squared Hellinger distance* between  $P$  and  $Q$ .

If  $P$  and  $Q$  have densities  $p$  and  $q$  w.r.t. a  $\sigma$ -finite measure  $\mu$ , then  $H^2(P, Q) = \int_{\mathcal{X}} (\sqrt{p} - \sqrt{q})^2 d\mu$  (example sheet).

4. If  $f(x) = \frac{|x-1|}{2}$ , then  $M_f = \frac{1}{2}$  and we have

$$\text{Div}_f(P, Q) \stackrel{\text{TODO:}}{=} \sup_{A \in \mathcal{A}} |P(A) - Q(A)| =: \text{TV}(P, Q),$$

the *total variation* divergence between  $P$  and  $Q$ .

All  $f$ -divergences are *jointly convex*: for all  $\lambda \in [0, 1]$  we have (see Example Sheet)

$$\text{Div}_f((1 - \lambda)P_1 + \lambda P_2, (1 - \lambda)Q_1 + \lambda Q_2) \leq (1 - \lambda) \text{Div}_f(P_1, Q_1) + \lambda \text{Div}_f(P_2, Q_2)$$

### 4.3 Le Cam's two point lemma

Plugging  $M = 1$  into eq. (6) yields the trivial result  $\mathcal{M} \geq 0$ . Surprisingly, when we plug in  $M = 2$ , we obtain Le Cam's two point lemma, which can often provide optimal rates for estimating real-valued parameters (though not optimal constants).

**Lemma 4.6.** *In the set-up of section 4.1, we have for any  $\vartheta_1, \vartheta_2 \in \Theta$  that*

$$\mathcal{M} \geq \frac{g(\eta)}{2} \{1 - \text{TV}(P_{\vartheta_1}, P_{\vartheta_2})\}.$$

*Proof.* For  $T \in \hat{\mathcal{T}}_2$ , let  $A := T^{-1}(\{1\})$ , then we have by eq. (6)

$$\begin{aligned} \mathcal{M} &\geq g(\eta) \left\{ 1 - \sup_{T \in \hat{\mathcal{T}}_2} \frac{P_{\vartheta_1}(T=1) + P_{\vartheta_2}(T=2)}{2} \right\} = \frac{g(\eta)}{2} \left\{ 1 - \sup_{T \in \hat{\mathcal{T}}_2} (P_{\vartheta_1}(T=1) - P_{\vartheta_2}(T=1)) \right\} \\ &\geq \frac{g(\eta)}{2} \{1 - \text{TV}(P_{\vartheta_1}, P_{\vartheta_2})\}. \end{aligned}$$

□

**Example 4.7.** Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\vartheta, 1)$  for some  $\vartheta \in \mathbb{R}$ . Let  $\vartheta_1 = 0$  and  $\vartheta_2 = cn^{-1/2}$  for some  $c > 0$ , and let  $P_{\vartheta_j} := N(\vartheta_j, 1)$  for  $j = 1, 2$ . Then we have (where  $\star$  denotes equalities that will be proved on the example sheet):

$$\text{TV}(P_{\vartheta_1}^{\times n}, P_{\vartheta_2}^{\times n}) \leq \sqrt{\frac{\text{KL}(P_0^{\times n}, P_{cn^{-1/2}}^{\times n})}{2}} \stackrel{\star}{=} \sqrt{\frac{n}{2} \text{KL}(P_0, P_{cn^{-1/2}})} \stackrel{\star}{=} \frac{c}{2}.$$

Plugging this into Le Cam's two point lemma yields (using the squared error loss  $L(x, y) = (x - y)^2$  that

$$\mathcal{M} = \inf_{\hat{\vartheta} \in \hat{\Theta}} \sup_{\vartheta \in \mathbb{R}} \mathbb{E}_{\vartheta} \left[ (\hat{\vartheta}(X_1, \dots, X_n) - \vartheta)^2 \right] \geq \sup_{c > 0} \frac{c^2}{8n} \left( 1 - \frac{c}{2} \right) = \frac{2}{27n}.$$

In this problem, it can be shown that  $\mathcal{M} = 1/n$ , so Le Cam's two point lemma does give the optimal rate, but it does not give the optimal constant.

**Example 4.8.** **TODO:** Understand remark 45 in lecture notes

### 4.4 Assouad's lemma

**Lemma 4.9.** *Let  $m \in \mathbb{N}$ ,  $\Phi := \{0, 1\}^m$  and  $\{P_{\varphi} \mid \varphi \in \Phi\}$  a family of probability measures on  $(\mathcal{X}, \mathcal{A})$ . Write  $\varphi \sim \varphi'$  when  $\varphi$  and  $\varphi'$  differ in precisely one coordinate and  $\varphi \sim_j \varphi'$  when this coordinate is the  $j$ -th.*

*Suppose the loss function is of the form*

$$L(\varphi', \varphi) = \sum_{j=1}^m L_j(\varphi', \varphi) = \sum_{j=1}^m g(d_j(\varphi', \varphi)),$$

*where  $d_1, \dots, d_m$  are semimetrics satisfying  $d_j(\varphi', \varphi) \geq \alpha_j$  whenever  $\varphi \sim_j \varphi'$ , and where  $g$  is increasing and satisfies  $g(x + y) \leq A\{g(x) + g(y)\}$  for all  $x, y \geq 0$  and some  $A > 0$ . Then*

$$\inf_{\hat{\varphi} \in \hat{\Phi}} \max_{\varphi \in \Phi} \mathbb{E}_{\varphi} L(\hat{\varphi}, \varphi) \geq \frac{1}{2A} \left\{ 1 - \max_{\varphi \sim \varphi'} \text{TV}(P_{\varphi}, P_{\varphi'}) \right\} \sum_{j=1}^m g(\alpha_j).$$

*Proof.* For any  $\varphi \in \Phi, j \in [m]$ , let  $\varphi^{[j]}$  be the unique element of  $\Phi$  with  $\varphi \sim_j \varphi^{[j]}$ . Letting  $\hat{\Phi}$  denote the set of measurable functions from  $\mathcal{X}$  to  $\Phi$ , we have

$$\max_{\varphi \in \Phi} \mathbb{E}_{\varphi} L(\hat{\varphi}, \varphi) \geq \frac{1}{2^m} \sum_{\varphi \in \Phi} \sum_{j=1}^m \mathbb{E}_{\varphi} L_j(\hat{\varphi}, \varphi) = \frac{1}{2^{m+1}} \sum_{j=1}^m \sum_{\varphi \in \Phi} \left\{ \mathbb{E}_{\varphi} L_j(\hat{\varphi}, \varphi) + \mathbb{E}_{\varphi^{[j]}} L_j(\hat{\varphi}, \varphi^{[j]}) \right\}, \quad (7)$$

where the last equality follows from the fact that in the sum we count every element of  $\varphi$  twice (so we must divide by 2). By the definition of  $L_j$  and the triangle inequality,

$$L_j(\hat{\varphi}, \varphi) + L_j(\hat{\varphi}, \varphi^{[j]}) \geq \frac{1}{A} g(d_j(\hat{\varphi}, \varphi) + d_j(\hat{\varphi}, \varphi^{[j]})) \geq \frac{1}{A} g(d_j(\varphi, \varphi^{[j]})) \geq \frac{g(\alpha_j)}{A}.$$

If we multiply and divide eq. (7) by  $\sum_{j=1}^m L_j(\hat{\varphi}, \varphi) + L_j(\hat{\varphi}, \varphi^{[j]})$ , we obtain, writing  $\mathcal{F}$  for the set of measurable functions from  $\mathcal{X} \rightarrow [0, 1]$ ,

$$\begin{aligned} \max_{\varphi \in \Phi} \mathbb{E}_{\varphi} L(\hat{\varphi}, \varphi) &\geq \left( \sum_{j=1}^m L_j(\hat{\varphi}, \varphi) + L_j(\hat{\varphi}, \varphi^{[j]}) \right) \frac{1}{2^{m+1}} \frac{\sum_{j=1}^m \sum_{\varphi \in \Phi} \left\{ \mathbb{E}_{\varphi} L_j(\hat{\varphi}, \varphi) + \mathbb{E}_{\varphi^{[j]}} L_j(\hat{\varphi}, \varphi^{[j]}) \right\}}{\sum_{j=1}^m L_j(\hat{\varphi}, \varphi) + L_j(\hat{\varphi}, \varphi^{[j]})} \\ &\stackrel{??}{\geq} \frac{\sum_{j=1}^m g(\alpha_j)}{2^{m+1} A} \sum_{\varphi \in \Phi} \inf_{f_1, f_2 \in \mathcal{F}: f_1 + f_2 = 1} \left\{ \mathbb{E}_{\varphi}(f_1) + \mathbb{E}_{\varphi^{[j]}}(f_2) \right\} \\ &= \frac{\sum_{j=1}^m g(\alpha_j)}{2^{m+1} A} \sum_{\varphi \in \Phi} \left( 1 - \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_{\varphi} f - \mathbb{E}_{\varphi^{[j]}} f \right\} \right) \\ &\stackrel{*}{\geq} \frac{1}{2A} \left\{ 1 - \max_{\varphi \sim \varphi'} \text{TV}(P_{\varphi}, P_{\varphi'}) \right\} \sum_{j=1}^m g(\alpha_j), \end{aligned}$$

**TODO:** explain  $\star$  (related to the alternative expression for TV as the divergence of  $|x - 1|/2$ ).  $\square$

**Example 4.10.** Let  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N_d(\vartheta, \Sigma) =: P_{\vartheta}$  for some  $\vartheta \in \mathbb{R}^d$ , where  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ . Fix  $c > 0$ , and for  $\varphi \in \{0, 1\}^d$  define  $\vartheta^{\varphi} \in \mathbb{R}^d$  by  $\vartheta_j^{\varphi} = c\sigma_j n^{-1/2} \mathbb{1}_{\varphi_j=1}$  for  $j \in [d]$ .

If we used the squared error loss  $L(\vartheta, \vartheta') = \|\vartheta - \vartheta'\|_2^2$ , then we have  $d_j(\vartheta, \vartheta') = |\vartheta_j - \vartheta_j'|$ , so  $\alpha_j = c\sigma_j n^{-1/2}$  and  $g(x) = x^2$ , and  $(x + y)^2 \leq 2(x^2 + y^2)$  implies  $A = 2$ .

Write  $\hat{\Theta}$  for the set of measurable functions from  $(\mathbb{R}^d)^{\times n}$  to  $\mathbb{R}^d$ , then Assouad's lemma tells us that, for any  $\hat{\vartheta}(X_1, \dots, X_n) = \hat{\vartheta} \in \hat{\Theta}$ ,

$$\begin{aligned} \sup_{\vartheta \in \mathbb{R}^d} \mathbb{E}_{\vartheta} \left( \left\| \hat{\vartheta} - \vartheta \right\|^2 \right) &\geq \max_{\varphi \in \Phi} \mathbb{E}_{\vartheta^{\varphi}} \left( \left\| \hat{\vartheta} - \vartheta^{\varphi} \right\|^2 \right) \\ &\geq \frac{1}{4} \left\{ 1 - \max_{\varphi \sim \varphi'} \text{TV}(P_{\vartheta^{\varphi}}, P_{\vartheta^{\varphi'}}) \right\} \sum_{j=1}^d \frac{c^2 \sigma_j^2}{n} \\ &\stackrel{*}{\geq} \frac{c^2}{4n} \left\{ 1 - \max_{\varphi \sim \varphi'} \sqrt{\frac{\text{KL}(P_{\vartheta^{\varphi}}^{\times n}, P_{\vartheta^{\varphi'}}^{\times n})}{2}} \right\} \sum_{j=1}^d \sigma_j^2 \\ &= \frac{c^2}{4n} \left( 1 - \frac{c}{2} \right) \sum_{j=1}^d \sigma_j^2, \end{aligned}$$

and taking the supremum over all  $c > 0$  on the right-hand side gives  $\frac{4}{27n} \sum_{j=1}^d \sigma_j^2$ . Here,  $\star$  follows from Pinsker's inequality (example sheet).

It can be shown that this is the optimal rate in terms of  $n$  and the  $\sigma_j$ , but again,  $4/27$  is not the optimal constant.

## 4.5 The data processing inequality

If  $(\mathcal{X}, \mathcal{A}, \mu)$  is a measure space and  $(\mathcal{Y}, \mathcal{B})$  a measurable space, and  $g: \mathcal{X} \rightarrow \mathcal{Y}$  is measurable, we denote the pushforward measure by  $\mu^g := \mu \circ g^{-1}$ .

**Lemma 4.11** (Data processing inequality). *Let  $(\mathcal{X}, \mathcal{A})$  and  $(\mathcal{Y}, \mathcal{B})$  be measurable spaces,  $P, Q$  probability measures on  $\mathcal{X}$ , and  $g: \mathcal{X} \rightarrow \mathcal{Y}$  measurable. Then for any  $f$ -divergence  $\text{Div}_f$  we have*

$$\text{Div}_f(P^g, Q^g) \leq \text{Div}_f(P, Q).$$

*Proof.* We decompose  $P = P_{\text{ac}} + P_{\text{sing}}$  w.r.t.  $Q$ . We will show that  $(P_{\text{ac}})^g \ll Q^g$  with Radon-Nikodym derivative  $\gamma$ , where

$$\gamma(y) := \mathbb{E}_Q \left( \frac{dP_{\text{ac}}}{dQ} \mid g(X) = y \right).$$

To see this, let  $B \in \mathcal{B}$ , then

$$\begin{aligned} (P_{\text{ac}})^g(B) &= P_{\text{ac}}(g^{-1}(B)) = \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(B)} \frac{dP_{\text{ac}}}{dQ} dQ = \mathbb{E}_Q \left[ \mathbb{E}_Q \left[ \mathbb{1}_{g^{-1}(B)}(X) \frac{dP_{\text{ac}}}{dQ}(X) \mid g(X) \right] \right] \\ &\stackrel{*}{=} \mathbb{E}_Q \left[ \mathbb{1}_{g^{-1}(B)}(X) \mathbb{E}_Q \left[ \frac{dP_{\text{ac}}}{dQ}(X) \mid g(X) \right] \right] \\ &= \int_{\mathcal{X}} \mathbb{1}_{g^{-1}(B)} \cdot (\gamma \circ g) dQ \stackrel{*}{=} \int_{\mathcal{Y}} \mathbb{1}_B \cdot \gamma dQ^g = \int_B \gamma dQ^g, \end{aligned}$$

where  $\star$  follows from the transformation theorem. This establishes the claim.

Now, writing  $(P_{\text{sing}})^g = ((P_{\text{sing}})^g)_{\text{ac}} + ((P_{\text{sing}})^g)_{\text{sing}}$  (the Lebesgue decomposition w.r.t.  $Q^g$ ), we have

$$P^g = (P_{\text{ac}})^g + (P_{\text{sing}})^g = (P_{\text{ac}})^g + ((P_{\text{sing}})^g)_{\text{ac}} + ((P_{\text{sing}})^g)_{\text{sing}},$$

and since  $(P_{\text{ac}})^g \ll Q^g$ , this gives the Lebesgue decomposition of  $P^g$  w.r.t.  $Q^g$ , namely  $(P^g)_{\text{ac}} = (P_{\text{ac}})^g$  and  $(P^g)_{\text{sing}} = ((P_{\text{sing}})^g)_{\text{sing}}$ . We now obtain, using  $f(x+y) \leq f(x) + M_f$ , that

$$\begin{aligned} \text{Div}_f(P^g, Q^g) &= \int_{\mathcal{Y}} f \left( \frac{d(P_{\text{ac}})^g}{dQ^g} + \frac{d((P_{\text{sing}})^g)_{\text{ac}}}{dQ^g} \right) dQ^g + ((P_{\text{sing}})^g)_{\text{sing}}(\mathcal{Y}) \cdot M_f \\ &\leq \int_{\mathcal{Y}} f \left( \frac{d(P_{\text{ac}})^g}{dQ^g} \right) dQ^g + (P_{\text{sing}})^g(\mathcal{Y}) \cdot M_f \\ &= \int_{\mathcal{Y}} f \circ \gamma dQ^g + (P_{\text{sing}})^g(\mathcal{Y}) \cdot M_f \\ &= \int_{\mathcal{X}} f \circ \gamma \circ g dQ + P_{\text{sing}}(\mathcal{X}) \cdot M_f \\ &= \mathbb{E}_Q \left[ f \left( \mathbb{E}_Q \left[ \frac{dP_{\text{ac}}}{dQ}(X) \mid g(X) \right] \right) \right] + P_{\text{sing}}(\mathcal{X}) \cdot M_f \\ &\leq \mathbb{E}_Q \left[ \mathbb{E}_Q \left\{ f \left( \frac{dP_{\text{ac}}}{dQ}(X) \mid g(X) \right) \right\} \right] + P_{\text{sing}}(\mathcal{X}) \cdot M_f \\ &= \int_{\mathcal{X}} f \left( \frac{dP_{\text{ac}}}{dQ} \right) dQ + P_{\text{sing}}(\mathcal{X}) \cdot M_f = \text{Div}_f(P, Q), \end{aligned}$$

where the last inequality follows from the conditional version of Jensen.  $\square$