# 1 Basic concepts

## 1.1 Parametric vs Nonparametric models

A statistical model postulates a family of possible data generating mechanisms. Examples include:

(i) Let $X_1, \ldots, X_n \sim^{\text{iid}} \Gamma(m, \theta)$ where $m$ is known and $\theta \in (0, \infty) := \Theta$;

(ii) Let $Y_i = \alpha + \beta x_i + \varepsilon_i$ for $i \in [n] := \{1, \ldots, n\}$, where $x_1, \ldots, x_n$ and $\varepsilon_1, \ldots, \varepsilon_n \sim^{\text{iid}} \mathcal{N}(0, \sigma^2)$. Here the unknown parameter is $\theta = (\alpha, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) := \Theta$.

If the parameter space $\Theta$ is finite-dimensional, we speak of a *parametric model*. When the model is correctly specified, i.e there exists $\theta_0 \in \Theta$ for which the data were generated from the distribution with parameter $\theta_0$, typically we can use the MLE $\hat{\theta}_n$ to estimate $\theta_0$, and expect $n^{1/2}(\hat{\theta}_n - \theta_0)$ to converge to a non-degenerate limiting distribution. On the other hand, when the model is misspecified, inferences may be very misleading.

Examples of nonparametric models include:

(i) Let $X_1, \ldots, X_n \sim^{\text{iid}} F$ for some unknown distribution function $F$;

(ii) Let $X_1, \ldots, X_n \sim^{\text{iid}} f$ for some density $f$ belonging to some unknown smoothness class;

(iii) Let $Y_i = m(x_i) + \varepsilon_i$ for $i \in [n]$, where $x_1, \ldots, x_n$ are known, $m$ belongs to some unknown smoothness class and $\varepsilon_1, \ldots, \varepsilon_n$ are iid with $\mathbb{E}(\varepsilon_1) = 0$, $\text{Var}(\varepsilon_1) = \sigma^2$.

Such infinite-dimensional models are much less vulnerable to model misspecification. Typically however, we will pay a price in terms of a slower rate of convergence.

## 1.2 Estimating an arbitrary distribution function

Let $\mathcal{F}$ denote the set of all distribution functions on $\mathbb{R}$. The *empirical distribution function* $\mathbb{F}_n$ of real-valued random variables $X_1, \ldots, X_n$ is defined by

$$\mathbb{F}_n(x) = \mathbb{F}_n(x, X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}.$$

**Theorem** (Glivenko-Cantelli Theorem)**.** *Let $X_1, \ldots, X_n \sim^{iid} F \in \mathcal{F}$ and let $\mathbb{F}_n$ denote the empirical distribution function of $X_1, \ldots, X_n$. Then*

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \xrightarrow{a.s} 0 \text{ as } n \to \infty.$$

*Proof.* Let $\varepsilon > 0$ and $k := \lceil \frac{1}{\varepsilon} \rceil$. Let $x_0 = -\infty$, $x_i = \inf\{x \in \mathbb{R} : F(x) \geq i/k\}$ for $i \in [k-1]$ and $x_k = \infty$. Writing $F(x-)$ for $\lim_{y \uparrow x} F(y)$, note that for $i \in [k]$

$$F(x_i-) - F(x_{i-1}) \leq \frac{i}{k} - \frac{i-1}{k} = \frac{1}{k} \leq \varepsilon.$$

Now define the event

$$\Omega_{n,\varepsilon} = \left\{ \max_{i \in [k]} \sup_{m \geq n} |\mathbb{F}_m(x_i) - F(x_i)| \leq \varepsilon \right\} \cap \left\{ \max_{i \in [k]} \sup_{m \geq n} |\mathbb{F}_m(x_i-) - F(x_i-)| \leq \varepsilon \right\}$$

Noting that both $\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq x\}$ and $\mathbb{F}_n(x-) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i < x\}$ are both sample averages of i.i.d random variables, we have by a union bound and the SLLN that

$$\begin{aligned}
&\mathbb{P}_F(\Omega_{n,\varepsilon}^c) \\
&\leq \sum_{i=1}^k \mathbb{P}_F \left( \sup_{m \geq n} |\mathbb{F}_m(x_i) - F(x_i)| > \varepsilon \right) + \sum_{i=1}^k \mathbb{P}_F \left( \sup_{m \geq n} |\mathbb{F}_m(x_i-) - F(x_i-)| > \varepsilon \right) \\
&\xrightarrow{a.s} 0.
\end{aligned}$$

Now let $x \in \mathbb{R}$ and find $i_* \in [k]$ such that $x \in [x_{i_*-1}, x_{i_*})$. Then for any $n_0 \in \mathbb{N}$ and $n \geq n_0$,

$$\begin{aligned}
\mathbb{F}_n(x) - F(x) &\leq \mathbb{F}_n(x_{i_*}-) - F(x_{i_*-1}) \\
&= \mathbb{F}_n(x_{i_*}-) - F(x_{i_*}-) + F(x_{i_*}-) - F(x_{i_*-1}) \\
&\leq \max_{i \in [k]} \sup_{m \geq n_0} |\mathbb{F}_m(x_i-) - F(x_i-)| + \varepsilon.
\end{aligned}$$

We also have

$$\begin{aligned}
F(x) - \mathbb{F}_n(x) &\leq F(x_{i_*}-) - \mathbb{F}_n(x_{i_*-1}) \\
&= F(x_{i_*}-) - F(x_{i_*-1}) + F(x_{i_*-1}) - \mathbb{F}_n(x_{i_*-1}) \\
&\leq \varepsilon + \max_{i \in [k]} \sup_{m \geq n_0} |\mathbb{F}_m(x_i) - F(x_i)|.
\end{aligned}$$

It follows that

$$\mathbb{P}_F \left( \sup_{n \geq n_0} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| > 2\varepsilon \right) \leq \mathbb{P}_F(\Omega_{n_0,\varepsilon}^c) \to 0 \text{ as } n_0 \to \infty.$$

Since $\varepsilon > 0$ was arbitrary, we conclude that

$$\begin{aligned}
\mathbb{P}_F \left( \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \to 0 \right) &= \mathbb{P}_F \left( \bigcap_{L=1}^\infty \bigcup_{n_0=1}^\infty \left\{ \sup_{n \geq n_0} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \leq \frac{1}{L} \right\} \right) \\
&= \lim_{L \to \infty} \lim_{n_0 \to \infty} \mathbb{P}_F \left( \sup_{n \geq n_0} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \leq \frac{1}{K} \right) \\
&= 1.
\end{aligned}$$

□

In fact, we can say much more:

**Theorem** (Dvoretzky-Kiefer-Wolfowitz Theorem)**.** *Under the conditions of the previous theorem, for every $\varepsilon > 0$,*

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

*Proof.* Not given. □

**Corollary** (Uniform Glivenko-Cantelli)**.** *Under the conditions of the Glivenko-Cantelli Theorem,*

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

*Proof.* By a union bound and the DKW inequality,

$$\begin{aligned}\sup_{F\in\mathcal{F}} \mathbb{P}_F\left(\sup_{m\geq n}\sup_{x\in\mathbb{R}} |\mathbb{F}_m(x) - F(x)| > \varepsilon\right) &\leq \sup_{F\in\mathcal{F}} \sum_{m\geq n} \mathbb{P}_F\left(\sup_{x\in\mathbb{R}} |\mathbb{F}_m(x) - F(x)| > \varepsilon\right) \\ &\leq 2\sum_{m\geq n} e^{-2m\varepsilon^2} \\ &\frac{2e^{-2n\varepsilon^2}}{1-e^{-2\varepsilon^2}} \to 0 \text{ as } n \to \infty.\end{aligned}$$

□

As another application of the DKW inequality, consider the problem of finding a confidence bound for $F$. Given $\alpha \in (0,1)$, set $\varepsilon_n = \varepsilon_n(\alpha) := \left\{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)\right\}^{1/2}$. Then, by the DKW Theorem,

$$\mathbb{P}_F\left(\max\{0, \mathbb{F}_n(x) - \varepsilon_n\} \leq F(x) \leq \min\{\mathbb{F}_n(x) + \varepsilon_n, 1\} \ \ \forall x \in \mathbb{R}\right) \geq 1 - \alpha.$$

In fact, let $U_1, \ldots, U_n \sim^{\text{iid}} \mathcal{U}[0,1]$ and let $\mathbb{G}_n$ denote their empirical distribution. Define the *quantile function* $F^{-1} : (0,1] \to (-\infty, \infty]$ by $F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}$ (i.e the *generalised inverse*). Since $F$ is increasing and right-continuous, we have $\{x \in \mathbb{R} : F(x) \geq p\} = [F^{-1}(p), \infty)$, so $\{U_i \leq F(x)\} = \{F^{-1}(U_i) \leq x\}$. Hence

$$\begin{aligned}\mathbb{G}_n(F(x)) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{U_i \leq F(x)\} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{F^{-1}(U_i) \leq x\} &=^d \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \leq x\} \\ &= \mathbb{F}_n(x).\end{aligned}$$

So

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

with equality if $F$ is continuous. Thus if $F$ is continuous, the distribution of $\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)|$ does not depend on $F$!

Other generalisations of the Glivenko-Cantelli Theorem include Uniform Laws of Large Numbes (ULLN). Let $X_1, X_2, \ldots, X_n$ be iid, taking values in some measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{G}$ denote a family of real-valued measurable functions on $\mathcal{X}$. We say $\mathcal{G}$ satisfies a *ULLN* if

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

**Remark.** $\mathcal{G}$ may be uncountable, so this random variable could be non measurable. But we can get around this by working with an outer probability $\mathbb{P}^*(A) = \inf\{\mathbb{P}(B) : B \in \mathcal{F},\ B \supseteq A\}$.

Thus, in the Glivenko-Cantelli Theorem we proved that $\{\mathbb{1}\{\cdot \leq x\} : x \in \mathbb{R}\}$ satisfies a ULLN. In general, proving a ULLN requires control of the 'size' of $\mathcal{G}$, which can be measured for instance through its metric entropy (see van de Geer, 2000).

## 1.3 Concentration Inequalities

**Definition.** A random variable $X$ is *sub-Gaussian with variance parameter* $\sigma^2$ if

$$\mathbb{E}(e^{tX}) \leq e^{t^2\sigma^2/2} \quad \forall t \in \mathbb{R}.$$

**Remark.** It can be shown (see Example Sheet) that any such random variable must satisfy $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) \leq \sigma^2$. Note also that we have equality in the definition if $X \sim \mathcal{N}(0, \sigma^2)$.

Here are some characterisations of sub-Gaussianity:

**Proposition.**

(a) If $X$ is sub-Gaussian with variance parameter $\sigma^2$, then

$$\max\{\mathbb{P}(X \geq x), \mathbb{P}(X \leq -x)\} \leq e^{-\frac{x^2}{2\sigma^2}} \qquad (*)$$

for every $x \geq 0$.

(b) If $X$ satisfies $(*)$, then for every $q \in \mathbb{N}$,

$$\mathbb{E}(X^{2q}) < 2q!(2\sigma^2)^q \leq q!(4\sigma^2)^q.$$

(c) Suppose that $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^{2q}) \leq q!C^{2q}$ for every $q \in \mathbb{N}$. Then $X$ is sub-Gaussian with variance parameter $4C^2$.

*Proof.*

(a) By Markov's inequality (i.e Chernoff's bound)

$$\mathbb{P}(X \geq x) \leq \inf_{t \geq 0} e^{-tX}\mathbb{E}(e^{tX}) \leq \inf_{t \geq 0} e^{-tx+t^2\sigma^2/2} = e^{-\frac{x^2}{2\sigma^2}}$$

for every $x \geq 0$, since the infimum is attained at $t = \frac{x}{\sigma^2}$. The bound for $\mathbb{P}(X \leq -x)$ is similar (since $-X$ is also $\sigma^2$-sub-Gaussian).

(b) We have

$$\begin{aligned}
\mathbb{E}(X^{2q}) = \int_0^\infty \mathbb{P}(X^{2q} \geq x)\mathrm{d}x &= 2q\int_0^\infty y^{2q-1}\mathbb{P}(|X| \geq y)\mathrm{d}x \\
&\leq 4q\int_0^\infty y^{2q-1}e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y \\
&= 2q(2\sigma^2)^q\int_0^\infty t^{q-1}e^{-t}\mathrm{d}t \qquad (t = \tfrac{y^2}{2\sigma^2}) \\
&= 2q!(2\sigma^2)^q
\end{aligned}$$

for $q \in \mathbb{N}$.

(c) Let $X'$ denote an independent copy of $X$. Then $X - X'$ has a symmetric distribution, so for every $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}(e^{t(X-X')}) &= \frac{1}{2}\left(\mathbb{E}(e^{t(X-X')}) + \mathbb{E}(e^{-t(X-X')})\right) \\
&= \mathbb{E}(\cosh(t(X-X'))) \\
&= \sum_{q \geq 0} \frac{t^{2q}\mathbb{E}[(X-X')^{2q}]}{(2q)!}
\end{aligned}$$

where the final step follows from Fubini's theorem, since the terms in the Taylor series of $x \mapsto \cosh(x)$ are non-negative. Moreover, $\mathbb{E}(e^{-tX}) \geq 1$ by Jensen's inequality. Hence by applying the inequality $(a+b)^r \leq \max(1, 2^{r-1})(a^r + b^r)$ for every $a, b, r \geq 0$ we have for every $t \in \mathbb{R}$

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

□

**Theorem** (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with $X_i$ having sub-Gaussian parameter $\sigma_i^2$ for $i \in [n]$. Then $\bar{X} := n^{-1}\sum_{i=1}^n X_i$ is sub-Gaussian with variance parameter $\frac{\bar{\sigma}^2}{n}$, where $\bar{\sigma}^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$. In particular,*

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

*for every $x \geq 0$.*

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

$$\mathbb{E}(e^{t\bar{X}}) = \prod_{i=1}^n \mathbb{E}(e^{\frac{t}{n}X_i}) \leq \prod_{i=1}^n e^{\frac{t^2\sigma_i^2}{2n^2}} = e^{\frac{t^2\bar{\sigma}^2}{n}}.$$

The second claim then follows from part (a) of the previous proposition. □

**Remark.** It is often convenient to state the conclusion of Hoeffding's inequality as an upper bound on the $(1-\delta)$th qualtile of the distribution of $\bar{X}$: for every $\delta \in [0, 1]$

$$\mathbb{P}\left(\bar{X} \geq \frac{2^{1/2}\bar{\sigma}\log^{1/2}(1/\delta)}{n^{1/2}}\right) \leq \delta.$$

**Remark.** Since we have the same bound on $\mathbb{P}(\bar{X} \leq -x)$ we have that

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

for every $x \geq 0$. Often, Hoeffding's inequality is stated in this (weaker) way.

**Lemma** (Hoeffding's Lemma)**.** *Let $X$ be a mean 0 bounded random variable taking values in $[a, b]$. Then $X$ is sub-Gaussian with variance parameter $(b-a)^2/4$.*

*Proof.* Example Sheet 1. □

**Corollary.** *Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}(X_i) = \mu_i$ and $X_i$ taking values in $[a_i, b_i]$ for $i \in [n]$. Writing $\bar{X} := n^{-1}\sum_{i=1}^n X_i$ and $\bar{\mu} = n^{-1}\sum_{i=1}^n \mu_i$, we have*

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

*for every $x \geq 0$.*

*Proof.* By Hoeffding's Lemma and Hoeffding's inequality, $\bar{X} - \bar{\mu}$ is sub-Gaussian with variance parameter $\sum_{i=1}^n \frac{(b_i-a_i)^2}{4n^2}$. □

The bound in the previous corollary may be loose when the variance of $X_i$ is small by comparison with $(b_i - a_i)^2$. This happens for instance when $X_i \sim$ Bernoulli$(p_i)$ with $p_i$ close to 0 or 1. In such circumstances, Bennett's inequality (which applies when the random variables have a bounded right tail) or Bernstein's inequality (which applied under a weaker integrability condition) may be preferred.

**Theorem** (Bennett's inequality)**.** *Let $X_1, \dots, X_n$ be independent with $\mathbb{E}X_i = 0$ and $X_i \leq b$ for some $b > 0$ and all $i \in [n]$. Write $S = \sum_{i=1}^n X_i$ and assume that $\nu = n^{-1}\sum_{i=1}^n Var(X_i) \in (0, \infty)$, and define $\varphi : \mathbb{R} \to \mathbb{R}$ by $\varphi(u) = e^u - 1 - u$. Then, for every $t \geq 0$*

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

*Hence, defining $h : [0, \infty) \to [0, \infty)$ by $h(u) = (1+u)\log(1+u) - u$, we have for every $x \geq 0$ that*

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

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

$$g(u) = \begin{cases} \frac{\varphi(u)}{u^2} & \text{if } u \neq 0 \\ \frac{1}{2} & \text{if } u = 0 \end{cases}.$$

We claim that $g$ is increasing. To see this, define $G : \mathbb{R} \to \mathbb{R}$ by

$$G(u) = \int_0^1 e^{su}\mathrm{d}s = \begin{cases} \frac{e^u - 1}{u} & \text{if } u \neq 0 \\ 1 & \text{if } u = 0 \end{cases}.$$

Then $G$ is convex, because if $\lambda \in (0,1)$ and $u, v \in \mathbb{R}$ we have

$$\begin{aligned} G(\lambda u + (1-\lambda)v) &= \int_0^1 e^{s(\lambda u + (1-\lambda)v)}\mathrm{d}s \\ &\leq \lambda \int_0^1 e^{su}\mathrm{d}s + (1-\lambda)\int_0^1 e^{sv}\mathrm{d}s \\ &= \lambda G(u) + (1-\lambda)G(v) \end{aligned}$$

by convexity of $u \mapsto e^{su}$. Hence

$$g(u) = \begin{cases} \frac{G(u) - G(0)}{u} & \text{if } u \neq 0 \\ G'(0) & \text{if } u = 0 \end{cases}$$

is increasing, as required.

It follows that for every $t \geq 0$,

$$e^{tX_i} - 1 - tX_i \leq X_i^2 \frac{e^{bt} - 1 - bt}{b^2} = X_i^2 \frac{\varphi(bt)}{b^2}.$$

Hence for every $t \geq 0$

$$\begin{aligned} \mathbb{E}(e^{tS}) = \sum_{i=1}^n \log \mathbb{E}(e^{tX_i}) &\leq \sum_{i=1}^n \log\left(1 + \frac{\mathbb{E}(X_i^2)\varphi(bt)}{b^2}\right) \\ &\leq n \log\left(1 + \frac{\nu}{b^2}\varphi(bt)\right) \qquad \text{(Jensen)} \\ &\leq \frac{n\nu}{b^2}\varphi(bt) \end{aligned}$$

where the final inequality uses the fact $\log(1+x) \leq x$ for $x > -1$.

Hence, by a Chernoff bound

$$\mathbb{P}\left(\frac{S}{n} \geq x\right) = \mathbb{P}(S \geq nx) \leq \inf_{t \geq 0} e^{-ntx + \frac{n\nu}{b^2}\varphi(bt)}.$$

But if $f(t) = -tx + \frac{\nu}{b^2}\varphi(bt)$ then $f'(t) = -x + \frac{\nu}{b}(e^{bt} - 1)$, so its infimum is attained at $t^* = \frac{1}{b}\log\left(1 + \frac{bx}{\nu}\right) \geq 0$, and

$$\begin{aligned}
\mathbb{P}\left(\frac{S}{n} \geq x\right) &\leq \exp\left[-\frac{nx}{b}\log\left(1+\frac{bx}{\nu}\right) + \frac{n\nu}{b^2}\left\{\left(1+\frac{bx}{\nu}\right) - 1 - \log\left(1+\frac{bx}{\nu}\right)\right\}\right] \\
&\leq \exp\left[-\frac{n\nu}{b^2}\left\{\left(1+\frac{bx}{\nu}\right)\log\left(1+\frac{bx}{\nu}\right) - \frac{bx}{\nu}\right\}\right] \\
&= \exp\left\{-\frac{n\nu}{b^2}h\left(\frac{bx}{\nu}\right)\right\}
\end{aligned}$$

for every $x \geq 0$, as required. □

Before we can state Bernstein's inequality, we first introduce the notion of sub-Gamma random variables.

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

$$\log \mathbb{E}(e^{tX}) \leq \frac{\sigma^2 t^2}{2(1-ct)}$$

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

**Proposition.**

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

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

for $x \geq 0$.

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

*Proof.*

(a) By a Chernoff bound, for all $x \geq 0$ we have

$$\mathbb{P}(X \geq x) \leq \inf_{t \in [0, 1/c)} e^{-tx + \frac{\sigma^2 t^2}{2(1-ct)}} \leq e^{-\frac{x^2}{2(\sigma^2 + cx)}}$$

where the final inequality follows by setting $t = \frac{x}{\sigma^2 + cx} \in [0, 1/c)$.

(b) Recall from the proof of the previous theorem (Bennett's inequality) that $\varphi(u) \leq u^2/2$ for $u \leq 0$. It follows that

$$\varphi(u) = e^u - 1 - u \leq \frac{u^2}{2} + \sum_{q \geq 3} \frac{u_+^q}{q!} \text{ for all } u \in \mathbb{R}.$$

Hence, for $t \in [0, 1/c)$, by Fubini's theorem

$$\begin{aligned} \log \mathbb{E}(e^{tX}) \leq \mathbb{E}(e^{tX}) - 1 &= \mathbb{E}\varphi(tX) && \text{(since } \mathbb{E}X = 0) \\ &= \frac{t^2}{2}\mathbb{E}(X^2) + \sum_{q \geq 3} \frac{t^q \mathbb{E}(X_+^q)}{q!} \\ &\leq \frac{\sigma^2}{2} \sum_{q \geq 2} t^q c^{q-2} \\ &= \frac{\sigma^2 t^2}{2(1-ct)} \end{aligned}$$

as required.

□

Finally, we are in a position to state Bernstein's inequality.

**Theorem** (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean 0 random variables with $\frac{1}{n}\sum_{i=1}^n Var(X_i) \leq \sigma^2$ and $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \leq \frac{q!}{2}\sigma^2 c^{q-2}$ for some $\sigma, c > 0$. 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, writing $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, we have for every $x \geq 0$ that*

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

*Proof.* As in the proof of (b) of the previous proposition, for every $t \in [0, 1/c)$,

$$\begin{aligned} \log \mathbb{E}(e^{tS}) &\leq \sum_{i=1}^n \left[\frac{t^2}{2}\mathbb{E}(X_i^2) + \sum_{q \geq 3} \frac{t^q}{q!}\mathbb{E}[(X_i)_+^q]\right] \\ &\leq \frac{n\sigma^2}{2} \sum_{q \geq 2} t^q c^{q-2} && \text{(Fubini)} \\ &= \frac{n\sigma^2}{2(1-ct)} \end{aligned}$$

as required. By part (a) of the previous proposition, for every $x \geq 0$ we have

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

as required. □

**Remark.** If in addition we have $X_i \leq b$ for some $b > 0$ and all $i \in [n]$, then for every integer $q \geq 3$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i)_+^q] \leq \frac{b^{q-2}}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] \leq \sigma^2 b^{q-2} \leq \frac{q!}{2}\sigma^2\left(\frac{b}{3}\right)^{q-3}.$$

Thus we can take $c = b/3$. In that case, Bernstein's inequality can be deduced from Bennett's inequality, since

$$h(u) \geq \frac{u^2}{2(1+u/3)}$$

for all $u > 0$; see Example Sheet.

# 2 Kernel density estimation

## 2.1 Introduction

Let $X_1, \ldots, X_n$ be independent, real-valued random variables with density $f$. The oldest, and most commonly used non-parametric density estimator is the *histogram*, usually formed by dividing the real lines into equal-sized intervals, known as *bins*.

If $I_x$ denotes the bin for which $x \in \mathbb{R}$ belongs, and $b > 0$ denotes the *binwidth*, then the histogram density estimator of $f$ is $\hat{f}_n^H = \hat{f}_{n,b}^H$, where

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

Drawbacks of histograms include:

- Difficulties of choosing the binwidth and the positioning of the bin edges;
- Suboptimal theoretic performance;
- Difficulties of graphical display in the multivariate case.

## 2.2 The univariate kernel density estimator

A *kernel* is a Borel-measurable function $k : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{-\infty}^{\infty} k(x)\mathrm{d}x = 1$. A *univariate kernel density estimator* of $f$ is of the form $\hat{f}_n = \hat{f}_{n,h,k}$, where

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

where $k$ is a kernel and $h > 0$ is called the *bandwidth*. It is convenient to define the *scaled kernel* $k_h(\cdot) := h^{-1}k(\frac{\cdot}{h})$, so that

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

Typically in practice, $k$ is chosen to be non-negative (though see later discussion), which ensures that $k$ (and hence $\hat{f}_n$) is a density; often $k$ is chosen to be symmetric about 0.

The main intuition is that, perhaps surprisingly, the choice of kernel is much less important than the choice of bandwidth.

## 2.3 Mean squared error of kernel estimators

If we think of $\hat{f}_n(x)$ as a point estimate of $f(x)$, then it is natural to try to choose $h$ and $k$ to minimise the *mean squared error* (MSE), defined by

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

This is often preferred to alternatives, such as the mean absolute error, due to its appealing decomposition into variance and squared bias terms:

$$\begin{aligned}\mathrm{MSE}(\hat{f}_n) &= \mathbb{E}[(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)])^2] + [\mathbb{E}(\hat{f}_n(x)) - f(x)]^2 \\ &= \mathrm{Var}(\hat{f}_n(x)) + \mathrm{Bias}^2\hat{\mathrm{f}}_{\mathrm{n}}(\mathrm{x}).\end{aligned}$$

The *convolution* of Borel-measurable functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ is $g_1 * g_2$, where

$$(g_1 * g_2)(x) = \int_{\mathbb{R}} g_1(x - z)g_2(z)\mathrm{d}z,$$

whenever this integral exists. We can now compute

$$\begin{aligned}\mathrm{Bias}\hat{f}_n(x) = \mathbb{E}[k_h(x - X_1)] - f(x) &= \int_{-\infty}^{\infty} k_h(x - z)f(z)\mathrm{d}z - f(x) \\ &= (k_h * f)(x) - f(x).\end{aligned}$$

Now, for $i \in [n]$, let $\xi_i = \xi_i(x) = k_h(x - X_i)$. Then $\xi_1, \ldots, \xi_n$ are iid, so

$$\begin{aligned}\mathrm{Var}\hat{f}_n(x) = \frac{1}{n}\mathrm{Var}(\xi_1) &= \frac{1}{n}[\mathbb{E}[\xi_1^2] - \mathbb{E}^2[\xi_1]] \\ &= \frac{1}{n}\left[(k_h^2 * f)(x) - (k_h * f)^2(x)\right].\end{aligned}$$

Hence

$$\mathrm{MSE}[\hat{f}_n(x)] = \frac{1}{n}\left[(k_h^2 * f)(x) - (k_h * f)^2(x)\right] + \left[(k_h * f)(x) - f(x)\right]^2.$$

Often however, we prefer to choose $h$ and $k$ based on how well $\hat{f}_n$ estimates $f$ as a function, as measured by the *mean integrated squared error* (MISE), defined by

$$\mathrm{MISE}(\hat{f}_n) = \mathbb{E}\left[\int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 \mathrm{d}x\right].$$

Since the integrand is non-negative, we have by Fubini's theorem that

$$\begin{aligned}\mathrm{MISE}(\hat{f}_n) &= \int_{\mathbb{R}} \mathrm{MSE}(\hat{f}_n(x))\mathrm{d}x \\ &= \frac{1}{n}\int_{\mathbb{R}} \frac{1}{n}\left[(k_h^2 * f)(x) - (k_h * f)^2(x)\right]\mathrm{d}x + \int_{\mathbb{R}} \left[(k_h * f)(x) - f(x)\right]^2 \mathrm{d}x.\end{aligned}$$

Although this expression is exact, it depends on the bandwidth in a complicated way. So we therefore seek finite-sample bounds to clarify this dependence.

## 2.4 Bounds on the variance and bias

For a kernel $k$, we write $R(k) := \int_{\mathbb{R}} k^2(u)\mathrm{d}u$.

**Proposition.** Let $\hat{f}_n$ denote a KDE with bandwidth $h$ and kernel $k$ constructed from $X_1, \ldots, X_n \sim^{\mathrm{iid}} f$. Then for every $x \in \mathbb{R}$,

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

*Proof.* We have

$$\begin{aligned}\mathrm{Var}\hat{f}_n(x) &= \frac{1}{n}\mathrm{Var}(\xi_1) \\ &\leq \frac{1}{n}\mathbb{E}(\xi_1^2) \\ &\leq \frac{1}{nh^2}\int_{\mathbb{R}} k^2\left(\frac{x-z}{h}\right) f(z)\mathrm{d}z \\ &= \frac{1}{nh}\int_{\mathbb{R}} k^2(u) f(x-uh)\mathrm{d}u \qquad (u = \tfrac{x-z}{h}) \\ &\leq \frac{1}{nh} R(k)\|f\|_\infty.\end{aligned}$$

□

To study the bias, we will need conditions on both $f$ and $k$.

**Definition.** Let $I$ be an interval, let $\beta, L > 0$ and let $m := \lceil \beta \rceil - 1$. The *Hölder class* $\mathcal{H}(\beta, L)$ on $I$ is the set of $m$-times differentiable functions $f : I \to \mathbb{R}$ satisfying

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

for all $x, x' \in I$. Where $I$ is unspecified we will take $I = \mathbb{R}$.

**Example.** For $\alpha \in (0,1)$, the function $x \mapsto x^\alpha$ belongs to $\mathcal{H}(\beta, 1)$ on $[0,1]$ for every $\beta \in (0, \alpha]$. Indeed, wlog suppose $0 \leq x' < x \leq 1$, and write $x' = \varepsilon x$ for some $\varepsilon \in (0,1)$. Then

$$\frac{x^\alpha - (x')^\alpha}{x - x'} = \frac{x^\alpha(1-\varepsilon^\alpha)}{x^\beta(1-\varepsilon)^\beta} \leq \frac{1-\varepsilon^\alpha}{(1-\varepsilon)^\beta} \leq (1-\beta) \leq 1.$$

On the other hand, by considering $x' = 0$, we see that this function does not belong to $\mathcal{H}(\beta, L)$ for any $\beta > 0$ and any $L > 0$.

The densities in $\mathcal{H}(\beta, L)$ are denoted $\mathcal{F}(\beta, L) := \{f \in \mathcal{H}(\beta, L) : \int_{\mathbb{R}} f = 1\}$.

**Definition.** For $\ell \in \mathbb{N}$, we say a kernel $K$ is *of order* $\ell$ if $\int_{\mathbb{R}} u^j k(u) \mathrm{d}u$ for all $j \in [\ell - 1]$.

Most kernels used in practice are of order 2. Kernels of order $\ell \geq 3$ cannot be non-negative, because $\int_{\mathbb{R}} u^2 k(u) \mathrm{d}u = 0$. This means that the resulting KDE is not necessarily a density.

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

$$\mu_\beta(k) := \int_{\mathbb{R}} |u|^\beta |k(u)| \mathrm{d}u < \infty.$$

If $\hat{f}_n$ denotes the KDE with kernel $K$ and bandwidth $h$ based on $X_1, \ldots, X_n \sim f$, then

$$|\mathrm{Bias} \hat{f}_n(x)| \leq \frac{L}{(\ell-1)!} \mu_\beta(k) h^\beta$$

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

*Proof.* We have

$$\begin{aligned} \mathrm{Bias} \hat{f}_n(x) &= \frac{1}{h} \int_{\mathbb{R}} k\left(\frac{x-z}{h}\right) f(z) \mathrm{d}z - f(x) \\ &= \int_{\mathbb{R}} k(u) \left\{ f(x - uh) - f(x) \right\} \mathrm{d}u. \qquad (u = (x-z)/h) \end{aligned}$$

By a Taylor expansion with the mean value form of the remainder, with $m := \lceil \beta \rceil - 1$, there exists $\tau \in [0,1]$ such that

$$f(x-uh) - f(x) = \sum_{r=1}^{m-1} \frac{(-uh)^r}{r!} f^{(r)}(x) + \frac{(-uh)^m}{m!} f^{(m)}(x - \tau uh).$$

Since $k$ is a kernel of order $\ell = m + 1$, we find that

$$\begin{aligned} \mathrm{Bias} \hat{f}_n(x) &= \frac{(-h)^m}{m!} \int_{\mathbb{R}} u^m k(u) f^{(m)}(x - \tau uh) \mathrm{d}u \\ &= \frac{(-h)^m}{m!} \int_{\mathbb{R}} u^m k(u) \left\{ f^{(m)}(x - \tau uh) - f^{(m)}(x) \right\} \mathrm{d}u. \end{aligned}$$

We conclude that

$$|\text{Bias}\hat{f}_n(x)| \leq \frac{Lh^m}{m!}\int_{\mathbb{R}} |u|^m |k(u)||\tau uh|^{\beta-m}\mathrm{d}u \leq \frac{L}{m!}\mu_\beta(k)h^\beta.$$

□

Observe that our upper bound on $\text{Var}\hat{f}_n(x)$ is decreasing in $h$, wheras the upper bound on the bias is increasing in $h$. This is a prototypical example of a bias-variance trade-off.

**Theorem.** *Assume that $k$ is a kernel of order $\ell := \lceil \beta \rceil$ satisfying $\max\{R(k), \mu_\beta(k)\} < \infty$. Then there exists $C = C(\beta, k) > 0$ and $h_n^* = h_n^*(\beta, L, k) > 0$ such that the KDE $\hat{f}_{n,h_n^*,k}$ based on iid observations $X_1, \ldots, X_n$ satisfies*

$$\sup_{x\in\mathbb{R}} \sup_{f\in\mathcal{F}(\beta,L)} \text{MSE}_f\{\hat{f}_{n,h_n^*,k}(x)\} \leq CL^{\frac{2}{\beta+1}} n^{-\frac{2\beta}{2\beta+1}}.$$

*Proof.* We first claim that there exists $C' = C'(\beta) > 0$ such that

$$\sup_{f\in\mathcal{F}(\beta,L)} \|f\|_\infty \leq C'\beta^{\frac{1}{\beta+1}}.$$

To see this, let $k^*$ denote a bounded kernel of order $\ell := \lceil \beta \rceil$ satisfying $\mu_\beta(k^*) < \infty$. Then by the triangle inequality and the previous proposition, we have for every $x \in \mathbb{R}$ and $f \in \mathcal{F}(\beta, L)$ that

$$\begin{aligned} f(x) &\leq \inf_{h>0} \int_{\mathbb{R}} |k^*(x-z)|f(z)\mathrm{d}z + \left| \int_{\mathbb{R}} k_h^*(x-z)f(z)\mathrm{d}z - f(x) \right| \\ &\leq \inf_{h>0} \left\{ \frac{\|k^*\|_\infty}{h} + \frac{L}{(\ell-1)!}\mu_\beta(k^*)h^\beta \right\} \\ &\leq C'L^{\frac{1}{\beta+1}} \end{aligned}$$

which establishes our first claim. Moreover, the RHS of the upper bound

$$\text{MSE}\{\hat{f}_n(x)\} \leq \frac{C'L^{\frac{1}{\beta+1}}}{nh}R(k) + \frac{L^2}{[(\ell-1)!]^2}\mu_\beta^2(k)h^{2\beta}$$

is minimised at

$$h_n^* = h_n^*(\beta, L, k) := \left( \frac{C'R(k)[(\ell-1)!]^2}{2\beta\mu_\beta^2(k)} \right) L^{-\frac{1}{\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

We conclude that

$$\text{MSE}(\hat{f}_{n,h_n^*,k}(x)) \leq \left( \frac{(C')^{2\beta}R(k)^{2\beta}\mu_\beta^2(k)(2\beta+1)^{2\beta+1}}{[(\ell=1)!]^2(2\beta)^{2\beta}} \right) L^{\frac{2}{\beta+1}} n^{-\frac{2\beta}{2\beta+1}}.$$

□

## 2.5 Bounds on the Integrated Variance and Squared Bias

**Proposition.** Let $\hat{f}_n$ denote the KDE with kernel $k$ and bandwidth $h$ based on iid observations $X_1, \ldots, X_n$ having an arbitrary distribution $P$ on $\mathbb{R}$. Then

$$\int_{\mathbb{R}} \mathrm{Var}(\hat{f}_n(x))\mathrm{d}x \leq \frac{R(k)}{nh}.$$

*Proof.* By the proof of the pointwise analologue and Fubini's theorem,

$$\begin{aligned}
\int_{\mathbb{R}} \mathrm{Var}\hat{f}_n(x)\mathrm{d}x &\leq \frac{1}{nh^2} \int_{\mathbb{R}} \int_{\mathbb{R}} k^2 \left( \frac{x-z}{h} \right) \mathrm{d}P(z)\mathrm{d}x \\
&= \frac{1}{nh^2} \int_{\mathbb{R}} \int_{\mathbb{R}} k^2 \left( \frac{x-z}{h} \right) \mathrm{d}x\mathrm{d}P(z) \\
&\leq \frac{R(k)}{nh}.
\end{aligned}$$

□

**Definition.** Fix $\beta, L > 0$ and let $m := \lceil \beta \rceil - 1$. The Nikol'ski class $\mathcal{N}(\beta, L)$ is the set of $(m-1)$-times differentiable $f : \mathbb{R} \to \mathbb{R}$ for which $f^{(m-1)}$ is locally absolutely continuous, with weak derivative $f^{(m)}$ satisfying

$$\left[\int_{\mathbb{R}} (f^{(m)}(x+t) - f^{(m)}(x))^2 \mathrm{d}x\right]^{1/2} \leq L|t|^{\beta - m}$$

for $t \in \mathbb{R}$. The densities in $\mathcal{N}(\beta, L)$ are denoted as $\mathcal{F}_{\mathcal{N}}(\beta, L) = \{f \in \mathcal{N}(\beta, L) : f \geq 0,\ \int_{\mathbb{R}} f = 1\}$.

**Theorem** (Generalised Minkowski Inequality)**.** *For $f : \mathbb{R}^2 \to \mathbb{R}$ measurable*

$$\left[\int_{\mathbb{R}} \left\{\int_{\mathbb{R}} f(x,y)\mathrm{d}y\right\}^2 \mathrm{d}x\right]^{1/2} \leq \int_{\mathbb{R}} \left\{\int_{\mathbb{R}} f^2(x,y)\mathrm{d}x\right\}^{1/2} \mathrm{d}y.$$

*Proof.* Not given (see Part II Analysis of Functions). □

**Proposition.** Fix $\beta, L > 0$ and let $k$ denote a kernel of order $\ell = \lceil \beta \rceil$. Then the kernel density estimator $\hat{f}_n$ with kernel $k$ and bandwidth $h > 0$ based on $X_1, \ldots, X_n \sim f \in \mathcal{F}_{\mathcal{N}}(\beta, L)$ satisfies

$$\int_{\mathbb{R}} \mathrm{Bias}^2 \hat{f}_n(x)\mathrm{d}x \leq \frac{L^2}{[(\ell-1)!]^2} \mu_\beta^2(k) h^{2\beta}.$$

*Proof.* First consider the case $m := \ell - 1 \geq 1$. Since $f^{(m-1)}$ is absolutely continuous on the line segment joining $x - uh$ and $x$, we have by a Taylor expansion with the integral form of remainder that for every $x, u \in \mathbb{R}$ and $h > 0$ we have

$$f(x-uh) - f(x) = \sum_{r=1}^{m-1} \frac{(-uh)^r}{r!} f^{(r)}(x) + \frac{(-uh)^m}{(m-1)!} \int_0^1 (1-\tau)^{m-1} f^{(m)}(x - \tau uh)\mathrm{d}\tau.$$

Since the kernel is of order $\ell = m + 1$, we have from the proof of the pointwise bound that

$$\begin{aligned}
\mathrm{Bias}\hat{f}_n(x) &= \int_{\mathbb{R}} k(u) \frac{(-uh)^m}{(m-1)!} \int_0^1 (1-\tau)^{m-1} f^{(m)}(x - \tau uh)\mathrm{d}\tau\mathrm{d}u \\
&= \int_{\mathbb{R}} k(u) \frac{(-uh)^m}{(m-1)!} \int_0^1 (1-\tau)^{m-1} (f^{(m)}(x - \tau uh) - f^{(m)}(x))\mathrm{d}\tau\mathrm{d}u.
\end{aligned}$$

Hence by two applications of the Generalised Minkowski Inequality

$$\begin{aligned}
&\int_{\mathbb{R}} \mathrm{Bias}^2 \hat{f}_n(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} k(u) \frac{(-uh)^m}{(m-1)!} \int_0^1 (1-\tau)^{m-1} (f^{(m)}(x-\tau uh) - f^{(m)}(x)) \mathrm{d}\tau \mathrm{d}u \right\}^2 \mathrm{d}x \\
&\leq \left( \int_{\mathbb{R}} |k(u)| \frac{|uh|^m}{(m-1)!} \left[ \int_{\mathbb{R}} \left\{ \int_0^1 (1-\tau)^{m-1} (f^{(m)}(x-\tau uh) - f^{(m)}(x)) \mathrm{d}\tau \right\}^2 \mathrm{d}x \right]^{1/2} \mathrm{d}u \right)^2 \\
&\leq \left[ \int_{\mathbb{R}} |k(u)| \frac{|uh|^m}{(m-1)!} \int_0^1 (1-\tau)^{m-1} \left\{ \int_{\mathbb{R}} (f^{(m)}(x-\tau uh) - f^{(m)}(x))^2 \mathrm{d}x \right\}^{1/2} \mathrm{d}\tau \mathrm{d}u \right]^2 \\
&\leq \left[ \int_{\mathbb{R}} |k(u)| \frac{|uh|^m}{(m-1)!} \int_0^1 (1-\tau)^{m-1} L |uh|^{\beta-m} \mathrm{d}\tau \mathrm{d}u \right]^2 \\
&= \frac{L^2}{(m!)^2} \mu_\beta^2(k) h^{2\beta}.
\end{aligned}$$

When $m = 0$ we can proceed directly via one application of the Generalised Minkowski Inequality:

$$\begin{aligned}
\int_{\mathbb{R}} \mathrm{Bias}^2 \hat{f}_n(x)\mathrm{d}x &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} k(u)(f(x-uh) - f(x))\mathrm{d}u \right\}^2 \mathrm{d}x \\
&\leq \left[ \int_{\mathbb{R}} |k(u)| \left\{ \int_{\mathbb{R}} (f(x-uh) - f(x))^2 \mathrm{d}x \right\}^{1/2} \mathrm{d}u \right]^2 \\
&\leq L^2 \mu_\beta^2(k) h^{2\beta}.
\end{aligned}$$

□

Putting our results together, we have the following:

**Theorem.** *Fix $\beta, L > 0$ and let $k$ be a kernel of order $\ell = \lceil \beta \rceil$. Then the KDE $\hat{f}_n$ with kernel $k$ and bandwidth $h > 0$ based on $X_1, \ldots, X_n \sim^{iid} f \in \mathcal{F}_{\mathcal{N}}(\beta, L)$ satisfies*

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

*In particular, if $\max\{R(k), \mu_\beta(k)\} < \infty$, then there exist $h_n^{**} = h_n^{**}(\beta, L, k) > 0$ and $C = C(\beta, K) > 0$ such that*

$$\mathrm{MISE}(\hat{f}_{n,h_n^{**},k}) \leq C L^{\frac{2}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}}.$$

*Proof.* The first part follows from previous propositions. For the second part, the upper bound on the MISE is minimised by setting

$$h_n^{**} := \left( \frac{R(k)[(\ell-1)!]^2}{(2\beta)\mu_\beta^2(k)} \right)^{\frac{1}{2\beta+1}} L^{-\frac{2}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

This gives

$$\mathrm{MISE}(\hat{f}_{n,h_n^{**},k}) \leq \left( \frac{R(k)^{2\beta}\mu_\beta^2(k)(2\beta+1)^{2\beta+1}}{[(\ell-1)!]^2(2\beta)^{2\beta}} \right)^{\frac{1}{2\beta+1}} L^{\frac{2}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}}.$$

□

## 2.6 Bandwidth Selection

The choice of $h_n^{**}$ in the previous theorem is not completely practical because typically we do not know $\beta$ or $L$. Here we present two bandwidth selection algorithms, one based on least squares cross-validation, and the other based on Lepski's method.

### 2.6.1 Least squares cross-validation

Recall that provided all the terms are finite

$$\mathrm{MISE}(\hat{f}_n) = \mathbb{E}\int_{\mathbb{R}} \hat{f}_n^2(x)\mathrm{d}x - 2\mathbb{E}\int_{\mathbb{R}} \hat{f}_n(x)f(x)\mathrm{d}x + \int_{\mathbb{R}} f^2(x)\mathrm{d}x.$$

Since the last term does not depend on $h$, it suffices to minimise the sum of the first two terms on the RHS. This depends on the unknown $f$, but an unbiased estimator for $n \geq 2$ is given by

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

where $\hat{f}_{n,-i}(x) = \hat{f}_{n,h,k,-i}(x) := \frac{1}{n-1}\sum_{j\neq i} k_h(x - x_j)$. The least squares cross-validation bandwidth selector minimises $\mathrm{LSCV}(h)$.

## 2.7 Lepski's method

Lepski's method is a very general approach to choosing tuning parameters in an adaptive way. We present here a version of the method, taylored to the problem of estimating a density $f$ at the point $x \in \mathbb{R}$.

Recall that if $f \in \mathcal{F}(\beta, L)$ where $\beta$ and $L$ are known, then the optimal bandwidth $h_n^*$ is of the form $C_{\beta,L,k} n^{-\frac{1}{2\beta+1}}$. Suppose now that $L$ is unknown and we only have an upper bound $\ell$ on $\beta$. Lepski's method chooses a bandwidth $\hat{h}$ adaptively from a pre-defined geometric sequence $\mathcal{H}_n$ of candidates. These extend from order $n^{-1}$ up to order $n^{-\frac{1}{2\ell+1}}$, up to logarithmic factors, so contain the optimal bandwidth $h_n^*$ when $n$ is large. The procedure constructs confidence intervals $\hat{I}_{n,h}$ for $\mathbb{E}\hat{f}_{n,h,k}(x)$ for $h \in \mathcal{H}_n$, whose widths are data-drived proxies for the standard deviation of $\hat{f}_{n,h,k}(x)$. The required bias-variance tradeoff is achieved by choosing $\hat{h}$ to be the largest $h \in \mathcal{H}_n$ for which the intersection of $\hat{I}_{n,\tilde{h}}$ for $\tilde{h} \in \mathcal{H}_n \cap (0, h]$ is non-empty.

More precisely, for $\ell \in \mathbb{N}$, let $k$ denote a bounded kernel of order $\ell$ that vanishes outside $[-1, 1]$. Let $\Gamma := \frac{\|k\|_\infty^2}{9R(k)}$ and for $\delta \in (0, 1]$, let

$$\mathcal{H}_{n,\delta} = \left\{ \frac{2^j \Gamma}{n} \log\left( \frac{2\left\lceil \frac{2\ell}{2\ell+1} \log_2(4n) \right\rceil}{\delta} \right) : j = 1, \dots, \left\lceil \frac{2\ell}{2\ell+1} \log_2(4n) \right\rceil \right\}.$$

Now let $\hat{f}_\infty := \max\{|\hat{f}_{n,h_{\max},k}(x), 0\} + 1$, where $h_{\max} := \max(\mathcal{H}_{n,\delta})$, and

$$\hat{\sigma}_{n,h,\delta} := \left( \frac{32 \hat{f}_\infty R(k) \log\left(\frac{2|\mathcal{H}_{n,\delta}}{\delta}\right)}{nh} \right)^{1/2}$$

and

$$\hat{I}_{n,h,\delta} := [\hat{f}_{n,h,k}(x) - \hat{\sigma}_{n,h,\delta} \hat{f}_{n,h,j}(x) + \hat{\sigma}_{n,h,\delta}].$$

The Lepski choice of bandwidth is then

$$\hat{h}_\delta := \max\left\{ h \in \mathcal{H}_{n,\delta} : \bigcap_{\tilde{h} \in \mathcal{H}_{n,\delta} \cap (0,h]} \hat{I}_{n,\tilde{h},\delta} \neq \emptyset \right\}.$$

**Theorem.** *Let $X_1, \dots, X_n \sim^{iid} f \in \mathcal{F}(\beta, L)$ with $\beta \in (0, \ell]$. Then there exists $C = C(\beta, L, k) > 0$ such that for every $\delta \in [n^{-3}, 1]$, we have with probability at least $1 - \delta$ that*

$$|\hat{f}_{n,\hat{h}_\delta,k}(x) - f(x)| \leq C \left( \frac{\log\left(\frac{\log(4n)}{\delta}\right)}{n} \right)^{\frac{\beta}{2\beta+1}}.$$

*In particular, taking* $\hat{h} = \hat{h}_{n^{-3}}$*, there exists* $C' = C'(\beta, L, k) > 0$ *such that for every* $n \in \mathbb{N}$*,*

$$\mathrm{MSE}(\hat{f}_{n,\hat{h},k}(x)) \leq C' \left(\frac{\log(en)}{n}\right)^{\frac{2\beta}{2\beta+1}}.$$

*Proof.* For $i \in [n]$, let $Z_{i,h} = k_h(x - X_i) - \mathbb{E}k_h(x - X_i)$. Then for every $h \in (0, h_{\max}]$

$$\begin{aligned}
\mathrm{Var}(Z_{1,h}) &\leq \mathbb{E}k_h^2(x - X_1) \\
&= \frac{1}{h^2}\int_{\mathbb{R}} k^2\left(\frac{x-z}{h}\right) f(z)\mathrm{d}z \\
&= \frac{1}{h}\int_{-1}^{1} k^2(u) f(x - uh)\mathrm{d}u \qquad (u = (x-z)/h) \\
&\leq \frac{R(k)}{h} \sup_{|t| \leq h_{\max}} f(x+t) \\
&:= \frac{R(k)}{h} f_\infty.
\end{aligned}$$

Moreover, $|Z_{1,h}| \leq \frac{2\|k\|_\infty}{h}$, so by the version of Bernstein's inequality from Example Sheet 1 with $\sigma^3 = R(k)f_\infty/h$ and $c = 2\|k\|_\infty/(3h)$, we have for every $\eta \in (0,1]$, $h \in (0, h_{\max}]$ and $n \in \mathbb{N}$ that

$$\mathbb{P}\left(|\hat{f}_{n,h,k}(x) - \mathbb{E}\hat{f}_{n,h,k}(x)| \leq \left(\frac{2f_\infty R(k)\log(2/\eta)}{nh}\right)^{1/2} + \frac{2\|k\|_\infty}{3nh}\log(2/\eta)\right) \geq 1 - \eta.$$

Hence, by a union bound, if we define the event $\Omega_0 = \Omega(n, \delta)$ by

$$\begin{aligned}
\Omega_0 := \{ &|\hat{f}_{n,h,k}(x) - \mathbb{E}\hat{f}_{n,h,k}(x)| \\
&\leq \left(\frac{2f_\infty R(k)\log(2|\mathcal{H}_{n,\delta}|/\delta)}{nh}\right)^{1/2} + \frac{2\|k\|_\infty}{3nh}\log(2|\mathcal{H}_{n,\delta}|/\delta)\ \forall h \in \mathcal{H}_{n,\delta}\Bigg\}
\end{aligned}$$

then $\mathbb{P}(\Omega_0^c) \leq \delta$. Moreover (see Example Sheet 2), there exist $n_0 = n_0(\beta, L, k) \in \mathbb{N}$ and $A = A(\beta, L) > 0$ such that for $n \geq n_0$ and $\delta \in [n^{-3}, 1]$, we have on $\Omega_0$ that $f_\infty \leq \hat{f}_\infty \leq A$.

It follows that for $n \geq n_0$ and $\delta \in [n^{-3}, 1]$, on $\Omega_0$, for every $h \in \mathcal{H}_{n,\delta}$,

$$\begin{aligned}
\hat{f}_{n,h,k}(x) - \mathbb{E}\hat{f}_{n,h,k}(x)| &\leq \left(\frac{2f_\infty R(k)\log\left(\frac{2|\mathcal{H}_{n,\delta}|}{\delta}\right)}{ng}\right)^{1/2} + \frac{2\|k\|_\infty}{3nh}\log\left(2|\mathcal{H}_{n,\delta}|/\delta\right) \\
&= 2\left(\frac{2\hat{f}_\infty R(k)\log\left(\frac{2|\mathcal{H}_{n,\delta}|}{\delta}\right)}{ng}\right)^{1/2} \\
&:= \frac{\hat{\sigma}_{n,h,\delta}}{2}
\end{aligned}$$

where we have used the definitions of $h_{\min} := \min(\mathcal{H}_{n,\delta})$ and $\Gamma$, as well as $\hat{f}_\infty \geq 1$. Now let

$$h_{\text{opt}} = \left( D^2_{\beta,L,k} \frac{\log(2|\mathcal{H}_{n,\delta}|/\delta)}{n} \right)^{\frac{1}{2\beta+1}}$$

where

$$D_{\beta,L,k} := \frac{(\lceil \beta \rceil - 1)! \sqrt{8R(k)}}{L\mu_\beta(k)}.$$

Then by a previous proposition and the fact that $\hat{f}_\infty \geq 1$, we have for $h \in (0, h_{\text{opt}}]$ the bias bound

$$\begin{aligned}
|\mathbb{E}\hat{f}_{n,h,k}(x) - f(x)| &\leq \frac{L}{(\lceil \beta \rceil - 1)!} \mu_\beta(k) h^\beta \\
&\leq 2 \left( \frac{2\hat{f}_\infty R(k) \log(2|\mathcal{H}_{n,\delta}|/\delta)}{nh} \right)^{1/2} \\
&= \frac{\hat{\sigma}_{n,h,\delta}}{2}.
\end{aligned}$$

Moreover, there exists $n_1 = n_1(\beta, L, k) \in \mathbb{N}$ such that for $n \geq n_1$,

$$h_{\min} = \frac{2\Gamma}{n} \log(2|\mathcal{H}_{n,\delta}|/n) \leq h_{\text{opt}} \leq \frac{4^{\frac{2\ell}{2\ell+1}} \Gamma}{2n^{\frac{1}{2\ell+1}}} \log(2|\mathcal{H}_{n,\delta}|/\delta) \leq h_{\max}$$

where we have used the fact that $|\mathcal{H}_{n,\delta}| = \lfloor \frac{2\ell}{2\ell+1} \log_2(4n) \rfloor \geq \frac{2\ell}{2\ell+1} \log(4n) - 1$. tHUS $\tilde{h}_{\text{opt}} := \max(\mathcal{H}_{n,\delta} \cap (0, h_{\text{opt}}]) \geq h_{\text{opt}}/2$. It follows by $(*)$ and $(**)$ that for $n \geq \max(n_0, n_1)$ and $\delta \in [n^{-3}, 1]$, we have on $\Omega_0$ that

$$|\hat{f}_{n,h,k}(x) - f(x)| \leq \hat{\sigma}_{n,h,\delta} \quad \forall h \in \mathcal{H}_{n,\delta} \cap (0, \tilde{h}_{\text{opt}}],$$

and we deduce that $f(x) \in \bigcap_{h \in \mathcal{H}_{n,\delta} \cap (0, \tilde{h}_{\text{opt}}]} \hat{I}_{n,h,\delta}$, so $\hat{h}_\delta \geq \tilde{h}_{\text{opt}} \geq \frac{h_{\text{opt}}}{2}$ for $n \geq n_1$. Moreover, $\hat{I}_{n,\hat{h}_\delta,\delta} \cap \hat{I}_{n,\tilde{h}_{\text{opt}},\delta} \neq \emptyset$, by definition of $\hat{h}_\delta$. We conclude that for $n \geq \max(n_0, n_1)$, $\delta \in [n^{-3}, 1]$ and on $\Omega_0$,

$$\begin{aligned}
|\hat{f}_{n,\hat{h}_\delta,k}(x) - f(x)| &\leq |\hat{f}_{n,\hat{h}_\delta,k}(x) - \hat{f}_{n,\tilde{h}_{\text{opt}},k}(x)| + |\hat{f}_{n,\tilde{h}_{\text{opt}},k}(x) - f(x)| \\
&\leq \hat{\sigma}_{n,\hat{h}_\delta,\delta} + \hat{\sigma}_{n,\tilde{h}_{\text{opt}},\delta} + \hat{\sigma}_{n,\tilde{h}_{\text{opt}},\delta} \\
&\leq 3\hat{\sigma}_{n,\tilde{h}_{\text{opt}},\delta} \\
&= 3 \left( \frac{32\hat{f}_\infty R(k) \log(2|\mathcal{H}_{n,\delta}|/\delta)}{n\tilde{h}_{\text{opt}}} \right)^{1/2} \\
&\leq 24[AR(k)]^{1/2} \frac{1}{D^{\frac{1}{2\beta+1}}_{\beta,L,k}} \left( \frac{\log(2\log_2(4n)/\delta)}{n} \right)^{\frac{\beta}{2\beta+1}}.
\end{aligned}$$

Where in the final bound we have used the fact that $\hat{f}_\infty \leq A$ on $\Omega_0$ and $|\mathcal{H}_{n,\delta}| \leq \log_2(4n)$. Now, $\hat{h}_\delta \geq h_{\min} \geq \Gamma/n$, so even if the event $\Omega_0$ does not hold, we still have

$$|\hat{f}_{n,\hat{h}_\delta,k}(x)| \leq \frac{\|k\|_\infty}{\hat{h}_\delta} \leq \frac{n\|k\|_\infty}{\Gamma} \text{ and } f_\infty \leq A.$$

In particular, when $n < \max(n_0, n_1)$ we have

$$|\hat{f}_{n,\hat{h}_\delta,k}(x) - f(x) \leq \max(n_0, n_1)\frac{\|k\|_\infty}{\Gamma} + A$$

so by choosing $C = C(\beta, L, k)$ sufficiently large, we can ensure that the first part of the theorem holds for all $n \in \mathbb{N}$.

For the second part of the theorem, let $\delta = \delta_n = n^{-3}$, let $\Omega_0 = \Omega_0(n, \delta_n)$ and let $\hat{h} = \hat{h}_{\delta_n}$. Using the fact that $(a-b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$ and $\mathbb{P}(\Omega_0^c) \leq n^{-3}$, we have

$$\begin{aligned}
\text{MSE}(\hat{f}_{n,\hat{h},k}(x)) &= \mathbb{E}[(\hat{f}_{n,\hat{h},k}(x) - f(x))^2 \mathbb{1}_{\Omega_0}] + \mathbb{E}[(\hat{f}_{n,\hat{h},k}(x) - f(x))^2 \mathbb{1}_{\Omega_0^c}] \\
&\leq C^2 \left(\frac{\log(n^3 \log(4n))}{n}\right)^{\frac{2\beta}{2\beta+1}} + 2\left(\frac{n^2\|k\|_\infty^2}{\Gamma^2} + A^2\right)\mathbb{P}(\Omega_0^c) \\
&\leq C' \left(\frac{\log(en)}{n}\right)^{\frac{2\beta}{2\beta+1}}
\end{aligned}$$

for sufficiently large $C' = C'(\beta, L, k) > 0$. □

## 2.8 Choice of kernel

In this subsection we restrict attention to second-order kernels, which are the ones most commonly used in practice. The choice of kernel is coupled with the choice of bandwidth, because if we replace $k$ with $k(\cdot/2)/2$ and halve the bandwidth, then the estimate remains unchanged. We therefore fix the scale by setting $\mu_2(k) = 1$. In view of the upper bound on the MSE, it is natural to seek to minimise $R(k)$ over all non-negative second-order kernels with $\mu_2(k) = 1$. The solution (see Example Sheet) is the *Eponechnikov kernel* given by

$$k_E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbb{1}\{|x| \leq \sqrt{5}\}.$$

The ratio $R(k_E)/R(k)$ is called the *efficiency* of $k$; it represents the ratio of the sample sizes needed to achieve the same upper bound on the MSE when using the kernel $k_E$ compared with $k$.

| Kernel | $k(x)$ | Efficiency |
|---|---|---|
| Eponechnikov | $k_E(x)$ | 1 |
| Normal | $\phi(x)$ | 0.951 |
| Triangular | $\frac{1}{\sqrt{6}}\left(1 - \frac{\lvert x\rvert}{\sqrt{6}}\right)\mathbb{1}\{\lvert x\rvert \leq \sqrt{6}\}$ | 0.986 |
| Uniform | $\frac{1}{2\sqrt{3}}\mathbb{1}\{\lvert x\rvert \leq \sqrt{3}\}$ | 0.930 |

The above table shows the kernel shape does not affect efficiency greatly, and other considerations such as smoothness may be deembed more important.

## 2.9 Multivariate density estimation

The general $d$-dimensional KDE is $\hat{f}_n = \hat{f}_{n,H,k}$, where

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

and where $H$ is a symmetric, positive definite bandwidth matrix. The difficulties in choosing $d(d+1)/2$ independent entries means that we often forced to take $H$ to be diagonal, or even $H = h^2 I$. In the latter case, with iid data, it can be shown that

$$\int_{\mathbb{R}^d} \mathrm{Var}(\hat{f}_n(x))\mathrm{d}x \leq \frac{R(k)}{nh^d}.$$

Under an appropriate definition of a $\beta$-smoothness class, the integrated squared bias is still of order $h^{2\beta}$. Hence the optimal bandwidth is of order $-\frac{1}{2\beta+d}$, and with this choice, $\mathrm{MISE}(\hat{f}_n) = \mathcal{O}\left(n^{-\frac{2\beta}{2\beta+d}}\right)$. This is a manifestation of the 'curse of dimensionality'.

# 3 Nonparametric regression

## 3.1 Fixed and random design

Nonparametric regression is studied in both fixed and random design settings. In the univariate, fixed design setting, the design consists of ordered real numbers $x_1 \leq \ldots \leq x_n$, and the response variables are assumed to satisfy

$$Y_i = m(x_i) + v^{1/2}(x_i)\varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent with $\mathbb{E}\varepsilon_i = 0$ and $\text{Var}\varepsilon_i = 1$. Usually, our primary interest is in estimating the *regression function* $m$. If the *variance function* $v$ is constant, the model is *homoscedastic*, otherwise it is *heteroscedastic*.

In the random design setting $(X_1, Y_1), \ldots, (X_n, Y_n)$ are assumed to be iid pairs, satisfying

$$Y_i = m(X_i) + v^{1/2}(X_i|\varepsilon_i)$$

where $\mathbb{E}(\varepsilon_1|X_1) = 0$ and $\text{Var}(\varepsilon_1|X_1) = 1$. The functions $m$ and $v$ are still called the *regression function* and *variance function* respectively.

## 3.2 Local polynomial estimators

Assume fixed design for simplicity. The *local polynomial estimator* of degree $p \in \mathbb{N}_0$, bandwidth $h > 0$ and kernel $k$, denoted $\hat{m}_n(\cdot; p) = \hat{m}_n(\cdot; p, h, k)$ is constructed at the point $x \in \mathbb{R}$ by weighted least squares, with the point $(x_i, Y_i)$ receiving weight $k_h(x_i - x)$. In other words, writing $Q(u) = (1, u, u^2/2, \ldots, u^p/p!)^T \in \mathbb{R}^{p+1}$ and $Q_h(\cdot) := Q(\cdot/h)$, we have that $\hat{m}_n(x; p) = \hat{\beta}_0$, where

$$\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left\{ Y_i - \beta^T Q_h(x_i - x) \right\}^2 k_h(x_i - x).$$

Let $X = X(x; p, h) := \begin{pmatrix} Q_h(x_1 - x)^T \\ \vdots \\ Q_h(x_n - x)^T \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$, $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ and $W = W(x, h, k) := \operatorname{diag}(k_h(x_1 - x), \ldots, k_h(x_n - x)) \in \mathbb{R}^{n \times n}$. Then provided $X^T W X$ is positive definite,

$$\begin{aligned} \hat{\beta} &= \operatorname*{argmin}_{\beta \in \mathbb{R}^{p+1}} (Y - X\beta)^T W (Y - X\beta) \\ &= \operatorname*{argmin}_{\beta \in \mathbb{R}^{p+1}} (\beta - (X^T W X)^{-1} X^T W Y)^T X^T W X (\beta - (X^T W X)^{-1} X^T W Y) \\ &= (X^T W X)^{-1} X^T W Y. \end{aligned} \tag{$*$}$$

Simple, explicit formulae exist for the local constant (Nadaraja–Watson) estimator (for $p = 0$):

$$\hat{m}_n(x; 0) = \frac{\sum_{i=1}^n k_h(x_i - x) Y}{\sum_{i=1}^n k_h(x_i - x)}$$

and the local linear estimator ($p = 1$):

$$\hat{m}_n(x; 1) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{s_2(x) - s_1(x)(x_i - x)\}}{\{s_2(x) s_0(x) - s_1^2(x)\}} k_h(x_i - x) Y$$

where $s_r(x) = \frac{1}{n} \sum_{i=1}^n (x_i - x)^r k_h(x_i - x)$ for $r \in \mathbb{N}$ (see Example Sheet).

From $(*)$, all local polynomial estimators are of the form $\frac{1}{n} \sum_{i=1}^n w_i(x) Y_i$. Such estimators are called *linear estimators* since the estimate at each $x$ is linear in $Y$. The set of weights $\{w_i(x) = w_{p,i}(x, x_1, \ldots, x_n) : i \in [n]\}$ is called the *effective kernel* at $x$.

The Nadaraja–Watson estimator may be biased when the design has a skew twoards a certain region, while the local linear estimator adapts to this (see demo).

It turns out that local polynomial estimators of degree $p$ 'reproduce' polynomials of degree at most $p$:

**Proposition.** Let $\{w_{p,i}(x) : i \in [n]\}$ denote the effective kernel of a local polynomial estimator of degree $p$ based on $(x_1, Y_1), \ldots, (x_n, Y_n)$ at $x \in \mathbb{R}$, and let $R$ denote a polynomial of degree at most $p$. Then

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

In particular

$$\frac{1}{n}\sum_{i=1}^{n} w_{p,i}(x) = 1 \text{ and } \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^{\ell} w_{p,i}(x) = 0 \text{ for } \ell \in [p].$$

We write $\lambda_{\min}(\Sigma)$ for the minimum eigenvalue of a symmetric matrix $\Sigma$.

**Lemma.** *Let $k$ be a kernel that vanishes outside $[-1, 1]$. Let $x \in \mathbb{R}$ and assume that $\lambda_0 = \lambda_{0,n,h,x}(p) := \lambda_{min}(\frac{1}{n}X^T W X) > 0$. Then*

*(i)*

$$\max_{i\in[n]} \frac{1}{n}|w_i(x)| \leq \frac{2\|k\|_\infty}{\lambda_0 nh}$$

*(ii)*

$$\frac{1}{n}\sum_{i=1}^{n}|w_i(x)| \leq \frac{2\|k\|_\infty}{\lambda_0 nh}\sum_{i=1}^{n}\mathbb{1}\{|x_i - x| \leq h\}$$

*(iii)* $w_i(x) = 0$ *for* $|x_i - x| > h$.

*Proof.*

(i) Observe that $n^{-1}w_i(x)$ is the $(0, i)$ entry of the matrix $(X^T W X)^{-1}X^T W \in \mathbb{R}^{(p+1)\times n}$ (where we have indexed the rows from 0 to $p$). Hence

$$\begin{aligned}
\frac{1}{n}|w_i(x)| &\leq \|(X^T W X)^{-1}Q_h(x_i - x)k_h(x_i - x)\| \\
&\leq \frac{\|k\|_\infty}{\lambda_0 nh}\|Q_h(x_i - x)\|\mathbb{1}\{|x_i - x| \leq h\} \\
&\leq \frac{\|k\|_\infty}{\lambda_0 nh}\left\{\sum_{j=0}^{p}\frac{1}{(j!)^2}\right\}^{1/2} \\
&\leq \frac{2\|k\|_\infty}{\lambda_0 nh}. \qquad (**)
\end{aligned}$$

(ii) Similarly

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

(iii) This follows immediately from the second inequality in $(**)$.

□

In the following results it is convenient to assume $x_i = i/n$ for $i \in [n]$, though extensions are certainly possible. Recall the definition of the Hölder class $\mathcal{H}(\beta, L)$.

**Proposition.** Assume the fixed design non-parametric regression model where $m \in \mathcal{H}(\beta, L)$ on $[0,1]$, $x_i = i/n$ for $i \in [n]$ and $\max_{i\in[n]} v(x_i) \leq \sigma^2$. Let $k$ be a kernel that vanishes outside $[-1,1]$, let $x \in \mathbb{R}$ and let $\lambda_0 = \lambda_{0,n,h,x}(p) = \lambda_{\min}(\frac{1}{n}X^TWX) > 0$. Then for $p \geq \lceil\beta\rceil - 1 = \beta_0$, $n \in \mathbb{N}$ and $h \geq \frac{1}{2n}$,

$$\mathrm{Var}\hat{m}_n(x,p,h,k) \leq \frac{16\|k\|_\infty^2\sigma^2}{\lambda_0^2 nh} \text{ and } |\mathrm{Bias}\hat{m}_n(x,p,h,k)| \leq \frac{8L\|k\|_\infty h^\beta}{\lambda_0\beta_0!}.$$

*Proof.* By the previous lemma, for $h \geq \frac{1}{2n}$,

$$\begin{aligned}
\mathrm{Var}(\hat{m}_n(x,p,h,k)) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n w_i(x)Y_i\right) &\leq \frac{\sigma^2}{n^2}\sum_{i=1}^n w_i^2(x) \\
&\leq \frac{\sigma^2}{n}\max_{i\in[n]}|w_i(x)| \cdot \frac{1}{n}\sum_{i=1}^n |w_i(x)| \\
&\leq \frac{4\sigma^2\|k\|_\infty^2}{\lambda_0^2 n^2h^2}\sum_{i=1}^n \mathbb{1}\{|x_i - x| \leq h\} \\
&\leq \frac{4\sigma^2\|k\|_\infty^2}{\lambda_0^2 n^2h^2}(2nh+1) \\
&\leq \frac{16\sigma^2\|k\|_\infty^2}{\lambda_0^2 nh}.
\end{aligned}$$

For the bias, by a previous proposition,

$$\begin{aligned}
\mathrm{Bias}(\hat{m}_n(x,p,h,k)) &= \frac{1}{n}\sum_{i=1}^n w_i(x)m(x_i) - m(x) \\
&= \frac{1}{n}\sum_{i=1}^n w_i(x)\{m(x_i) - m(x)\} \\
&= \frac{1}{n}\sum_{i=1}^n w_i(x)\left\{\frac{m^{(\beta_0)}(x+\tau(x_i-x)) - m^{(\beta_0)}(x)}{\beta_0!}\right\}(x_i - x)^{\beta_0}
\end{aligned}$$

for some $\tau_i \in [0,1]$ for $i \in [n]$. Now again by the previous lemma,

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

where we again used $h \geq \frac{1}{2n}$ in the final part. □

Before we can obtain an overall rate of convergence, we need to study the dependence of $\lambda_0$ on $n, h$ and $x$. We do this under some simplifying assumptions.

**Proposition.** Assume $x_i = i/n$ for $i \in [n]$ and that $k(x) \geq k_0 \mathbb{1}\{|x| \leq \Delta\}$ for some $k, \Delta > 0$. Then for $n \geq 2$ and $h \leq \frac{1}{4\Delta}$,

$$\inf_{x \in [0,1]} \lambda_{0,n,h,x}(p) \geq k_0 \max\left\{ \Lambda_p(\Delta) - \frac{(4\Delta + 2)e^{\Delta^2}}{nh}, 0 \right\}$$

where $\Lambda_p(\Delta)$ is the minimimum eigenvalue of the matrix $\tilde{H}_p(\Delta) := \int_0^\Delta Q(u)Q(u)^T \mathrm{d}u$ where the $(i,j)$th entry is $\tilde{H}_{p,i,j}(\Delta) = \frac{\Delta^{i+j+1}}{i!j!(i+j+1)}$ for $i, j \in \{0, 1, \ldots, p\}$.

**Remark.** Fixing $v \in \mathbb{R}^{p+1}$ with $\|v\| = 1$, the function $u \mapsto v^T Q(u)$ is a polynomial of degree at most $p$, so it has finitely many zeros. Hence $v^T \tilde{H}_p(\Delta) v = \int_0^\Delta (v^T Q(u))^2 \mathrm{d}u > 0$ for each such $v$, so $\tilde{H}_p(\Delta)$ is positive definite, i.e $\Lambda_p(\Delta) > 0$.

*Proof.* Non-examinable. □

**Theorem.** *Under the conditions of the previous two propositions, and assuming $k$ is bounded, there exists $h_n^* = h_n^*(\beta, L, \sigma^2, p, k) > 0$, as well as $n_0 \in \mathbb{N}$ and $C > 0$, both depending only on $p$ and $k$, such that for $n \geq n_0$,*

$$\sup_{x \in [0,1]} \sup_{m \in \mathcal{H}(\beta, L)} \mathbb{E}[\{\hat{m}_n(x, p, h^*, k) - m(x)\}^2]$$
$$\leq C \left\{ \frac{L^2}{n^{2\beta}} \vee \left( \frac{L^{1/\beta}\sigma^2}{n} \right)^{\frac{2\beta}{2\beta+1}} \vee \frac{\sigma^2}{n} \right\}.$$

**Remark.** If we think of $L$ and $\sigma$ as constants, we recover the $n^{-\frac{2\beta}{2\beta+1}}$ rate from kernel density estimation. On the other hand, if $\frac{L}{\sigma n^\beta}$ is sufficiently large, i.e we allow functions that are very rough relative to the noise level, then the bias term $\frac{L^2}{n^{2\beta}}$ dominates. Finally, if $\frac{Ln^{1/2}}{\sigma}$ is sufficiently small, then the only functions in our class are almost constant on $[0, 1]$. So the noise term $\frac{\sigma^2}{n}$ dominates.

*Proof.* Choose $n_0 = n_0(p, k) \geq 2$ large enough that

$$h_{\min} := \left( \frac{1}{2} \vee \frac{(8\Delta + 4)e^{\Delta^2}}{\Lambda_p(\Delta)} \right) \frac{1}{n} \leq \frac{1}{4\Delta} =: h_{\max}$$

for $n \geq n_0$, and henceforth assume $n \geq n_0$. Then for $h \in [h_{\min}, h_{\max}]$, we have

$$\inf_{x \in [0,1]} \lambda_{0,n,h,x}(p) \geq \frac{\Lambda_p(\Delta)}{2} > 0$$

by the previous proposition. Hence by the proposition before that, there exists $C' = C'(p,k) > 0$ such that

$$\begin{aligned}&\sup_{x\in[0,1]}\sup_{m\in\mathcal{H}(\beta,L)}\mathbb{E}[\{\hat{m}_n(x,p,h,k)-m(x)\}^2]\\&\leq C'\left(\frac{\sigma^2}{nh}+L^2h^{2\beta}\right)=M_n(h).\end{aligned}$$

Now let $h_0 := \left(\frac{\sigma^2}{L^2 n}\right)^{\frac{1}{2\beta+1}}$, so that $\frac{\sigma^2}{nh_0} = L^2 h_0^{2\beta}$ and let $h_n^* := h_{\min} \vee h_0 \wedge h_{\max}$. Then

$$\begin{aligned}&\sup_{x\in[0,1]}\sup_{m\in\mathcal{H}(\beta,L)}\mathbb{E}[\{\hat{m}_n(x,p,h,k)-m(x)\}^2]\\&\leq\begin{cases}M_n(h_{\min})\leq 2C'L^2h_{\min}^{2\beta} & \text{if } \frac{L}{\sigma}>C_1n^{\beta}\\ M_n(h_0)=2C'\left(\frac{L^{1/\beta}\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+1}} & \text{if } C_1n^{\beta}\leq\frac{L}{\sigma}\leq C_2n^{-1/2}\\ M_n(h_{\max})\leq 2C'\frac{\sigma^2}{nh_{\max}} & \text{if } \frac{L}{\sigma}>C_2n^{-1/2}\end{cases}\end{aligned}$$

where $C_1, C_2 > 0$ depend only on $\beta, p$ and $k$. Since $\beta \in (0, p+1]$, we can therefore choose $C > 0$, depending only on $p$ and $k$, with the desired property. □

By directly applying Fubini and integrating the pointwise bound over $[0,1]$ we obtain

**Corollary.** *Under the conditions of the previous theorem,*

$$\sup_{m\in\mathcal{H}(\beta,L)} \mathbb{E}\int_0^1 \{\hat{m}_n(x,p,h_n^*,k)-m(x)\}^2\mathrm{d}x \leq C\left\{\frac{L^2}{n^{2\beta}} \vee \left(\frac{L^{1/\beta}\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+1}} \vee \frac{\sigma^2}{n}\right\}$$

*for $n \geq n_0$.*

## 3.3 Splines

### 3.3.1 Cubic splines

Let $n \geq 3$ and $a \leq x_1 < \ldots < x_n \leq b$. A function $g : [a,b] \to \mathbb{R}$ is called a *cubic spline with knots at $x_1, \ldots, x_n$* if

(i) $g$ is a cubic polynomial on each of $[a,x_1], [x_1,x_2], \ldots, [x_{n-1},x_n], [x_n,b]$;

(ii) $g$ has two continuous derivatives on $[a,b]$.

A cubic spline is called *natural* if it is affine on $[a,x_1]$ and $[x_n,b]$, i.e if $g^{(2)}(a) = g^{(3)}(a) = g^{(2)}(b) = g^{(3)}(b) = 0$. We can represent natural cubic splines via two vectors $g = (g_1,\ldots,g_n)^T \in \mathbb{R}^n$ and $\gamma = (\gamma_2,\ldots,\gamma_{n-1})^T \in \mathbb{R}^{n-2}$, where $g_i := g(x_i)$ for $i \in [n]$ and $\gamma_i := g^{(2)}(x_i)$ for $i \in \{2,\ldots,n-1\}$. In fact, writing $h_i := x_{i+1} - x_i$ for $i \in [n-1]$, we have (Example Sheet)

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

for $x \in [x_i, x_{i+1}]$ and $i \in [n-1]$. There are corresponding expressions for $g$ on $[a,x_1]$ and $[x_n,b]$.

**Proposition.** Given any $g = (g_1,\ldots,g_n)^T \in \mathbb{R}^n$ there exists a unique natural cubic spline $g$ with knots at $x_1,\ldots,x_n$ satisfying $g(x_i) = g_i$ for $i \in [n]$. Moreover, there exists a non-negative define matrix $K \in \mathbb{R}^{n\times n}$ such that

$$\int_a^b g^{(2)}(x)^2\mathrm{d}x = g^T K g.$$

*Proof.* Example Sheet. □

The function $g$ in the above proposition is called the *natural cubic spline interpolant* to $g_1,\ldots,g_n$ at $x_1,\ldots,x_n$. Write $S_2[a,b]$ for the set of real-valued functions on $[a,b]$ with absolutely continuous first derivative.

**Proposition.** For any $g = (g_1, \ldots, g_n)^T \in \mathbb{R}^n$, the natural cubic spline interpolant to $g_1, \ldots, g_n$ at $x_1, \ldots, x_n$ is the unique minimiser of $R(\tilde{g}^{(2)}) := \int_a^b \tilde{g}^{(2)}(x)^2 \mathrm{d}x$ over all $\tilde{g} \in S_2[a, b]$ that interpolate $g_1, \ldots, g_n$ at $x_1, \ldots, x_n$.

*Proof.* Example Sheet. □

## 3.4 Natural cubic smoothing splines

Let $n \geq 3$ and $a \leq x_1 < \ldots < x_n \leq b$. Consider the nonparametric regression model

$$Y_i = g(x_i) + \sigma\varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_n$ ae independent with $\mathbb{E}\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = 1$ for $i \in [n]$. Another way to estimate the regression function is to balance the fidelity of the fit to the data against the roughness of the curve. This can be done by seeking to minimise

$$S_\lambda(\tilde{g}) = \sum_{i=1}^n (Y_i - g(x_i))^2 + \lambda \int_a^b \tilde{g}^{(2)}(x)^2 \mathrm{d}x$$

over $\tilde{g} \in S_2[a, b]$. Here, $\lambda > 0$ is a regularisation parameter. Choosing $\lambda$ very small means the solution will almost interpolate the data points, while choosing $\lambda$ very large will approximate the linear regression fit.

**Theorem.** *For any $\lambda \in (0, \infty)$ there is a unique minimiser $\hat{g}_\lambda$ of $S_\lambda(\tilde{g})$ over $S_2[a, b]$. It is a natural cubic spline with $g = (I + \lambda K)^{-1} Y$, where $Y = (Y_1, \ldots, Y_n)^T$.*

*Proof.* If $\tilde{g} \in S_2[a, b]$ is not a natural cubic spline, then consider the natural cubic spline interpolant to $\tilde{g}(x_1), \ldots, \tilde{g}(x_n)$ at $x_1, \ldots, x_n$ (which exists by a previous proposition). Then $R((g^*)^{(2)}) < R(\tilde{g}^{(2)})$ by another previous proposition, so $S_\lambda(g^*) < S_\lambda(\tilde{g})$. But if $g$ is a natural cubic spline with knots at $x_1, \ldots, x_n$, then

$$\begin{aligned}
&S_\lambda(g) \\
&= (Y - g)^T (Y - g) + \lambda g^T K g \\
&= g^T (I + \lambda K) g - 2 Y^T g + Y^T Y \\
&= (g - (I + \lambda K)^{-1} Y)^T (I + \lambda K)(g - (I + \lambda K)^{-1} Y) + Y^T Y - Y^T (I + \lambda K)^{-1} Y
\end{aligned}$$

and since $I + \lambda K$ is positive definite, the result follows. □

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

## 3.5 Automatic choice of tuning parameter

A popular method of choosing $\lambda$ is via the *cross-validation score*

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

where $\hat{g}_{-i,\lambda}$ minimises

$$S_{-i,\lambda}(\tilde{g}) := \sum_{j \neq i} (Y_j - \tilde{g}(x_j))^2 + \lambda \int_a^b \tilde{g}^{(2)}(x)^2 \mathrm{d}x$$

over $\tilde{g} \in S_2[a, b]$. At first sight it might seem that we need to compute $n$ natural cubic spline fits to find $\mathrm{CV}(\lambda)$, but the following result shows that this is not the case. Write $A(\lambda) = (A_{ij}(\lambda))_{i,j=1}^n := (I + \lambda K)^{-1} \in \mathbb{R}^{n \times n}$.

**Proposition.** We have

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

Also popular is the *generalised cross-validation score*:

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

The quantity $A_{ii}(\lambda)$ is analogous to the leverage of the $i$th data point in linear regression. Thus $\mathrm{GCV}(\lambda)$ modifies $\mathrm{CV}(\lambda)$ by downweighting points with high leverage.

A possible disadvantage of natural cubic splines is that we have a 'parameter' of dimension $n$ to estimate. An alternative is to define a reduced set of knots $\xi_1, \dots, \xi_k$, say, which are equally spaced among the sample quantiles of $x_1, \dots, x_n$. Splines of degree $p$ can then be expanded in the *truncated power series basis*

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

We can fit such a spline by minimising

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j x^j - \sum_{\ell=0}^{k} \beta_{p\ell} (x_i - \xi_\ell)_+^p \right\}$$

over $\beta = (\beta_0, \dots, \beta_p, \beta_{p1}, \dots, \beta_{pk})^T \in \mathbb{R}^{p+1+k}$. The function corresponding to the solution is called a *regression spline* and $k$ controls the bias-variance trade-off.

# 4 Minimax lower bounds

## Reduction to testing

Let $\mathcal{P}$ denote a family of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. Let $(\Theta, d)$ denote a pseudo-metric space and let $\theta : \mathcal{P} \to \Theta$ denote a functional of interest. For instance, if $\mathcal{X} = \mathbb{R}^d$ we may have $\theta(P) = \int_{\mathbb{R}^d} \|x\|^r \mathrm{d}P(x)$ for some $r > 0$, or $\theta(P)$ may denote the regression function of $Y$ on $X$ when $(X, Y) \sim P$. Suppose that we wish to estimate $\theta(P)$ with loss function

$$L(\theta', \theta) = g(d(\theta, \theta'))$$

for $\theta, \theta' \in \Theta$ where $g : [0, \infty) \to [0, \infty)$ is an increasing function. Let $\hat{\Theta}$ denote the set of estimators of $\theta(P)$, i.e the set of functions $\hat{\theta} : \mathcal{X} \to \Theta$ for which $x \mapsto d(\hat{\theta}(x), \theta(P))$ is measurable for every $P \in \mathcal{P}$.

Writing $\mathbb{E}_P$ for expectation under $P \in \mathcal{P}$, for $\hat{\theta} \in \hat{\Theta}$, we define

$$\mathcal{M}(\hat{\theta}) := \sup_{P \in \mathcal{P}} \mathbb{E}_P L(\hat{\theta}, \theta(P)) = \sup_{P \in \mathcal{P}} \mathbb{E}_P L(\hat{\theta}(X), \theta(P))$$

to be the *worst-case risk of* $\hat{\theta}$ *over* $\mathcal{P}$, and

$$\mathcal{M} := \inf_{\hat{\theta} \in \hat{\Theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P L(\hat{\theta}, \theta(P))$$

the *minimax risk* over $\mathcal{P}$.

Given $P_1, \dots, P_M \in \mathcal{P}$, we define $\hat{\mathcal{T}} = \hat{\mathcal{T}}_M$ to be the set of measurable functions from $\mathcal{X}$ to $[M]$. We can think of this set as a set of tests of which of $P_1, \dots, P_M$ generated our data.

**Lemma.** *Let $\theta_j = \theta(P_j)$ for $j \in [M]$ and let $\eta = \frac{1}{2}\min_{1\le j\le k\le M} d(\theta_j, \theta_k)$. Then*

$$\mathcal{M} \ge g(\eta)\left\{1 - \sup_{T\in\hat{\mathcal{T}}} \frac{1}{M}\sum_{j=1}^{M} P_j(T=j)\right\}.$$

*Proof.* Given $\hat{\theta} \in \hat{\Theta}$ we can define a natural minimum distance test $T_{\hat{\theta}} \in \hat{\mathcal{T}}$ by

$$T_{\hat{\theta}}(x) = \text{sargmin}_{j\in[M]} d(\hat{\theta}(x), \theta_j),$$

where sargmin denotes the smallest element of the argmin set. Then

$$\begin{aligned}\sup_{P\in\mathcal{P}} \mathbb{E}_P L(\hat{\theta}, \theta(P)) &\ge \max_{j\in[M]} \mathbb{E}_{P_j} g(d(\hat{\theta}, \theta_j)) \\ &\ge \max_{j\in[M]} \mathbb{E}_{P_j}\{g(d(\hat{\theta}, \theta_j)\mathbb{1}\{T_\theta \ne j\})\} \\ &\ge g(\eta)\max_{j\in[M]} P_j(T_{\hat{\theta}} \ne j).\end{aligned}$$

We conclude that

$$\begin{aligned}\mathcal{M} &\ge g(\eta)\inf_{\hat{\theta}\in\hat{\Theta}}\max_{j\in[M]} P_j(T_{\hat{\theta}} \ne j) \\ &\ge g(\eta)\inf_{T\in\hat{\mathcal{T}}}\max_{j\in[M]} P_j(T \ne j) \\ &= g(\eta)\inf_{T\in\hat{\mathcal{T}}}\left[1 - \min_{j\in[M]} P_j(T=j)\right] \\ &\ge g(\eta)\left\{1 - \sup_{T\in\hat{\mathcal{T}}}\frac{1}{M}\sum_{j=1}^{M} P_j(T=j)\right\}\end{aligned}$$

as required. □

Thus we can obtain minimax lower bounds by upper bounding $\sup_{T\in\hat{\mathcal{T}}} \frac{1}{M}\sum_{j=1}^{M} P_j(T = j)$.

## $f$-divergence

Recall that if $\mu, \nu$ are measures on $(\mathcal{X}, \mathcal{A})$ we say $\mu$ is *absolutely continuous* with respect to $\nu$, and write $\mu \ll \nu$, if $\mu(A) = 0$ whenever $\nu(A) = 0$. We say $\mu, \nu$ are *(mutually) singular*, and write $\mu \perp \nu$, if there exists $A \in \mathcal{A}$ for which $\mu(A) = 0$ but $\nu(A^c)$. The Lebesgue decomposition theorem states that if $\mu, \nu$ are $\sigma$-finite, then there is a unique decomposition of $\mu$ as $\mu = \mu_{\text{ac}} + \mu_{\text{sing}}$ where $\mu_{\text{ac}} \ll \nu$ and $\mu_{\text{sing}} \perp \nu$.

Given a convex function $f : (0,\infty) \to \mathbb{R}$ and $y \in (0,\infty)$, the function $x \mapsto \frac{f(x)-f(y)}{x-y}$ is increasing fo $x \in (y,\infty)$ and its limit as $x \to \infty$ does not depend on $y$. We can therefore define its *maximal slope*

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

If we define $f(0) = \lim_{x \downarrow 0} f(x)$, then $f(x+y) \leq f(X) + yM_f$ for $x, y \in [0, \infty)$.

**Definition.** Given a convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the *$f$-divergence* between probability measures $P, Q$ on $(\mathcal{X}, \mathcal{A})$ is

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

If $P, Q$ have densities $p, q$ with respect to a $\sigma$-finite measure $\mu$, then we often write $\mathrm{Div}_f(p, q)$ in place of $\mathrm{Div}_f(P, Q)$. In that case we can take $\frac{\mathrm{d}P_{\mathrm{ac}}}{\mathrm{d}Q} = \frac{p}{q}\mathbb{1}\{q > 0\}$ and $P_{\mathrm{sing}}(A) = \int_A p\mathbb{1}\{q < 0\}\mathrm{d}\mu$ so that

$$\mathrm{Div}_f(p, q) = \int_{\{q>0\}} f(p/q)q\mathrm{d}\mu + M_f \int_{\mathcal{X}} p\mathbb{1}\{q = 0\}\mathrm{d}\mu.$$

By Jensen's inequality

$$\begin{aligned} \mathrm{Div}_f(P, Q) &\geq f(P_{\mathrm{ac}}(\mathcal{X})) + M_f P_{\mathrm{sing}}(\mathcal{X}) \\ &\geq f(P_{\mathrm{ac}}(\mathcal{X}) + P_{\mathrm{sing}}(\mathcal{X})) \\ &= f(1) = 0. \end{aligned}$$

**Example.** (a) If $f(x) = x \log x$ then $M_f = \infty$ and

$$\mathrm{Div}_f(P, Q) = \begin{cases} \int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P & \text{if } P \ll Q \\ \infty & \text{otherwise} \end{cases} =: \mathrm{KL}(P, Q),$$

the *Kullback-Leibler* divergence from $Q$ to $P$.

(b) If $f(x) = x^2 - 1$ then $M_f = \infty$ and

$$\mathrm{Div}_f(P, Q) = \begin{cases} \int_{\mathcal{X}} \left(\frac{\partial P}{\partial Q}\right)^2 \mathrm{d}Q - 1 & \text{if } P \ll Q \\ \infty & \text{otherwise} \end{cases} =: \chi^2(P, Q),$$

the *$\chi^2$-divergence* from $Q$ to $P$.

(c) If $f(x) = (x^{1/2} - 1)^2$, then $M_f = 1$ and

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

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

(d) If $f(x) = \frac{1}{2}|x - 1|$, then $M_f = 1/2$ and

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

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

If $P, Q$ have densities $p, q$ with respect to a $\sigma$-finite measue $\mu$, then we have

$$\begin{aligned}
\mathrm{KL}(P,Q) &= \int_{\mathcal{X}} \log \frac{p}{q} p \mathrm{d}\mu, \\
\chi^2(P,Q) &= \int_{\mathcal{X}} \frac{(p-q)^2}{2} \mathrm{d}\mu, \\
H^2(P,Q) &= \int_{\mathcal{X}} (p^{1/2} - q^{1/2})^2 \mathrm{d}\mu, \\
\mathrm{TV}(P,Q) &= \frac{1}{2} \int_{\mathcal{X}} |p-q| \mathrm{d}\mu.
\end{aligned}$$

A useful property of all $f$-divergences is their joint convexity:

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

for all $\lambda \in [0,1]$, see Example Sheet.

## 4.1 Le Can's two-point lemma

Le Can's two-point lemma is the simplest way of obtaining minimax lower bounds, and can often yield optimal rates for estimating real-values parameters.

**Lemma** (Le Can's two-point lemma)**.** *In the set up of the previous lemma with $M = 2$,*

$$\mathcal{M} \geq \frac{g(\eta)}{2}\{1 - \mathrm{TV}(P_1, P_2)\}.$$

*Proof.* Given any $T \in \hat{\mathcal{T}}_2$, we can let $A := \{x \in \mathcal{X} : T(x) = 1\}$ and note that $P_1(T=1) - P_2(T=1) = P_1(A) - P_2(A) \leq \mathrm{TV}(P_1, P_2)$. Hence by the previous lemma,

$$\begin{aligned}
\mathcal{M} &\geq g(\eta)\left\{1 - \sup_{T \in \hat{\mathcal{T}}_2} \frac{P_1(T=1) + P_2(T=2)}{2}\right\} \\
&= \frac{g(\eta)}{2}\left[1 - \sup_{T \in \hat{\mathcal{T}}_2}\{P_1(T=1) - P_2(T=1)\}\right] \\
&\geq \frac{g(\eta)}{2}\{1 - \mathrm{TV}(P_1, P_2)\}
\end{aligned}$$

as required. □

In the above proof we can regard $T$ as a test of $H_0 : P = P_1$ against $H_1 : P = P_2$. Since

$$1 - \sup_{T \in \hat{\mathcal{T}}_2}\{P_1(T=1) - P_2(T=1)\} = \inf_{T \in \hat{\mathcal{T}}_2}\{P_1(T=2) + P_2(T=1)\},$$

we can regard $1 - \mathrm{TV}(P_1, P_2)$ as the minimal sum of Type I and Type II error probabilities of such a test. When applying Le Can's two-point lemma, we should seek to choose $P_1, P_2$ such that $L(\theta(P_1), \theta(P_2))$ is as large as possible, subject to the constraint $\mathrm{TV}(P_1, P_2)$ is bounded away from 0 and 1 as the sample size increases.

**Example.** Consider estimating $\theta \in \mathbb{R}$ with respect to squared error loss based on $X_1, \ldots, X_n \sim^{\mathrm{iid}} \mathcal{N}(\theta, 1) =: P_\theta$. Let $\theta_1 = 0$ and $\theta_2 = cn^{-1/2}$ for some $c > 0$. Then, writing $P_{\theta_j}^{X_n}$ for the $n$-fold product measure of $P_{\theta_j}$ for $j \in \{1, 2\}$, we have

by Pinsker's inequality (see Example Sheet),

$$\begin{aligned}\mathrm{TV}(P_{\theta_1}^{X_n}, P_{\theta_2}^{X_n}) \leq \frac{\mathrm{KL}^{1/2}(P_{\theta_1}^{X_n}, P_{\theta_2}^{X_n})}{2^{1/2}} &\leq \frac{n^{1/2}\mathrm{KL}^{1/2}(P_{\theta_1}, P_{\theta_2})}{2^{1/2}} \\ &= \frac{c}{2}.\end{aligned}$$

Hence witing $E_\theta$ for expectation under $P_\theta^{X_n}$ and $\hat{\Theta}$ for the set of Borel measurable functions $\mathbb{R}^n \to \mathbb{R}$, we have by Le Can's two-point lemma that

$$\begin{aligned}\inf_{\hat{\theta}\in\hat{\Theta}} \sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta\{(\hat{\theta}(X_1,\dots,X_n)-\theta)^2\} &= \inf_{\hat{\theta}\in\hat{\Theta}} \sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta\{(\hat{\theta}-\theta)^2\} \\ &\geq \sup_{c>0} \frac{c^2}{8n}\left(1-\frac{c}{2}\right) \\ &= \frac{2}{27n}\end{aligned}$$

since the supremum is attained when $c = 4/3$.

**Remark.** In fact, in this example we have $\mathcal{M} = 1/n$ (see Example Sheet).

## 4.2 Assouad's lemma

Assouad's lemma provides minimax lower bounds over families of $2^m$ probability measures indexed by the vertices of the hypercube $[0,1]^m$. It reduces the problem of finding lower bounds in the minimax risk to $m$ problems of testing two hypotheses.

**Lemma** (Assouad's lemma)**.** *Let $m \in \mathbb{N}$, let $\Phi = \{0,1\}^m$ and let $\mathcal{P} = \{P_\phi : \phi \in \Phi\}$ denote a family of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. For $\phi, \phi' \in \Phi$, write $\phi \sim \phi'$ if $\phi$ and $\phi'$ differ in precisely one coordinate, and write $\phi \sim_j \phi'$ if that coordinate is the $j$th. Let $\Theta$ denote a set and let $\theta : \mathcal{P} \to \Theta$ denote a parameter of interest. Suppose our loss function is of the form*

$$L(\theta_1,\theta_2) = \sum_{j=1}^m L_j(\theta_1,\theta_2) := \sum_{j=1}^m g(d_j(\theta_1,\theta_2))$$

*for $\theta_1, \theta_2 \in \Theta$, where $d_1, \dots, d_m$ are pseudo-metrics on $\Theta$ satisfying*

$$d_j(\theta(P_\phi), \theta(P_{\phi'})) \geq \alpha_j$$

*whenever $\phi \sim_j \phi'$ and where $g$ is an increasing function with $g(x+y) \leq A\{g(x)+g(y)\}$ for all $x, y \in [0,\infty)$ and some $A > 0$. Then*

$$\inf_{\hat{\theta}\in\hat{\Theta}} \max_{\phi\in\Phi} E_\phi L(\hat{\theta}, \theta(P_\phi)) \geq \frac{1}{2A}\left\{1 - \max_{\substack{\phi,\phi'\in\Phi \\ \phi\sim\phi'}} \mathrm{TV}(P_\phi, P_{\phi'})\right\} \sum_{j=1}^n g(\alpha_j)$$

*where $\hat{\Theta}$ denotes the set of estimators of $\theta(P)$.*

**Remark.** In the common situation where $g(x) = x^q$ for some $q > 0$, we can take $A = \max(2^{q-1}, 1)$. In that special case, Assouad's lemma can be regarded as a slight strengthening of Le Can's two-point lemma, with the lower bound being larger by a factor of $\min(2^q, 2)$.

*Proof.* For $j \in [m]$ and $\phi \in \Phi$, write $\phi^j$ for the unique element of $\Phi$ for which $\phi \sim_j \phi^j$. Then for any $\hat{\theta} \in \hat{\Theta}$,

$$\begin{aligned}\max_{\phi\in\Phi} E_\phi L(\hat{\theta}, \theta(P_\phi)) &= \max_{\phi\in\Phi} \sum_{j=1}^m E_\phi L_j(\hat{\theta}, \theta(P_\phi)) \\ &\geq \frac{1}{2^m} \sum_{\phi\in\Phi} \sum_{j=1}^m E_\phi L_j(\hat{\theta}, \theta(P_\phi)) \\ &= \frac{1}{2^{m+1}} \sum_{j=1}^m \sum_{\phi\in\Phi} \{E_\phi L_j(\hat{\theta}, \theta(P_\phi)) + E_\phi L_j(\hat{\theta}, \theta(P_{\phi^j}))\}\end{aligned}$$

where equality holds since we have double counted each term and divided by 2. Now

$$\begin{aligned}L_j(\hat{\theta}, \theta(P_\phi)) + L_j(\hat{\theta}, \theta(P_{\phi^j})) &\geq \frac{1}{A} g(d_j(\hat{\theta}, \theta(P_\phi)) + d_j(\hat{\theta}, \theta(P_{\phi^j}))) \\ &\geq \frac{1}{A} g(d_j(\theta(P_\phi), \theta(P_{\phi^j}))) \\ &\geq \frac{g(\alpha_j)}{A}.\end{aligned}$$

Hence, writing $\mathcal{F}$ for the set of measurable functions from $\mathcal{X}$ to $[0, 1]$,

$$\begin{aligned}&E_\phi L_j(\hat{\theta}, \theta(P_\phi)) + E_{\phi^j} L_j(\hat{\theta}, \theta(P_{\phi^j})) \\ &\geq \frac{g(\alpha_j)}{A} \inf_{\substack{f_1, f_2 \in \mathcal{F} \\ f_1 + f_2 = 1}} \left[E_\phi f_1 + E_{\phi^j} f_2\right] \\ &= \frac{g(\alpha_j)}{A} \{1 - \mathrm{TV}(P_\phi, P_{\phi^j})\}.\end{aligned}$$

Here, the equality follows because if $g_\phi, g_{\phi'}$ denote densities of $P_\phi, P_{\phi'}$ with respect to a $\sigma$-finite dominating measure $\mu$, then writing $B := \{x \in \mathcal{X} : g_\phi(x) > g_{\phi'}(x)\}$,

$$\begin{aligned}\sup_{f\in\mathcal{F}} \int_{\mathcal{X}} f \mathrm{d}(P_\phi - P_{\phi'}) &= \sup_{f\in\mathcal{F}} \int_{\mathcal{X}} f(g_\phi - g_{\phi'}) \mathrm{d}\mu \\ &= \int_B (g_\phi - g_{\phi'}) \mathrm{d}\mu \\ &= \frac{1}{2} \int_{\mathcal{X}} |g_\phi - g_{\phi'}| \mathrm{d}\mu \\ &= \mathrm{TV}(P_\phi, P_{\phi'}).\end{aligned}$$

We conclude that

$$\begin{aligned}\inf_{\hat{\theta}\in\hat{\Theta}}\max_{\phi\in\Phi} E_\phi L(\hat{\theta},\theta(P_\phi)) &\geq \frac{1}{2^{m+1}}\sum_{\phi\in\Phi}\sum_{j=1}^{m}\frac{g(\alpha_j)}{A}\mathrm{TV}(P_\phi,P_{\phi^j}) \\ &\geq \frac{1}{2A}\left\{1-\max_{\substack{\phi,\phi'\in\Phi\\ \phi\sim\phi'}}\mathrm{TV}(P_\phi,P_{\phi'})\right\}\sum_{j=1}^{m}g(\alpha_j)\end{aligned}$$

as required. □