Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 1

1. For $n \in \mathbb{N}$ let $X_1, \dots, X_n$ be independent and identically distributed random variables on the measurable space $(\mathcal{X}, \mathcal{A})$. Furthermore, let $\mathcal{F}$ be a set of real valued functions on $(\mathcal{X}, \mathcal{A})$ such that $\mathbb{E}[|f(X_1)|] < \infty$ for all $f \in \mathcal{F}$.

   For functions $l, u \in L^1(\mathcal{X})$ we define the *bracket*

   $$[l, u] := \{f : \mathcal{X} \to \mathbb{R} : l(x) \leqslant f(x) \leqslant u(x) \text{ for all } x \in \mathcal{X}\}.$$

   If $\mathbb{E}[|u(X_1) - l(X_1)|] \leqslant \varepsilon$, we call $[l, u]$ an *ε-bracket*.

   Prove: If for any $\varepsilon > 0$ the class $\mathcal{F}$ can be covered by finitely many ε-brackets, then the *Glivenko–Cantelli* property holds:

   $$\sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{k=1}^{n} f(X_k) - \mathbb{E}[f(X_k)] \Big| \xrightarrow{a.s.} 0.$$

2. Apply exercise 1 to show the *Glivenko–Cantelli theorem*:

   The almost sure convergence of the empirical distribution function to the true distribution function holds uniformly in $\mathbb{R}$, that is for $n \to \infty$

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

3. For a kernel function $K \in L^1(\mathbb{R}^d)$ and any bounded real valued function $f : \mathbb{R}^d \to \mathbb{R}$ prove the following:

   (a) For any $x \in \mathbb{R}^d$ the convolution $(K_h * f)(x)$ is well-defined.
   (b) If f is continuous at x, then $(K_h * f - f)(x) \to 0$ as $h \to 0$.
   (c) The continuity assumption is necessary (by example).

4. Show: For any $m \in \mathbb{N}$ there is a unique polynomial P_m with degree smaller or equal to m such that

   $$K(x) := \begin{cases} P_m(x), & \text{for } x \in [-1, 1], \\ 0, & \text{otherwise} \end{cases}$$

   is a square-integrable kernel on $\mathbb{R}$ of order m.

Submit before the lecture on Thursday, April 25.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 2

1. Let $X_1, \dots, X_n$ be i.i.d. with Lebesgue-density $f : \mathbb{R} \to [0, \infty)$. Define the sinc kernel by $K(x) := \mathcal{F}\left[\frac{1}{2\pi}\mathbf{1}_{[-1,1]}\right](x)$.

   (a) Show that a continuous version of K is $K(x) = \frac{\sin(x)}{\pi x}$ $x \neq 0$, $K(0) = \pi^{-1}$ and that $K \in L^2(\mathbb{R}) \backslash L^1(\mathbb{R})$.

   (b) The kernel density estimator satisfies $\mathcal{F}\hat{f}_{n,h}(u) = \hat{\varphi}_n(u)\mathbf{1}_{[-h^{-1},h^{-1}]}(u)$, where $\hat{\varphi}_n(u) = n^{-1}\sum_{j=1}^n e^{iuX_j}$.
   Note: $\mathcal{F}\mathcal{F}f(x) = 2\pi f(-x)$ *in dimension* $d = 1$.

   (c) If $f \in H^s(\mathbb{R})$ for arbitrary $s > 0$, then

$$\mathrm{MISE}_{\mathbb{R}}(\hat{f}_{n,h}) \lesssim \|f\|_{H^s}^2 h^{2s} + \frac{1}{nh}.$$

2. Let $X_1, \dots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_1] = 0$ and $|X_1| \leq S$ a.s. Set $Y_n := \sum_{i=1}^n X_i$ and let $t > 0$.

   (a) Show for any $\alpha > 0$

$$P(Y_n \geq t) \leq e^{-\alpha t}\left(\mathbb{E}\left[e^{\alpha X_1}\right]\right)^n.$$

   (b) Use a series representation of e^x along with $(1 + A)^n \leq e^{An}$ to derive

$$P(Y_n \geq t) \leq \exp\Big(-\alpha t + \mathrm{Var}(Y_n)S^{-2}(e^{tS} - 1 - \alpha S)\Big).$$

   (c) Conclude that the absolute value of Y_n satisfies

$$P(|Y_n| \geq t) \leq 2\exp\left(-\frac{t^2}{4(\mathrm{Var}(Y_n) + tS)}\right).$$

3. Let $X_1, \dots, X_n$ be random variables, satisfying $\max_{1\leq k\leq n}\mathbb{E}[\exp(aX_k^2)] \leq C$ for constants $a, C > 0$. Use Jensen's inequality to prove

$$\mathbb{E}\left[\max_{1\leq k\leq n} X_k^2\right] \leq \frac{1}{a}\log(Cn).$$

4. Consider the orthonormal basis $(\varphi_j)_{j\geq 1}$ of $L^2([0,1])$, where $\varphi_1(x) = 1$ and

$$\varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx), \quad \varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx), \quad k \in \mathbb{N}\,.$$

Let further $V_m = \operatorname{span}(\varphi_1, \ldots, \varphi_m)$ and Π_{V_m} be the orthogonal projection of $L^2([0,1])$ onto V_m. For $X_1, \ldots, X_n$ i.i.d. with Lebesgue-density $f \in L^2([0,1])$ define the projection estimator of f by

$$\hat{f}_{n,m}(x) := \sum_{j=1}^{m} \langle \varphi_j, \hat{\mu}_n \rangle \varphi_j(x) := \sum_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i) \right) \varphi_j(x), \quad x \in [0,1].$$

(a) Show the following decomposition:

$$\mathbb{E}\left[\|\hat{f}_{n,m} - f\|_{L^2}^2 \right] = \|f - \Pi_{V_m} f\|_{L^2}^2 + \frac{1}{n} \sum_{j=1}^{m} \int_0^1 \varphi_j^2(x) f(x)\,dx - \frac{1}{n} \|\Pi_{V_m} f\|_{L^2}^2.$$

(b) For $\alpha \in \mathbb{N}$ and $L > 0$ let $H^\alpha_{\mathrm{per}}([0,1], L)$ consist of all $g \in C^{\alpha-1}([0,1]) \cap L^2([0,1])$ with $g^{(j)}(0) = g^{(j)}(1)$, $j = 0, \ldots, \alpha - 1$, and such that $g^{(\alpha-1)}$ has a weak derivative $g^{(\alpha)}$ with $\|g^{(\alpha)}\|_{L^2} \leq L$.
Show $H^\alpha_{\mathrm{per}}([0,1], L) \subseteq \mathcal{W}^\alpha([0,1], L)$, where

$$\mathcal{W}^\alpha([0,1], L) := \left\{ g \in L^2([0,1]) : \sum_{j \geq 1} a_j^2 \langle \varphi_j, g \rangle^2 \leq \frac{L^2}{\pi^{2\alpha}} \right\}$$

with $a_j = \mathbf{1}(j \text{ even}) j^\alpha + \mathbf{1}(j \text{ odd})(j-1)^\alpha$.

(c) Conclude that any $f \in \mathcal{W}^\alpha([0,1], L)$ satisfies

$$\|f - \Pi_{V_m} f\|_{L^2}^2 \leq \frac{L^2}{\pi^{2\alpha}} m^{-2\alpha}$$

and

$$\mathbb{E}\left[\|\hat{f}_{n,m} - f\|_{L^2}^2 \right] \lesssim L^2 m^{-2\alpha} + \frac{m}{n}.$$

5. For $\alpha \in (0,1]$ and $L, R > 0$ let $\mathcal{H}_{[0,1]}(\alpha; L, R)$ be a Hölder ball. Consider for $f \in \mathcal{H}_{[0,1]}(\alpha; L, R)$ the following observation models.

(i) *Gaussian white noise model:* For a Brownian motion $W = (W_t)_{t \in [0,1]}$ observe

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}} dW_t, \quad t \in [0,1], \tag{$\mathcal{E}_{1,n}$}$$

which means that the Gaussian process $(Y_{1,g})_{g \in L^2}$ is observed, where $Y_{1,g} := \int_0^1 g(t) dY_t \sim \mathcal{N}(\langle f, g \rangle, \frac{\sigma^2}{n} \|g\|_{L^2}^2)$ and $\operatorname{Cov}(Y_{1,g}, Y_{1,h}) = \frac{\sigma^2}{n} \langle g, h \rangle$.

(ii) *Nonparametric regression model:* For $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d. observe

$$Y_{2,i} = f\Big(\frac{i}{n}\Big) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{$\mathcal{E}_{2,n}$}$$

For $I_i := \left[\frac{i-1}{n}, \frac{i}{n}\right)$ construct the random variables $\widetilde{Y}_{2,i} := nY_{1,\mathbf{1}_{I_i}},\ i = 1, \dots, n$.

(a) Show $\widetilde{Y}_{2,i} \sim \mathcal{N}\left(n\int_{I_i} f(t)dt, \sigma^2\right)$, $i = 1, \dots, n$, and $\mathrm{Cov}(\widetilde{Y}_{2,i}, \widetilde{Y}_{2,j}) = \sigma^2\delta_{ij}$.

(b) Prove that the Kullback-Leibler distance under f satisfies

$$\mathrm{KL}_f\left(\mathcal{L}(Y_{2,i}), \mathcal{L}\left(\widetilde{Y}_{2,i}\right)\right) = \frac{1}{2\sigma^2}\left(f\left(\frac{i}{n}\right) - n\int_{I_i} f(t)dt\right)^2, \quad i = 1, \dots, n.$$

(c) Denote by $\mathrm{TV}(\bullet, \bullet)$ the total variation distance. Give sufficient conditions on α, L, R such that $\mu_n := \mathcal{L}((Y_{2,i})_{1\le i\le n})$ and $\tilde{\mu}_n := \mathcal{L}((\widetilde{Y}_{2,i})_{1\le i\le n})$ obey

$$\sup_{f\in\mathcal{H}(\alpha;L,R)}\left(\mathrm{KL}_f(\mu_n, \tilde{\mu}_n) + \mathrm{TV}_f(\mu_n, \tilde{\mu}_n)\right) \to 0, \quad \text{as } n \to \infty.$$

Hint: Use results on KL and TV from the book 'Methoden der Statistik' or 'Introduction to Nonparametric Estimation' by A. Tsybakov (online available).

6. Consider the models $\mathcal{E}_{1,n}$ and $\mathcal{E}_{2,n}$ from Exercise 5 and let $\vartheta := \int_0^1 f(t)w(t)dt$, for $w \in L^2([0,1])$.

(a) Show that the estimator $\hat{\vartheta}_{1,n} := \int_0^1 w(t)dY_t$ in $\mathcal{E}_{1,n}$ obeys

$$\sqrt{n}(\hat{\vartheta}_{1,n} - \vartheta) \sim \mathcal{N}\left(0, \sigma^2\|w\|_{L^2}^2\right).$$

(b) Find an estimator $\hat{\vartheta}_{2,n}$ in $\mathcal{E}_{2,n}$ and regularity assumptions on w and f such that

$$\sqrt{n}(\hat{\vartheta}_{2,n} - \vartheta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\|w\|_{L^2}^2\right), \quad \text{as } n \to \infty.$$

Is there an estimator $\hat{\vartheta}_{2,n}$ which has these asymptotics for any $w \in L^2([0,1])$?

Problems 1 & 2 are bonus exercises due before the lecture on Monday, April 29.
Problems 3 - 6 have to be submitted before the lecture on Monday, May 6.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 3

1. For $f \in L^2([0,1])$ consider the signal-in-white-noise model

$$dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}} dW_t, \quad t \in [0,1]. \tag{0.1}$$

Let $K \in L^2(\mathbb{R})$ be a kernel of order $\langle \alpha \rangle$ and of compact support. Set

$$\hat{f}^{\mathrm{NW}}_{n,h}(x) := \int_0^1 K_h(x-t)dY_t, \quad x \in (0,1).$$

(a) Show that for h sufficiently small

$$\mathbb{E}[(\hat{f}^{\mathrm{NW}}_{n,h}(x) - f(x))^2] = ((K_h * f - f)(x))^2 + \frac{\sigma^2 \|K\|^2_{L^2}}{nh}.$$

Compare the above risk to the pointwise risk in density estimation.

(b) For $f \in \mathscr{H}(\alpha; R, L)$ with $\alpha \in (0,1]$ find the asymptotically optimal bandwidth $h_n^* \to 0$, as $n \to \infty$, and derive

$$\mathbb{E}[(\hat{f}^{\mathrm{NW}}_{n,h}(x) - f(x))^2] \lesssim n^{-\frac{2\alpha}{2\alpha+1}}.$$

(c) Show that $h \to 0$, $nh \to \infty$ and $n^{\frac{1}{2\alpha+1}} h \to 0$ imply

$$\sqrt{nh}(\hat{f}^{\mathrm{NW}}_{n,h}(x) - f(x)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 \|K\|^2_{L^2})$$

and construct an asymptotic $(1-\gamma)$-confidence interval for $\gamma \in (0,1)$.

2. For f with derivative $f' \in \mathscr{H}_{[0,1]}(\alpha; R, L)$, $\alpha \in (0,1]$, consider the model (0.1) from Exercise 1. Let $K \in L^2(\mathbb{R})$ be a kernel with support in $[-1,1]$ and derivative $K' \in L^2(\mathbb{R})$.

(a) Show that the estimator $(\widehat{f'})^{\mathrm{NW}}_{n,h}(x) := \int_0^1 \frac{\partial}{\partial x} K_h(x-t)dY_t$ satisfies for $x \in (0,1)$ with $x > h$ and $x - 1 < -h$:

$$\mathbb{E}[((\widehat{f'})^{\mathrm{NW}}_{n,h}(x) - f'(x))^2] = ((K_h * f' - f')(x))^2 + \frac{\sigma^2 \|K'\|^2_{L^2}}{nh^3}.$$

(b) Find the asymptotically optimal bandwidth $h_n^* \to 0$, as $n \to \infty$, such that

$$\mathbb{E}[((\widehat{f'})^{\mathrm{NW}}_{n,h}(x) - f'(x))^2] \lesssim n^{-\frac{2\alpha}{(2\alpha+3)}}.$$

3. For $f \in \mathscr{H}_{[0,1]}(\alpha; R, L)$ consider the regression model

$$Y_i = f\Big(\frac{i}{n}\Big) + \varepsilon_i, \quad i = 1, \dots, n.$$

Let $\hat{f}_{n,h}$ be the local polynomial estimator of degree $\langle \alpha \rangle$ and suppose that the kernel K is Lipschitz continuous with support $[0, 1]$. Moreover, assume $\|B(x)^{-1}\|_2 \leq C$, where $C > 0$ and $\|\bullet\|_2$ is the Frobenius norm. Show that

$$\mathbb{E}[\|\hat{f}_{n,h} - f\|_\infty^2] \lesssim h^{2\alpha} + \frac{\log n}{nh}.$$

and determine the optimal rate.

Hint: Use Exercise 2.3 for the variance term.

4. Let $(X_i, Y_i)_{i=1,\dots,n}$ be observation in a nonparametric regression model.

   (a) Assuming $\sum_{i=1}^n K_h(x - X_i) > 0$, prove the following representation of the Nadaraya-Watson estimator

$$\hat{f}_{n,h}^{NW}(x) = \operatorname{argmin}_{y \in \mathbb{R}} \Big(\sum_{i=1}^n (y - Y_i)^2 K_h(x - X_i) \Big).$$

   (b) Let K be the rectangular kernel. Which estimator is obtained by minimising $\sum_{i=1}^n |y - Y_i| K_h(x - X_i)$?

Submit before the lecture on Monday, May 13.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 4

1. In a regression model with deterministic design let $\hat{f}_{n,K}$ be the projection estimator of $f \in L^2([0,1])$ based on $\varphi_1, \dots, \varphi_K \in L^2([0,1])$. Denote by $\|\bullet\|_n$ the empirical norm and assume that $\sigma^2 := \mathbb{E}[\varepsilon_i^2] < \infty$. Show:

   (a) If Π_K denotes the orthogonal projection onto $\mathrm{span}(\varphi_1, \dots, \varphi_K)$ then

   $$\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] = \|f - \Pi_K f\|_n^2 + \frac{\sigma^2 K}{n}.$$

   Discuss the dependence on K and n. What happens in case of $K = n$?

   (b) For random design with $\sigma^2(x) := \mathrm{Var}_f(Y_i | X_i = x_i) < \infty$ P^{X_i}-a.s. it holds

   $$\mathbb{E}_f[\|\hat{f}_{n,K} - f\|_n^2] = \mathbb{E}_f[\|f - \Pi_K f\|_n^2] + \frac{1}{n}\sum_{k=1}^{K} \mathbb{E}_f[\sigma^2(X_i)\varphi_k^2(X_i)].$$

2. Show that the following sets are approximation spaces in $L^2([0,1])$ and give their order of approximation.

   (a) $V_K = \{\varphi : [0,1] \to \mathbb{R} \,|\, \varphi = \sum_{m=1}^{K} c_m \mathbf{1}_{[(m-1)/K, m/K)}, c \in \mathbb{R}^K\}$.

   (b) $W_K = \{\varphi \in L^2([0,1]) : \varphi|_{[(m-1)/K, m/K)} \in \mathrm{Pol}_M, k = 1, \dots, M\}$.

3. Let $a < x_1 < \dots < x_n < b$, $n \geq 2$, and denote by S_n the space of natural cubic splines on $[a,b]$ with knots at $x_1, \dots, x_n$, i.e. any $g \in S_n$ lies in $C^2([a,b])$, is a polynomial of degree three on each interval $[x_i, x_{i+1}]$ and is linear on the boundary intervals $[a, x_1]$, $[x_n, b]$.

   (a) Let $g \in S_n$ and consider any $f \in C^2([a,b])$ with $g(x_i) = f(x_i)$. Show $\int_a^b f''(x)^2 dx \geq \int_a^b g''(x)^2 dx$ with equality if and only if $f = g$.
   Hint: Use integration by parts to prove $\int_a^b g''(x)(f''(x) - g''(x))dx = 0$.

   (b) For $\lambda > 0$ and $y \in \mathbb{R}^n$ consider the penalized least squares problem

   $$\min_{f \in C^2([a,b])} \Big(\sum_{i=1}^{n} (y_i - f(xi))^2 + \lambda \int_a^b f''(x)^2 dx \Big).$$

   Use (a) to argue that the minimiser must be a natural cubic spline. Determine the minimiser.

4. In the usual i.i.d. regression model $(x_i, Y_i)_{i=1,\dots,n}$ let $(\varphi_j)_{j=1,\dots,n}$ be an orthonormal system with respect to the empirical scalar product. We consider the Lasso estimator in the special case where the dictionary is given by (φ_j). Hence, we define for $p \in \{0,1\}$ and $\lambda > 0$

$$\hat{\beta}_{n,p} := \arg\min_{\beta \in \mathbb{R}^n} \left\{ \Big\| Y - \sum_{j=1}^n \beta_j \varphi_j \Big\|_n^2 + \lambda |\beta|_{\ell^p} \right\}$$

with $|\beta|_{\ell^1} := \sum_{j=1}^n |\beta_j|$ and $|\beta|_{\ell^0} := \sum_{j=1}^n \mathbf{1}_{\{\beta_j \neq 0\}}$.

(a) Show for $p = 1$

$$\hat{\beta}_{n,1} = \Big(\big(\langle Y, \varphi_j \rangle_n - \lambda/2 \big)_+ - \big(\langle Y, \varphi_j \rangle_n + \lambda/2 \big)_- \Big)_{j=1,\dots,n},$$

denoting $a_+ := \max\{0, a\}$ and $a_- := |a| - a_+$ for $a \in \mathbb{R}$. The estimator $\hat{\beta}_{n,1}$ is called *soft-thresholding* estimator.

(b) For $p = 0$ verify

$$\hat{\beta}_{n,0} = \Big(\langle Y, \varphi_j \rangle_n \mathbf{1}_{\{|\langle Y, \varphi_j \rangle_n| \geqslant \sqrt{\lambda}\}} \Big)_{j=1,\dots,n},$$

which is referred to as *hard-thresholding* estimator.

Submit before the lecture on Monday, May 20.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 5

1. In the regression model $Y_i = f(x_i) + \varepsilon_i, i = 1, \dots, n$, let ε_i be i.i.d. and symmetric with $\sigma^2 := \mathrm{Var}(\varepsilon_1) < \infty$ and $Q := \mathbb{E}[|\varepsilon_1|^q] < \infty$ for some $q > 2$. Moreover, let $(\psi_j)_{j \geqslant 1} \subseteq L^2(D)$ satisfy $\|\psi_j\|_n = 1$ and $\tau := \sup_j \|\psi_j\|_\infty < \infty$. Supposing $p \to \infty$ and $n^{2-q}(\log p)^q \to 0$ as $n \to \infty$, we prove for $\kappa > \sqrt{8}$

$$\lim_{n\to\infty} \mathbb{P}\Big(\max_{j=1,\dots,p} |\langle \varepsilon, \psi_j \rangle_n| > \kappa\sigma \sqrt{\frac{\log p}{n}} \Big) = 0. \tag{0.2}$$

   Proceed as follows:

   (a) Assume $|\varepsilon_1| \leq R$, apply Bernstein's inequality and prove that

$$\mathbb{P}\Big(\max_{1\leq j\leq p} |\langle \varepsilon, \psi_j \rangle_n| > \kappa\sigma \sqrt{\frac{\log p}{n}} \Big) \leq 2p \exp\Big(-\frac{\kappa^2 \log(p)}{4(1 + \kappa\sqrt{\log(p)/(n\sigma^2)}\tau R)} \Big).$$

   (b) Drop the assumption $|\varepsilon_1| \leq R$, set $\tilde{\varepsilon}_i := \varepsilon_i \mathbf{1}_{|\varepsilon_i| \leq R}$, $i = 1, \dots, n$, and show:

$$\mathbb{P}(\varepsilon \neq \tilde{\varepsilon}) \leq nQR^{-q}.$$

   (c) Chose R appropriately to obtain (0.2). How does the statement changes if $\tau_p := \sup_{j\leq p} \|\psi_j\|_\infty \to \infty$?

2. In the regression model $Y_i = f(x_i) + \varepsilon_i, i = 1, \dots, n$, with ε_i i.i.d., $\mathbb{E}[\varepsilon_1] = 0$ and $\sigma^2 := \mathbb{E}[\varepsilon_1^2] < \infty$, let $(\psi_j)_{j=1,\dots,p}$ be an orthonormal system with respect to $\langle \bullet, \bullet \rangle_n$. In the *ridge regression* we estimate f by $\hat{f}^{\mathrm{ridge}} := \sum_{j=1}^p \hat{\beta}_{n,j}^{ridge} \psi_j$ where

$$\hat{\beta}_n^{\mathrm{ridge}} := \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \Big\| Y - \sum_{j=1}^p \beta_j \psi_j \Big\|_n^2 + \lambda |\beta|_{\ell^2}^2 \Big\}$$

   for some $\lambda > 0$.

   (a) Prove

$$\hat{\beta}_n^{\mathrm{ridge}} = \Big(\frac{\langle Y, \psi_j \rangle_n}{1 + \lambda} \Big)_{j=1,\dots,p}.$$

   (b) Decompose $\mathbb{E}[\|\hat{f}^{\mathrm{ridge}} - f\|_n^2]$ into bias and variance and determine both errors in terms of λ and $\beta_j = \langle f, \psi_j \rangle_n$.

3. Let $\mathbb{P}, \mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{F})$ possessing densities p and q, respectively, with respect to some dominating measure μ. Prove for the total variation distance $\| \mathbb{P} - \mathbb{Q} \|_{\mathrm{TV}} := \sup_{A \in \mathscr{F}} | \mathbb{P}(A) - \mathbb{Q}(A)|$ the following:

   (a) $\| \mathbb{P} - \mathbb{Q} \|_{\mathrm{TV}} = \frac{1}{2} \int_{\mathcal{X}} |p - q|(x) \mu(\mathrm{d}x) = \int_{\mathcal{X}} \max(p(x), q(x)) \mu(\mathrm{d}x) - 1.$

   (b) It holds that $\mathrm{KL}(\mathbb{P} \,|\, \mathbb{Q}) \leqslant \chi^2(\mathbb{P}, \mathbb{Q})$, where

$$\chi^2(\mathbb{P}, \mathbb{Q}) := \begin{cases} \int \left(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1 \right)^2 d\mathbb{Q}, & \text{if } \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{else.} \end{cases}$$

   (c) For $\mathbb{P} \sim \mathcal{N}(\mu, \sigma^2)$ and $\mathbb{Q} \sim \mathcal{N}(\tilde{\mu}, \sigma^2)$ calculate $\| \mathbb{P} - \mathbb{Q} \|_{\mathrm{TV}}$, $\mathrm{KL}(\mathbb{P} \,|\, \mathbb{Q})$ and $\chi^2(\mathbb{P}, \mathbb{Q})$.

4. Let $\mathbb{P}, \mathbb{Q}$ be probability measures on $(\mathcal{X}, \mathscr{F})$ possessing densities p and q, respectively, with respect to some dominating measure μ.

   (a) Prove that the sum of type I and II error satisfies

$$\inf_{\psi} \big(\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \big) = \big(1 - \| \mathbb{P}_0 - \mathbb{P}_1 \|_{\mathrm{TV}} \big),$$

   where the infimum is taken over all tests, i.e. measurable functions $\psi : \mathcal{X} \to \{0, 1\}$, of the hypothesis $H_0 : \mathbb{P} = \mathbb{P}_0$ versus the alternative $H_1 : \mathbb{P} = \mathbb{P}_1$.

   (b) Construct a test that attains the infimum in (a).

   (c) Give an example such that the sum of type I and II error equals zero and one, respectively.

5.^{PE} Practical exercise (Voluntary):

Let the regression function be given by

$$f: \quad [0, 1] \to \mathbb{R}, \quad x \mapsto 5 \sin \Big(\frac{4\pi}{x + 1} \Big).$$

Generate a sample

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

using random design points $X_i \sim \mathcal{U}([0, 1])$ and standard normal errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n$. Implement the soft-thresholding and the hard-thresholding estimator based on the Haar basis. Discuss the loss of each approach for varying sample sizes n, penalty parameters λ and noise levels σ.

Submit before the lecture on Monday, May 27.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 6

1. Consider the regression model on the interval $[0,1]$ with deterministic design $x_i = \frac{i}{n+1}, i = 1,\dots,n$, and with standard normal errors. Prove for $\alpha, L > 0$ and $x_0 \in (0,1)$ the lower bound

$$\liminf_{n\to\infty} \inf_{\hat{f}_n} \sup_{f\in\mathscr{H}_{[0,1]}(\alpha;L)} n^{2\alpha/(2\alpha+1)}\, \mathbb{E}_f\left[(\hat{f}_n(x_0) - f(x_0))^2\right] > 0.$$

   Proceed as follows: Let $h = h_0 n^{-\frac{1}{2\alpha+1}}$ for some $h_0 > 0$ and let $\varphi \in C^\infty$ be a function with support in $[-1,1]$ and with $\varphi(0) > 0$. Show:

   (a) For $f_{0,n} := 0$ and $f_{1,n} := \gamma_0 h^\alpha \varphi(h^{-1}(\bullet - x_0))$, for some $\gamma_0 \in \mathbb{R}$, it holds

$$|f_{0,n}(x_0) - f_{1,n}(x_0)| \geq C n^{-\frac{\alpha}{2\alpha+1}},$$

   where $C > 0$.

   (b) For some $n_0 \in \mathbb{N}$ and a suitable choice of h_0 it holds that

$$\mathrm{KL}(P_{f_{0,n}} | P_{f_{1,n}}) \leq 1, \quad \forall n \geq n_0.$$

   (c) Apply the results from the lecture to obtain the lower bound.

   Bonus: Give an example of φ as above.

2. Consider the regression model from Exercise 1. Prove the lower bound

$$\liminf_{n\to\infty} \inf_{\widehat{(f')}_n} \sup_{f\in C^1 : f'\in\mathscr{H}_{[0,1]}(\alpha;L)} n^{2\alpha/(2\alpha+3)}\, \mathbb{E}_f\left[(\widehat{(f')}_n(x_0) - f'(x_0))^2\right] > 0.$$

   You are alternatively allowed to work in the signal-in-white-noise model, where Girsanov's theorem gives the likelihood/density.

Submit before the lecture on Monday, June 3.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 7

1. For $m \geq 8$ let $E := \{0,1\}^m$ and set $\rho(\varepsilon, \tilde{\varepsilon}) := |\varepsilon - \tilde{\varepsilon}|_{\ell_0}$, $\varepsilon, \tilde{\varepsilon} \in E$. Prove that there is a subset $\{\varepsilon^{(0)}, \dots, \varepsilon(J)\} \subseteq E$ with $J \geq 2^{m/8}$ and $\varepsilon^{(0)} = (0, \dots, 0)$, such that $\rho(\varepsilon^{(k)}, \varepsilon^{(l)}) \geq m/8$, for any $k \neq l$. Proceed as follows:

   (a) Set $D := \lfloor m/8 \rfloor$ and let $E_1 := \{\varepsilon \in E | \rho(\varepsilon, \varepsilon^{(0)}) > D\}$. Set further iteratively $E_j := \{\varepsilon \in E_{j-1} | \rho(\varepsilon, \varepsilon^{(j-1)}) > D\}$, where $\varepsilon^{(j-1)} \in E_{j-1}$ arbitrary. Denote by J the last index j such that $E_j \neq \emptyset$. Show that $\rho(\varepsilon^{(k)}, \varepsilon^{(l)}) \geq m/8$, for any $k \neq l$.

   (b) Verify the bound $|E_j \backslash E_{j+1}| \leq \sum_{i=0}^{D} \binom{m}{i}$ and conclude

$$(J+1) \sum_{i=0}^{D} \binom{m}{i} \geq 2^m \quad \text{and} \quad \sum_{i=0}^{D} \binom{m}{i} 2^{-m} \geq (J+1)^{-1}.$$

   (c) Find the statement of Hoeffding's inequality and apply it to $P(S_m \leq D)$, where $S_m \sim \mathrm{Bin}(m, 1/2)$, to derive $J \geq 2^{m/8}$.

2. In the lecture the minimax optimal rate $n^{-\alpha/(2\alpha+1)}$ for pointwise loss was derived in the density estimation problem with $f \in \mathscr{H}_D(\alpha; R, L), D \subseteq \mathbb{R}$. Show that the same alternative, i.e. $f_h(x) = f(x) + \gamma_0 h^\alpha \varphi(h^{-1}(x - x_0))$, yields only the suboptimal rate $n^{-1/2}$ as lower bound for the MISE, given that $f(x_0) > 0$.

3. Consider the regression model on the interval $[0,1]$ with deterministic design $x_i = \frac{i}{n+1}, i = 1, \dots, n$, and with standard normal errors. Prove for $\alpha, L > 0$ the lower bound

$$\liminf_{n\to\infty} \inf_{\hat{f}_n} \sup_{f \in \mathscr{H}_{[0,1]}(\alpha;L)} \Big(\frac{n}{\log n}\Big)^{2\alpha/(2\alpha+1)} \mathbb{E}_f \left[\|\hat{f}_n - f\|^2_{L^\infty([0,1])} \right] > 0.$$

   Hint: Let $\gamma_0 > 0$ be constant, $h_n = M_n^{-1}$ for some integer $M_n > 1$ and $\xi_j = \frac{j-1/2}{M_n}, j = 1, \dots, M_n$. Consider the alternatives

$$f_{0,n} \equiv 0 \quad \text{and} \quad f_{j,n}(x) = \gamma_0 h_n^\alpha \varphi\Big(\frac{x - \xi_j}{h_n}\Big), j = 1, \dots, M_n,$$

   with a suitable regular function φ satisfying $\varphi(x) > 0$ if and only if $x \in (-1/2, 1/2)$.

4. For $\alpha \in (0,1], L > 0$ and standard normal random variables (ε_i) we consider the regression model

$$Y_i = f(x_i) + \sigma\varepsilon_i, \quad i = 1, \dots, n, \quad \text{for } f \in \mathscr{H}_{[0,1]}(\alpha; L)$$

with deterministic design points $x_i = \frac{i}{n+1}$ and unknown noise level $\sigma > 0$. Let $\hat{f}_{n,h}$ be the Nadaraja–Watson estimator with continuous and compact supported kernel and with bandwidth $h > 0$. To estimate σ^2, we define

$$\hat{\sigma}^2 := \|Y - \hat{f}_{n,h}\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \hat{f}_{n,h}(x_i)\big)^2.$$

Using the estimate for the risk of $\hat{f}_{n,h}$, verify

$$\mathbb{E}\left[|\hat{\sigma}^2 - \sigma^2|\right] \leqslant C\big(n^{-1/2} + (h^{2\alpha} + (nh)^{-1})\big)$$

for some constant $C > 0$. Which convergence rate can be achieved by an optimal choice of h?

Submit before the lecture on Monday, June 17.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 8

1. In the lecture an outline for the proof of the inequality

$$\mathbb{P}(Z - p \geq 2\sqrt{px} + 2x) \leq e^{-x}, \quad x \geq 0,$$

was given, where $Z \sim \chi^2_p$. Carry out the single steps precisely and write down a complete proof.

2. In the Gaussian sequence space model $(y_k)_{k\geq 1}$ the quantity $\mathcal{I}(K) := -\sum_{k=1}^K y_k^2 + 2K\frac{\sigma^2}{n}$ was introduced in the lecture. Consider the function

$$\psi(K) := \mathcal{I}(K) - \mathcal{I}(K^*), \quad K \in \mathbb{N},$$

for some $K^* \in \mathbb{N}$. Calculate the mean and the standard deviation of $\psi(K)$.

3. Let $\hat{f}_{n,h}$ be a kernel density estimator based on a continuous kernel K with support in $[-1,1]$ and $2K(0) > \|K\|^2_{L^2}$. Introduce the *cross-validation criterion*

$$CV_n(h) := \int_{R^d} \hat{f}_{n,h}(x)^2 dx - 2\hat{I}_n(h),$$

where $\hat{I}_n(h) := \frac{1}{n}\sum_{j=1}^n \hat{f}^{(-j)}_{n,h}(X_j)$ and $\hat{f}^{(-j)}_{n,h} := \frac{1}{n-1}\sum_{i\neq j} K_h(x - X_i)$. Show

$$\operatorname{argmin}_{h>0} \mathbb{E}_f[CV_n(h)] = \operatorname{argmin}_{h>0} \mathbb{E}_f[\|\hat{f}_{n,h} - f\|^2_{L^2}].$$

4. In the set-up of Exercise 3, why does the naive approach to replace $\hat{I}_n(h)$ by

$$\hat{I}^{naive}_n(h) := \frac{1}{n}\sum_{j=1}^n \hat{f}_{n,h}(X_j)$$

lead to undersmoothing? To this end study $CV^{naive}_n(h) := \int_{\mathbb{R}} \hat{f}_{n,h}(x)^2 dx - 2\hat{I}^{naive}_n(h)$ for $h \to 0$.

Submit before the lecture on Monday, June 24.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 9

1. Consider the Gaussian sequence space model

$$y_k = f_k + \sigma\zeta_k, \quad k \geqslant 1,$$

and let $K_{\max} \in \mathbb{N}$. Denote by S_I the subspace of sequences $(x_k)_{k\geq 1}$ such that $x_k = 0$, $\forall k \in \mathbb{N}\setminus I$, where $I \subseteq \{1,\dots,K_{\max}\}$, and let Π_{S_I} denote the orthgonal projection onto S_I. Show that the hard thresholding estimator satisfies

$$\hat{f}^{(\mathrm{hard})} = \Pi_{S_{\tilde{I}}} y, \text{ where } \tilde{I} = \mathrm{argmin}_{I\subseteq\{1,\dots,K_{max}\}} \Big(-\|\Pi_{S_I} y\|^2 + \tau \dim(S_I)\Big),$$

for some suitable $\tau > 0$.

2. In the setting of Exercise 1 apply the general oracle inequality for $\mathbb{E}[\|\hat{f} - f\|^2]$, derived for projection estimators with a penalised least squares criterion. Compare this bound with the specific oracle inequality derived for hard thresholding estimators.

3. Consider the function $f(x) = |x - 1/3|$, $x \in [0,1]$. Calculate the Haar wavelet coefficients $\langle f, \psi_{jk}\rangle$ and decide for which $s > 0$ and $p, q \in [1,\infty]$ it holds that

$$|f|_{s,p,q} := \Big\| \Big(2^{j(s+1/2-1/p)} \|(\langle f, \psi_{jk}\rangle)_k\|_{\ell^p} \Big)_{j\geq 0} \Big\|_{\ell^q} < \infty.$$

4. Given a triplet (s,p,q) find out for which triplets (s',p',q') we have the bound

$$|f|_{s',p',q'} \lesssim |f|_{s,p,q}$$

for all functions f with $|f|_{s,p,q} < \infty$.
Remark: This is a general form of the Sobolev embedding theorem.

Submit before the lecture on Monday, July 1.

Course *Nonparametric Statistics*
Summer term 2019
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß
Sebastian Holtz

Exercise sheet 10

1. Motivate and derive the kernel density estimation approach. Give examples of possible kernel choices.

2. For a density $f : \mathbb{R}^d \to \mathbb{R}$ and a kernel density estimator of f consider the MISE. What is meant by 'bias-variance tradeoff'? Give an upper bound of the MISE for an estimator of your choice and discuss the role of the dimension d.

3. State the nonparametric regression model with deterministic design and Gaussian errors on $[0, 1]$. Write down two related models and explain briefly the connection between the three models. Give an example in which the usage of one model as working basis is preferable over the other two.

4. What is the main idea behind the Nadaraya-Watson estimator $\hat{f}^{NW}_{n,h}$? What is the difference between $\hat{f}^{NW}_{n,h}$ and a local polynomial estimator?

5. In a regression model with deterministic design introduce the projection estimator and discuss its MISE. Explain by example the quantities 'approximation space' and 'order of approximation'.

6. Are the derived upper bounds for pointwise density estimation optimal? Give an outline of the key steps to verify optimality.

7. Regard a penalised estimation approach of your choice. To this end give possible representations and discuss the risk in view of the penalty.

8. Present at least two methods for choosing the regularisation parameter adaptively. Give an example in which rate optimality is attained.

Prepare brief solutions that you are able to present in the class on Monday, July 8.

1 Penalized least squares criterion for projection estimators

1.1 Setting

Observations: Gaussian sequence model $y_k = f_k + \frac{\sigma}{\sqrt{n}}\zeta_k$, $k \geqslant 1$, $\zeta_k \sim N(0,1)$ i.i.d.

For $m \geqslant 1$ let $S_m \subseteq \ell^2(\mathbb{N})$ be a d_m-dimensional subspace and $\hat{f}^{(m)} = \Pi_{S_m} y$ be the projection estimator onto S_m, where $\Pi_{S_m} y$ denotes the orthogonal projection of the data y onto S_m. Unbiased risk estimaton leads to choosing m adaptively via

$$\tilde{m} = \operatorname{argmin}_m \Big(- \|\hat{f}^{(m)}\|^2 + 2 d_m \tfrac{\sigma^2}{n} \Big).$$

We consider more generally:

1.1 Definition. The penalised least squares criterion for selecting among projection estimators is given by

$$\tilde{m} = \operatorname{argmin}_m \Big(- \|\hat{f}^{(m)}\|^2 + \mathrm{Pen}(d_m) \Big)$$

with a penalty function $\mathrm{Pen} : \mathbb{N} \to [0, \infty)$.

1.2 Lemma. *(Fundamental Inequality) For any $m^* \geqslant 1$ and $\hat{f} = \hat{f}^{(\tilde{m})}$, $f^* = \Pi_{S_{m^*}} f$ we have*

$$\|\hat{f} - f\|^2 \leqslant \|f^* - f\|^2 + \mathrm{Pen}(d_{\tilde{m}}) + 2\tfrac{\sigma}{\sqrt{n}} \langle \zeta, \hat{f} - f^* \rangle - \mathrm{Pen}(d_{\tilde{m}}).$$

Proof. By definition of $\tilde{m}$ we have

$$-\|\hat{f}\|^2 + \mathrm{Pen}(d_{\tilde{m}}) \leqslant -\|\hat{f}^{(m^*)}\|^2 + \mathrm{Pen}(d_{m^*}).$$

We therefore obtain

$$\begin{aligned}
&\|\hat{f} - f\|^2 + \mathrm{Pen}(d_{\tilde{m}}) - \mathrm{Pen}(d_{m^*}) \\
&= -\|\hat{f}\|^2 + \mathrm{Pen}(d_{\tilde{m}}) - \mathrm{Pen}(d_{m^*}) + 2\langle \hat{f}, \hat{f} - f \rangle + \|f\|^2 \\
&\leqslant -\|\hat{f}^{(m^*)}\|^2 + 2\langle \hat{f}, \hat{f} - f \rangle + \|f\|^2 \\
&\leqslant \|\hat{f}^{(m^*)} - f^*\|^2 - \|\hat{f}^{(m^*)}\|^2 + 2\langle \hat{f}, \hat{f} - f \rangle + \|f\|^2 \\
&= -\|f^*\|^2 + 2\langle f^* - \hat{f}^{(m^*)}, f^* \rangle + 2\langle \hat{f}, \hat{f} - f \rangle + \|f\|^2 \\
&= \|f^* - f\|^2 + 2\tfrac{\sigma}{\sqrt{n}} \langle \zeta, \hat{f} - f^* \rangle,
\end{aligned}$$

where in the last line $\|f\|^2 - \|f^*\|^2 = \|f - f^*\|^2$ by Pythagoras and $\langle \hat{f}, \hat{f} - f \rangle = \langle \hat{f}, y - f \rangle$, $\langle f^* - \hat{f}^{(m^*)}, f^* \rangle = \langle f - y, f^* \rangle$ by orthogonal projection properties. □

1.3 Lemma. *For $Z \sim \chi^2(p)$ and $x > 0$ we have the exponential deviation inequality*

$$P(Z - p \geqslant 2\sqrt{px} + 2x) \leqslant e^{-x}.$$

Proof. From $\mathbb{E}[e^{\alpha Z}] = (1-2\alpha)^{-p/2}$ for $0 \leqslant \alpha < 1/2$ we obtain by Markov's inequality

$$\begin{aligned} P(Z - p \geqslant 2\sqrt{px} + 2x) &= P(\exp(\alpha Z) \geqslant \exp(\alpha(p + 2\sqrt{px} + 2x))) \\ &\leqslant (1-2\alpha)^{-p/2} \exp(-\alpha(p + 2\sqrt{px} + 2x)) \\ &= \exp\Big(-\tfrac{p}{2}\log(1-2\alpha) - \alpha(p + 2\sqrt{px} + 2x)\Big). \end{aligned}$$

With $\kappa = 1 + 2\sqrt{x/p} + 2x/p$ the right-hand side is minimised by $\alpha = (1-\kappa^{-1})/2$ such that

$$P(Z - p \geqslant 2\sqrt{px} + 2x) \leqslant \exp\Big(\tfrac{p}{2}\log(\kappa) - \tfrac{\kappa-1}{2}p\Big).$$

Because of $e^{2\sqrt{x/p}} = 1 + 2\sqrt{x/p} + 2x/p + \cdots \geqslant \kappa$ we obtain further

$$P(Z - p \geqslant 2\sqrt{px} + 2x) \leqslant \exp\Big(\tfrac{p}{2}(2\sqrt{x/p}) - (\sqrt{x/p} + x/p)p\Big) = e^{-x}.$$

□

1.4 Remark. Consider the standardisation $(Z-p)/\sqrt{2p}$ and compare with $N(0,1)$-deviations, which are achieved as $p \to \infty$.

1.5 Theorem. *Assume* $\mathrm{Pen}(d) \geqslant C\frac{\sigma^2}{n}(d+1)$ *for some* $C > 1$ *and set*

$$\varepsilon_{C,\tau} := e^{-\tau} \sum_{m \geqslant 1} e^{-c^2(d_m+1)} \text{ with } c = \frac{1}{2}(\sqrt{1 + 2(C-1)/(C+1)} - 1)$$

and some $\tau > 0$. *Then with probability at least* $1 - \varepsilon_{C,\tau}$ *we have the oracle inequality*

$$\|\hat{f} - f\|^2 \lesssim \inf_{m^* \geqslant 1} \Big(\|f - \Pi_{S_{m^*}} f\|^2 + \mathrm{Pen}(d_{m^*}) \Big) + \tau\frac{\sigma^2}{n},$$

where the constant involved in '$\lesssim$' *is deterministic and depends on* C *only, in particular not on* f.

Proof. By the fundamental inequality and the assumption on the penalty we need to control for any $m^* \geqslant 1$

$$2\tfrac{\sigma}{\sqrt{n}}\langle \zeta, \hat{f} - f^* \rangle - \mathrm{Pen}(d_{\tilde{m}}) \leqslant 2\tfrac{\sigma}{\sqrt{n}}\langle \zeta, \hat{f} - f^* \rangle - C\tfrac{\sigma^2}{n}(d_{m^*} + 1).$$

Let S'_m denote the at most (d_m+1)-dimensional subspace spanned by S_m and f^*. Then

$$\langle \zeta, \hat{f} - f^* \rangle = \langle \Pi_{S'_m}\zeta, \hat{f} - f^* \rangle \leqslant \|\Pi_{S'_m}\zeta\| \|\hat{f} - f^*\|$$

follows. We use $2AB \leqslant \eta A^2 + \eta^{-1}B^2$ for any $A, B \in \mathbb{R}$ and $\eta > 0$ to deduce

$$2\tfrac{\sigma}{\sqrt{n}}\langle \zeta, \hat{f} - f^* \rangle - \mathrm{Pen}(d_{\tilde{m}}) \leqslant \tfrac{2}{1+C}\|\hat{f} - f^*\|^2 + \tfrac{\sigma^2}{n} \sup_{m \geqslant 1} \Big(\tfrac{1+C}{2}\|\Pi_{S'_m}\zeta\|^2 - C(d_m + 1) \Big).$$

Introduce $x = \tau + c^2 p$ and calculate from the definition that $c + c^2 = \frac{C-1}{2C+2}$. We derive from $\sqrt{1+y} \leqslant 1 + y/2$

$$2\sqrt{px} + 2x = 2cp\sqrt{1 + \tau/(cp)} + 2\tau + 2c^2 p \leqslant \tfrac{C-1}{C+1}p + 3\tau.$$

Noting $\|\Pi_{S'_m}\zeta\|^2 \sim \chi^2(\dim(S'_m))$, the preceding lemma thus shows

$$P\Big(\|\Pi_{S'_m}\zeta\|^2 - (d_m+1) \geqslant \tfrac{C-1}{C+1}(d_m+1) + 3\tau\Big) \leqslant e^{-\tau - c^2(d_m+1)}.$$

We apply the union bound and obtain

$$\begin{aligned}
&P\Big(\sup_{m\geqslant 1}(\|\Pi_{S'_m}\zeta\|^2 - \tfrac{2C}{1+C}(d_m+1)) \geqslant 3\tau\Big) \\
&\leqslant \sum_{m\geqslant 1} P\Big(\|\Pi_{S'_m}\zeta\|^2 - (d_m+1) \geqslant \tfrac{C-1}{C+1}(d_m+1) + 3\tau\Big) \\
&\leqslant e^{-\tau}\sum_{m\geqslant 1} e^{-c^2(d_m+1)} = \varepsilon_{C,\tau}.
\end{aligned}$$

This means that with probability at least $1-\varepsilon_{C,\tau}$

$$2\tfrac{\sigma}{\sqrt{n}}\langle \zeta, \hat{f} - f^*\rangle - \mathrm{Pen}(d_{\hat{m}}) \leqslant \tfrac{2}{C+1}\|\hat{f} - f^*\|^2 + \tfrac{3}{2}(C+1)\tfrac{\sigma^2}{n}\tau$$

and by the fundamental inequality

$$(1 - \tfrac{2}{C+1})\|\hat{f} - f\|^2 \leqslant \tfrac{2}{C+1}\|f^* - f\|^2 + \mathrm{Pen}(d_{m^*}) + \tfrac{3}{2}(C+1)\tfrac{\sigma^2}{n}\tau.$$

Dividing by $1 - \frac{2}{C+1} = \frac{C-1}{C+1} > 0$ and taking the minimiser m^*, we arrive at the claimed oracle inequality. □

1.6 Corollary. *Consider for $m = K$ the K-dimensional subspaces $S_K = \{(y_k)_{k\geqslant 1} \,|\, \forall\, k > K : y_k = 0\}$ and the projection estimators $\hat{f}^{(K)}$ with $\hat{f}^{(K)}_k = y_k \mathbf{1}(k \leqslant K)$. Then with the penalty $\mathrm{Pen}(K) = 2\frac{\sigma^2}{n}K$ according to unbiased risk estimation we obtain the oracle inequality*

$$\mathbb{E}[\|\hat{f} - f\|^2] \lesssim \inf_{K\geqslant 1} \mathbb{E}[\|\hat{f}^{(K)} - f\|^2].$$

Proof. Insert $d_K = K$ and $C = 2$ in the preceding theorem to infer $\varepsilon_{C,\tau} \lesssim e^{-\tau}$. For this note (rather check!) that the statement also holds under the weaker condition $\mathrm{Pen}(d) \geqslant C\frac{\sigma^2}{n}d$, only with a different constant. Therefore with probability at least $1-\varepsilon_{C,\tau}$

$$\|\hat{f} - f\|^2 \lesssim \inf_{K\geqslant 1}\Big(\|f - \Pi_{S_K} f\|^2 + \tfrac{\sigma^2}{n}K\Big) + \tau\tfrac{\sigma^2}{n}.$$

For a nonnegative random variable Z we have $\mathbb{E}[Z] = \int_0^\infty P(Z \geqslant \tau)d\tau$ and thus

$$\mathbb{E}[\|\hat{f} - f\|^2] \lesssim \inf_{K\geqslant 1}\Big(\|f - \Pi_{S_K} f\|^2 + \tfrac{\sigma^2}{n}K\Big) + \tfrac{\sigma^2}{n}\int_0^\infty \varepsilon_{C,\tau}d\tau.$$

Due to $\varepsilon_{C,\tau} \lesssim e^{-\tau}$ the integral is finite and we conclude by adjusting the constant and using the bias-variance decomposition for the projection estimator:

$$\mathbb{E}[\|\hat{f} - f\|^2] \lesssim \inf_{K\geqslant 1}\Big(\|f - \Pi_{S_K} f\|^2 + \tfrac{\sigma^2}{n}K\Big) = \inf_{K\geqslant 1}\mathbb{E}[\|\hat{f}^{(K)} - f\|^2].$$

□

1.7 Corollary. *Consider the projection estimators from the preceding corollary and assume that the Gaussian sequence model is obtained by the Fourier coefficients of a function $f \in L^2([0,1])$ observed in white noise. Then for functions f in the periodic Sobolev spaces $H^s_{per}([0,1])$, $s > 0$, $\hat{f}$ adapts automatically to the unknown regularity s:*

$$\mathbb{E}[\|\hat{f} - f\|^2] \lesssim n^{-2s/(2s+1)}.$$

Proof. By previous results the projection estimator with minimax optimal choice of K attains the rate $n^{-s/(2s+1)}$ and by the preceding theorem we lose at most a constant factor in the risk when using the adaptive estimator $\hat{f}$. □

1.8 Remark. Compared to the Lasso estimator we do not even lose a logarithmic factor. Although the size of $\log n$ is in most applications reasonably small, the log factor still indicates some more loss in precision, which is confirmed by the better finite sample properties of unbiased risk estimation compared to Lasso if the coefficients (f_k) have certain decay properties (as for Fourier coefficients of Sobolev functions). Nowadays, results with exact constants, e.g. a factor $1 + o(1)$ in front of the oracle risk, are available. These exact oracle inequalities convey, of course, a much more precise description.