Bayesian wavelet estimators in nonparametric regression

Natalia Bochkina

University of Edinburgh

Lecture 2. Classical minimax consistency and concentration of posterior measures

1. Decision-theoretic approach to classical consistency and concentration of posterior measures
2. Classical consistency
   - Bayes and minimax estimators
   - Rate of convergence
   - Lower bounds
   - Adaptivity
3. Concentration of posterior measures

1. Decision-theoretic approach to consistency and concentration of posterior measures

Decision theory for estimation:

- set of outcomes: $f_0 \in \mathcal{F}$ - “true” state of nature

  not observed directly but with an error: $Y_i = f(x_i) + \epsilon_i, i = 1, \ldots, n$
- set of decisions: $\delta(Y) = \hat{f}_n$ - estimators of f, where $Y = (Y_1, \ldots, Y_n)^T$.
- a loss function: $Q(f, \delta(Y))$

Classical approach:

Risk function $R(\delta, f) = \mathbb{E}_f Q(f, \delta(Y))$,

Aim: choose the decision function $\delta(Y)$ that minimises the risk $R(\delta, f)$.

f is unknown, want to choose $\delta(Y)$ that “works” for all $f \in \mathcal{F}$.

Two approaches: Bayes and minimax.

Bayes and minimax risks

Minimax risk: $R^M(\delta, \mathcal{F}) = \sup_{f \in \mathcal{F}} R(\delta, f)$.

Definition 1. *Decision $\delta^M(Y)$ is called minimax iff*

$$R^M(\delta^M, \mathcal{F}) = \inf_{\delta} R^M(\delta, \mathcal{F}) = \inf_{\delta} \sup_{f \in \mathcal{F}} \mathbb{E}_f Q(f, \delta(Y)).$$

Worst case scenario.

Bayes risk: suppose we have a probability measure π over $\mathcal{F}$, then the corresponding Bayes risk is

$$R^\pi(\delta, \mathcal{F}) = \mathbb{E}_\pi R(\delta, f) = \int_{\mathcal{F}} R(\delta, f)\pi(df) = \int_{\mathcal{F}} \int_{\mathbb{R}^n} Q(\delta(Y), f) p(Y \mid f) dY \pi(df).$$

Definition 2. *Given a probability measure π over $\mathcal{F}$, decision $\delta^\pi(Y)$ is called Bayes iff*

$$R^\pi(\delta^\pi, \mathcal{F}) = \inf_{\delta} R^\pi(\delta, \mathcal{F}).$$

δ^π minimises average risk with respect to π.

Both risks are frequentist, as the loss function is averaged over data Y.

Connection between minimax and Bayes estimators

Minimax estimators are often Bayes estimators.

Lemma 1. *Suppose that prior measure* π *is such that* $R^{\pi}(\delta, \mathcal{F}) = R^{M}(\delta, \mathcal{F})$*, i.e.*

$$\int_{\mathcal{F}} R(\delta, f)\pi(df) = \sup_{f \in \mathcal{F}} R(\delta, f).$$

Then,

- δ^{π} *is minimax*
- π *is a least favourable prior, i.e.* $R^{\pi}(\delta, \mathcal{F}) \geq R^{\pi'}(\delta, \mathcal{F})$ *for all probability measures* π'*.*

Bayesian decision theory

Optimal decision $\delta(Y)$ is for the given data Y:

$$\begin{aligned} \delta(Y) &= \arg\min_{\delta} \mathbb{E}[Q(\delta(Y), f) \mid Y] = \arg\min_{\delta} \int_{\mathcal{F}} Q(\delta(Y), f)\pi(df \mid Y) \\ &= \arg\min_{\delta} \int_{\mathcal{F}} Q(\delta(Y), f) p(Y \mid f)\pi(df). \end{aligned}$$

In practice the optimal Bayesian and frequentist decision rules coincide.

Bayesian estimators:

- $Q(\delta(Y), f) = I(\delta(Y) \neq f)$, optimal estimator: posterior mode (MAP) estimator
- $Q(\delta(Y), f) = ||\delta(Y) - f||_2^2$, optimal estimator: posterior mean
- $Q(\delta(Y), f) = ||\delta(Y) - f||_1$, optimal estimator: posterior median

2. Consistency of point estimators

Definition 3. *$\hat{f}_n$ is a (weakly) consistent estimator of f (with respect to distance d) iff for any $\epsilon > 0$,*

$$\mathbb{P}(d(\hat{f}_n, f) > \epsilon) \to 0 \quad \textit{as} \quad n \to \infty.$$

Commonly considered distances:

- error of estimation at a point: $d(\hat{f}_n, f) = |\hat{f}_n(x_0) - f(x_0)|$, for some $x_0 \in [0, 1]$
- integrated error: $d(\hat{f}_n, f) = ||\hat{f}_n(x_0) - f(x_0)||_u$,

where $||g||_u = \left(\int_0^1 |g(x)|^u dx \right)^{1/u}$ - norm of $L^u([0, 1])$, $u \in [1, \infty]$.

Consistency of point estimators

For $u = 2$, sufficient condition: variance and bias of $\hat{f}_n$ go to 0 as $n \longrightarrow \infty$:

$$\begin{aligned}\mathbb{P}[|\hat{f}_n(x_0) - f(x_0)| > \epsilon] &\leq \epsilon^{-2}\mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2] \\ &= \epsilon^{-2}[\mathbb{E}|\hat{f}_n(x_0) - \mathbb{E}\hat{f}_n(x_0)|^2 + |\mathbb{E}\hat{f}_n(x_0) - f(x_0)|^2].\end{aligned}$$

Convergence in mean (of $\mathbb{E}[d(\hat{f}_n, f)]^u$ for some $u > 0$) implies consistency.

Minimax rates of convergence

Look for estimators that achieve consistency over a set of functions $\mathcal{F}$.

To prove consistency in distance d it is sufficient to show convergence in mean, i.e. of $\mathbb{E}[d(\hat{f}_n, f)]^u$ for some $u > 0$.

This would lead to the “optimal” choice of tuning parameters for a chosen type of estimator, e.g. kernel, local polynomial or projection estimators.

Illustrate on consistency of the local polynomial estimator over Hölder spaces.

Hölder spaces

Definition 4. *Let $\beta > 0$, $M > 0$. The Hölder class $\mathbb{H}^\beta(M)$ on $[0,1]$ is defined as the set of $k = \lfloor \beta \rfloor$ times differentiable functions f : $[0,1] \to \mathbb{R}$ whose derivative $f^{(k)}$ satisfies*

$$|f^{(k)}(x) - f^{(k)}(y)| \le M|x-y|^{\beta-k}, \quad \forall x, y \in [0,1].$$

Local polynomial estimator

Definition 5. *Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel, $h > 0$ be a bandwidth, and $k > 0$ be an integer. The statistic $\hat{f}_n(x) = U^T(0)\widehat{\theta}_n(x)$ with*

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

is called a local polynomial estimator of order k of $f(x)$, or $LP(k)$ estimator of $f(x)$ for short.

Recall that

$$U(u) \quad = \quad (1, u, u^2/2!, \ldots, u^k/k!)^T.$$

Local polynomial estimator

For a fixed x the LP estimator is a weighted least squares estimator. Indeed, we can write $\hat{\theta}_n(x)$ as follows:

$$\hat{\theta}_n(x) = \arg\min_{\theta \in \mathbb{R}^{k+1}} (-2\theta^T a_{nx} + \theta^T B_{nx} \theta),$$

where the matrix B_{nx} and the vector a_{nx} are defined by the formulas

$$\begin{aligned} B_{nx} &= \frac{1}{nh} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^T \left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right), \\ a_{nx} &= \frac{1}{nh} \sum_{i=1}^{n} Y_i U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right). \end{aligned}$$

Hence, if matrix B_{nx} is invertible, $LP(k)$ estimator at x exists and is unique: $\hat{\theta}_n(x) = B_{nx}^{-1} a_{nx}$.

Assumptions (LP)

(LP1) There exist a real number $\lambda_0 > 0$ and a positive integer n_0 such that the smallest eigenvalue $\lambda_{\min}(B_{nx})$ of B_{nx} satisfies

$$\lambda_{\min}(B_{nx}) \geq \lambda_0$$

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

(LP2) There exists a real number $a_0 > 0$such that for any interval $A \subseteq [0, 1]$ and all $n \geq 1$,

$$\frac{1}{n} \sum_{i=1}^{n} I(X_i \in A) \leq a_0 \max(\mu(A), 1/n),$$

where $\mu(A)$ denotes the Lebesgue measure of A.

(LP3) The kernel K has compact support belonging to $[-1, 1]$ and there exists a number $K_{\max} < \infty$ such that $|K(u)| \leq K_{\max}, \forall u \in \mathbb{R}$.

Variance and bias for $LP(k)$ estimator

Denote bias $b(x_0) = \mathbb{E}_f \hat{f}_n(x_0) - f(x_0)$ and variance $\sigma^2(x_0) = \mathbb{E}_f |\hat{f}_n(x_0) - \mathbb{E}_f \hat{f}_n(x_0)|^2$.

Proposition 1. *Suppose that $f \in \mathbb{H}^\beta(M)$ on $[0,1]$, with $\beta > 0$ and $M > 0$. Let $\hat{f}_n$ be the $LP(k)$ estimator of f with $k = \lfloor \beta \rfloor$.*

Assume also that:

(i) the design points $X_1, \ldots, X_n$ are deterministic;

(ii) assumptions (LP1)-(LP3) hold;

(iii) the random variables ϵ_i are independent and such that for all $i = 1, \ldots, n$,

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

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

$$|b(x_0)| \leq q_1 h^\beta, \quad \sigma^2(x_0) \leq \frac{q_2}{nh},$$

where $q_1 = C_* L/k!$ and $q_2 = \sigma_{\max}^2 C_*$, $C_* = \frac{2K_{\max}}{\lambda_0} \max 1, 2a_0$.

Consistency of the local polynomial estimator over Holder spaces

Proposition 1 implies that

$$MSE = b^2(x_0) + \sigma^2(x_0) \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}$$

and that the minimizer h_n with respect to h of this upper bound on the risk is given by

$$h_n = \left(\frac{q_2}{2\beta q_1^2} \right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

Theorem 1. *Under the assumptions of Proposition 1 and if the bandwidth is chosen to be* $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}$, $\alpha > 0$, *the following upper bound holds:*

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}^\beta(M)} \sup_{x_0\in[0,1]} \mathbb{E}_f \left[n^{\frac{\beta}{2\beta+1}} |f(x_0) - \hat{f}_n(x_0)| \right]^2 \leq C < \infty,$$

where C *is a constant depending only on* β, M, a_0, λ_0, $\sigma^2_{\max}$, $K_{\max}$ *and* α.

Corollary 1. *Under the assumptions of Theorem 1 we have*

$$\limsup_{n\to\infty} \sup_{f\in\mathbb{H}^\beta(M)} \mathbb{E}_f \left[n^{\frac{\beta}{2\beta+1}} ||f - \hat{f}_n||_2 \right]^2 \leq C < \infty,$$

Rate of convergence

Rate of convergence of $\hat{f}_n$ in distance $d(f,g)$ over a set of functions $\mathcal{F}$:

$$r_n = \inf\left\{\epsilon_n > 0 : \quad \lim_{n\to\infty} \sup_{f\in\mathcal{F}} \mathbb{E}_f[\epsilon_n^{-1} d(\hat{f}_n, f)]^u \leqslant C < \infty\right\}.$$

Aims:

- find a consistent estimator of f over $\mathcal{F}$ that achieves the best possible rate of convergence
- ideally: characterisation of the set of all consistent estimators of f over $\mathcal{F}$ with the the best possible rate of convergence

Need to determine the best possible rate of convergence.

Lower bounds

Given $d(\hat{f}_n, f)$ and set of functions $\mathcal{F}$, find the best possible rate of convergence r_n:

$$\forall \hat{f}_n, \quad \sup_{f \in \mathcal{F}} \mathbb{E}_f [d(\hat{f}_n, f)]^u \geq C(\mathcal{F}, u) r_n^u.$$

Definition 6. *A positive sequence $\{r_n\}_{n=1}^{\infty}$ is called an optimal rate of convergence of estimators on $(\mathcal{F}, d)$ iff $\exists 0 < c(\mathcal{F}, u) \leq C(\mathcal{F}, u) < \infty$:*

$$\exists \hat{f}_n : \quad \sup_{f \in \mathcal{F}} \mathbb{E}_f [d(\hat{f}_n, f)]^u \leq C(\mathcal{F}, u) r_n^u$$

$$\forall \hat{f}_n, \quad \sup_{f \in \mathcal{F}} \mathbb{E}_f [d(\hat{f}_n, f)]^u \geq c(\mathcal{F}, u) r_n^u.$$

An estimator $\hat{f}_n$ satisfying

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f [d(\hat{f}_n, f)]^u \geq C' r_n^u,$$

where $\{r_n\}$ is the optimal rate of convergence and $C' < \infty$ is a constant, is called a rate optimal estimator on $(\mathcal{F}, d)$.

Optimal rates of convergence: global estimation, Holder spaces

Lower bound (Tsybakov, 2009), $d(\hat{f}_n, f) = ||\hat{f}_n - f||_2$, $u > 0$, $\mathcal{F} = \mathbb{H}^\beta(M)$.

Theorem 2. *Let $r > 0$ and $M > 0$, and assume that $Y_i = f(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, with deterministic x_i and iid ε_i: $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 < \infty$, with density $p_\varepsilon(u)$ wrt Lebesgue measure on $\mathbb{R}$ such that*

$$\exists p_\star > 0,\ v_0 > 0 : \int p_\varepsilon(u) \log \frac{p_\varepsilon(x)}{p_\varepsilon(u+v)} du \leq p_\star v^2$$

for all $|v| \leq v_0$.

Then,

$$\lim_{n\to\infty} \inf_{\hat{f}_n} \inf \sup_{f \in \mathbb{H}^r(M)} \mathbb{E}_f \left[n^{\frac{\beta}{2\beta+1}} ||f - \hat{f}_n||_2 \right]^2 \geq c(\beta, M, p_\star) > 0.$$

Optimal rates of convergence for LP estimators over Hölder spaces

Hence, if $x_i = i/n$, the local polynomial estimator of order $k = \lfloor \beta \rfloor$ with kernel K satisfying

$$\exists K_{\min} > 0,\ \Delta > 0,\ K_{\max} < \infty : \quad K_{\min} I(|u| \le \Delta) \le K(u) \le K_{\max} I(|u| \le \Delta)\,\forall u \in$$

and bandwidth $h_n = \alpha n^{-\frac{1}{2\beta+1}}$ for some $\alpha > 0$,

is rate optimal on $(\mathbb{H}^{\beta}(M), ||\cdot||_2)$,

and $r_n = n^{-\frac{\beta}{2\beta+1}}$ is the optimal rate of convergence for $(\mathbb{H}^{\beta}(M), ||\cdot||_2)$.

Optimal rates of convergence over Hölder spaces for $||\cdot||_u$

For $u \in [1, \infty)$: the optimal rate of convergence over $(\mathbb{H}^\beta(M), ||\cdot||_u)$ is also $r_n = n^{-\frac{\beta}{2\beta+1}}$.

However, for $u = \infty$: $||g||_\infty = \sup_{x\in[0,1]} |g(x)|$, the optimal rate of convergence (in the minimax sense) under Gaussian iid errors is

$$r_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}.$$

(Tsybakov, 2009).

Estimation at a point

For lower bound for the pointwise estimation over Hölder class $\mathbb{H}^\beta(M)$, with $d(f, \hat{f}_n) = |f(x_0) - \hat{f}_n(x_0)|$, is given by

$$\inf_{\tilde{f}_n} \sup_{f \in \mathbb{H}^\beta(M)} \mathbb{E}|\tilde{f}_n(x_0) - f(x_0)|^u \asymp n^{-\frac{u\beta}{2\beta+1}},$$

(Cai, 2003 - for white noise model, Bochkina & Sapatinas, 2009 - for nonparametric regression model).

Hence, the local polynomial estimator with kernel and bandwidth specified above is also locally rate optimal on $(\mathbb{H}^\beta(M), |f(x_0) - g(x_0)|)$.

Adaptivity

In order to be rate optimal over $\mathbb{H}^\beta$, $\hat{f}_n$ has to depend on smoothness of f, β.

In practice, β may not be known.

Definition 7. *An estimator $\hat{f}_n$ of f is called asymptotically efficient on the class $\mathcal{F}$ iff*

$$\lim_{n\to\infty} \frac{\sup_{f\in\mathcal{F}} \mathbb{E}_f ||\hat{f}_n - f||_2^2}{\inf_{\tilde{f}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f ||\hat{f}_n - f||_2^2} = 1,$$

where the infimum is over all estimators.

Definition 8. *An estimator $\hat{f}_n$ of f is called adaptive in the exact minimax sense on the family of classes $\{\mathbb{H}^\beta(M),\ \beta > 0, M > 0\}$ if it is asymptotically efficient for all classes $\mathbb{H}^\beta(M)$, $\beta > 0$,$M > 0$, simultaneously.*

Optimality of adaptive estimators

Now consider the subset of estimators $\mathcal{A}$ that do not depend on β or M.

Adaptive rate of convergence $r_{n,A}$ over $(\mathbb{H}^{\beta}(M), d)$:

$$\exists \hat{f}_n \in \mathcal{A} : \quad \sup_{f \in \mathbb{H}^{\beta}(M)} \mathbb{E}_f [d(\hat{f}_n, f)]^2 \leq C(\beta, M) r_{n,A}^2 < \infty$$

$$\forall \hat{f}_n \in \mathcal{A}, \quad \sup_{f \in \mathbb{H}^{\beta}(M)} \mathbb{E}_f [d(\hat{f}_n, f)]^2 \geq c(\beta, M) r_{n,A}^2 > 0.$$

Payment for adaptation in pointwise convergence rate over Hölder spaces.

Lepski (1990), white noise model: showed that the adaptive pointwise rate of convergence over Hölder spaces $\mathbb{H}^{\beta}(L)$ is $r_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$, i.e. there is an additional log factor.

Proposition 2. *(Cai, 2003) Let $u \in [1, \infty)$. Consider two Hölder classes $\mathbb{H}^{\beta_i}(M_i)$ for $i = 1, 2$. Let $\beta 1 > \beta_2 > 0$. If an estimator $\hat{f}_n$ attains a rate of $n^{-\rho}$ over $\mathbb{H}^{\beta_1}(M_1)$ with $\rho > u\beta_2/(1 + 2\beta_2)$, in particular, if $\hat{f}_n$ is rate-optimal over $\mathbb{H}^{\beta_1}(M_1)$, then*

$$\lim_{n\to\infty} \inf \left(\frac{n}{\log n}\right)^{u\beta_2/(1+2\beta_2)} \sup_{f\in\mathbb{H}^{\beta_2}(M_2)} \mathbb{E}_f|\hat{f}_n(x_0) - f(x_0)|^u > 0.$$

For the integral convergence rate, it is possible to avoid payment for adaptation (Lepski & Spokoiny, 1997, Cai, 2000). Studied by Cai(2008).

The rate is $\left(\frac{\log n}{n}\right)^{\beta/(1+2\beta)}$ called *adaptive pointwise minimax rate* over Hölder class $\mathbb{H}^{\beta}(M)$, and it is possible to avoid...

Lepski method.

Choice of data-driven bandwidth on $[h_{\min}, h_{\max}]$, where $h_{\min} = \frac{\log n}{n}$ is the smallest bandwidth for which $f_n(x; h_{\min})$ is still a consistent estimator of f, $h_{\max} = 1$.

- Start with a kernel estimator $f_n(x;h) = \int K_h(x-y) dP_n(y \mid f)$
- Choose a discrete 'logarithmic' grid H of candidate bandwidths:

$$H = \left\{ h_0 = h_{\max},\ h_{k+1} = \frac{h_k}{1 + [d(h_k)]^{-1/2}},\ k = 0, 1, \ldots \right\},$$

  where $d(h) = \sqrt{\max(1, c \log(h_{\max}/h))}$.

- Select a data-driven bandwidth $\hat{h}_n$ to be the maximal element of H such that

  $$\hat{h}_n = \max\{h \in H : |f_n(x_0, h) - f_n(x_0, g)| \leq (1 + d(g)^{-1/2}) \sigma_n(g) d(g) \forall g < h, g \in$$

  Here $\sigma_n^2(h) = \frac{||K||_2^2}{nh}$ - variance of kernel estimator $f_n(x_0, h)$.

- Use $\hat{f}_n^L(x) = f_n(x, \hat{h}_n)$ as the fully data driven estimator of f.

Adaptive kernel estimator

Then, if s is the order of the kernel K, for every $\beta \in (0, s]$,

$$\sup_{f \in \mathbb{H}^\beta(L)} \mathbb{E}_f \sup_{x \in [0,1]} |\hat{f}_n^L(x) - f(x)| \leq C r_n(\beta),$$

where $r_n = \left(\frac{\log n}{n}\right)^{\beta/(1+2\beta)}$.

(Lepski & Spokoiny, 1997).

Further questions

How realistic is the model white noise/Gaussian assumption?

Assume: $Y_i \sim \mathcal{N}(f(x_i), \sigma)$ independent, derive an “optimal” estimator of f.

If the model assumption is wrong, how far can the data deviate from this model in order for the estimator to remain optimal?

Golubev & Spokoiny (2009) - for parametric estimation.

Generalisations

- non-Gaussian errors (Chichignoud, 2010; Gannaz (2011): GLM with ℓ_1 norm penalty)
- multivariate case: Goldenshluger & Lepski (2008) Universal pointwise selection rule in multivariate function estimation. Bernoulli, 14(4), 1150-1190.
- multivariate case with composite functions (Juditsky, Lepski, Tsybakov, 2009).

In the current work: some restrictive assumptions.

Confidence regions, in the context of density estimation:

- L^p balls [Hoffmann and Lepski (2002, AoS), Baraud (2004, AoS), Cai and Low (2006, AoS), Robins and van der Vaart (2006, AoS)],
- uniform (L^∞) confidence bands (Gine & Nickl).

Equivalence of experiments

Lucien Le Cam (1986) Asymptotic methods in statistical decision theory, Springer.

Consider two statistical problems, $\mathcal{P}_1$ and $\mathcal{P}_2$, with the sample spaces $\mathcal{X}_i$, $i = 1, 2$ (and suitable σ-fields), but with the same parameter space Θ.

Let $\mathcal{D}$ be any (measurable) decision space and let $Q : \Theta \times \mathcal{D} \to [0, \infty)$ denote a loss function. Let $||Q|| = \sup[Q(\theta, d) : \theta \in \Theta, d \in \mathcal{D}]$. δ^i will be the generic symbol for a decision procedure in the ith problem. $R^{(i)}(\delta^i, Q, \theta)$ is the risk using procedure δ^i under loss Q and true parameter θ. Le Cam metric is

$$\Delta(\mathcal{P}_1, \mathcal{P}_2) = \max \left[\inf_{\delta^1} \sup_{\delta^2} \sup_{\theta} \sup_{Q:\,||Q||=1} |R^1(\delta^1, Q, \theta) - R^2(\delta^2, Q, \theta)|, \right.$$
$$\left. \inf_{\delta^2} \sup_{\delta^1} \sup_{\theta} \sup_{Q:\,||Q||=1} |R^1(\delta^1, Q, \theta) - R^2(\delta^2, Q, \theta)| \right].$$

Thus, if $\Delta(\mathcal{P}_1, \mathcal{P}_2) \leq \epsilon$, this means that for every procedure δ^i in problem i $\exists$ a procedure δ^j in problem j ($i \neq j$), with risk differing by at most ϵ, uniformly over all Q and θ.

3. Concentration of posterior measures

Nonparametric regression model: $Y_i \sim \mathcal{N}(f(x_i), \sigma^2)$, independent.

Assume f_0 is the “true” function that generated the data.

1. A. van der Vaart & H. Zanten (2008) Rates of contraction of posterior distributions based on Gaussian process priors. Annals of Statistics, 36(3), 1435 - 1463.

The rate of convergence (rate of contraction) is smallest r_n such that

$$\mathbb{P}(d(f, f_0) > M r_n \mid Y_1, \ldots, Y_n) \to 0 \quad \text{as} \quad n \to \infty,$$

for sufficiently large $M > 0$.

Concentration of posterior measures

Theorem 3. *(van der Vaart & Zanten, 2008). Assume* $Y_i \mid f \sim \mathcal{N}(f(x_i), \sigma^2)$ *iid,* $x_i \in \mathcal{X}$ *are fixed, with "true" values* (f_0, σ_0)*.*

Let prior on f *be a zero mean Gaussian process* W *with bounded sample paths and RKHS* $\mathcal{H}$*, and suppose that* $f_0 \in supp(W)$*. Furthermore, take an absolutely continuous prior on* σ *to be supported on* $[a, b] \subset (0, \infty)$ *with a Lebesgue density that is bounded away from 0, such that* $\sigma_0 \in [a, b]$*.*

Then, the posterior distribution satisfies $\mathbb{E}_{f_0,\sigma_0}\mathbb{P}(||f - f_0||_n + |\sigma - \sigma_0| > Mr_n \mid Y_1, \ldots, Y_n) \to 0$ *for any sufficiently large constant* M *and* r_n *is defined as follows:*

$$\inf_{f \in \mathcal{H}:\, ||f - f_0||_n < r_n} ||f||^2_{\mathcal{H}} - \log \mathbb{P}(||W||_n < r_n) \leq nr_n^2,$$

where $||g||_n^2 = \frac{1}{n}\sum_{i=1}^n |g(x_i)|^2$*.*

Example: Integrated Brownian motion prior

Define $I^1_{0+}f$ as the function $t \to \int_0^t f(s)ds$ and $I^m_{0+}f$ as $I^1_{0+}(I^{m-1}_{0+}f)$.

Theorem 4. *(Theorem 4.1, A. van der Vaart & H. Zanten, 2008). Let W be a standard Brownian motion and $Z_0, \ldots, Z_k$ independent standard normal random variables. The RKHS of the process*

$$t \to I^k_{0+}W(t) + \sum_{i=0}^{k} Z_i t^i / i!$$

is the Sobolev space $W^{k+1}_2[0,1]$ with norm
$||g||^2_{\mathcal{H}} = ||g^{(k)}||^2_2 + \sum_{i=0}^{k}[g^{(i)}(0)]2$*.*

If $f \in C^\beta$ with $\beta = k + 1/2$, then the contraction rate is $r_n \asymp n^{-\frac{\beta}{2\beta+1}}$.

Also extended to non-iid observations:

Subhashis Ghosal and Aad van der Vaart (2007). Convergence rates of posterior distributions for non-iid observations. Ann. Statist. 35, 192223.

Concentration of posterior measures

2. Gine & Nickl (2011) On the uniform consistency of nonparametric Bayes estimates.

White noise model: $dY(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t)$.

Gaussian process prior with wavelet-based kernel that a priori belongs to a slightly modified Hölder class $\mathbb{H}^{\beta}(L)$ with probability 1:

$$f(t) = \sum_{k=0}^{N} \xi_k \varphi_{Lk}(t) + \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \sqrt{\mu_j} \xi_{jk} \psi_{jk}(t)$$

where $\xi_k, \xi_{jk} \sim \mathcal{N}(0,1)$ iid, $\mu_L = 1$ and $\mu_j = j^{-1} 2^{-j(2r+1)}\ \forall j > L$.

Concentration of posterior measures: Gine & Nickl (2011) (cont)

Rate of contraction in L^∞ norm $||f||_\infty = \sup_x |f(x)|$: $r_n = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$.

Theorem 5. *Let (φ, ψ) be scaling and wavelet Daubechies functions of regularity $s > r > 0$. Let $f_0 \in C^{r,\infty}([0,1])$, and suppose we observe $dY_0(t) = f_0(t)dt + \frac{1}{\sqrt{n}}dW(t)$.*

Then, there exist $C > 0$ and $M_0 < \infty$ depending only on wavelet basis, r and $||f_0||_{\alpha,\infty}$ such that, for every $M_0 \leq M < \infty$, and for all $n \in \mathbb{N}$,

$$\mathbb{E}_{Y_0}\mathbb{P}(f : ||f - f_0||_\infty > Mr_n \mid Y_0) \leq \exp\{-C^2(M - M_0)^2 \log n\}.$$