Bayesian Adaptation

Aad van der Vaart

http://www.math.vu.nl/ aad

Vrije Universiteit Amsterdam

Joint work with Jyri Lember

Adaptation

Given a collection of possible models
find a single procedure
that works well for all models

Adaptation

Given a collection of possible models
find a single procedure
that works well for all models

as well as a procedure specifically
targetted to the correct model

Adaptation

Given a collection of possible models
find a single procedure
that works well for all models

as well as a procedure specifically
targetted to the correct model

correct model is the one
that contains the true distribution of the data

Adaptation to Smoothness

Given a random sample of size n from a density p_0 on $\mathbb{R}$ that is known to have α derivatives,

there exist estimators $\hat{p}_n$ with rate $\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$

Adaptation to Smoothness

Given a random sample of size n from a density p_0 on $\mathbb{R}$ that is known to have α derivatives,

there exist estimators $\hat{p}_n$ with rate $\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$

i.e. $\mathrm{E}_{p_0} d^2(\hat{p}_n, p_0)^2 = O(\epsilon_{n,\alpha}^2)$,
uniformly in p_0 with $\int (p_0^{(\alpha)})^2 \, d\lambda$ bounded (if $d = \|\cdot\|_2$)

Distances

Global distances on densities

d can be one of:

Hellinger: $$h(p,q) = \sqrt{\int |\sqrt{p} - \sqrt{q}|^2 \, d\mu},$$

Total variation: $$\|p - q\|_1 = \int |p - q| \, d\mu,$$

L_2: $$\|p - q\|_2 = \sqrt{\int |p - q|^2 \, d\mu}.$$

Adaptation

Data	$X_1, \ldots, X_n$ i.i.d. p_0
Models	$\mathcal{P}_{n,\alpha}$ for $\alpha \in A$, countable
Optimal rates	$\epsilon_{n,\alpha}$

Adaptation

Data	$X_1, \ldots, X_n$ i.i.d. p_0
Models	$\mathcal{P}_{n,\alpha}$ for $\alpha \in A$, countable
Optimal rates	$\epsilon_{n,\alpha}$

p_0 contained in or close to $\mathcal{P}_{n,\beta}$, some $\beta \in A$

Adaptation

Data	$X_1, \ldots, X_n$ i.i.d. p_0
Models	$\mathcal{P}_{n,\alpha}$ for $\alpha \in A$, countable
Optimal rates	$\epsilon_{n,\alpha}$

p_0 contained in or close to $\mathcal{P}_{n,\beta}$, some $\beta \in A$

We want procedures that (almost) attain rate $\epsilon_{n,\beta}$, but we do not know β

Adaptation-NonBayesian

Main methods:

- Penalization
- Cross validation

Adaptation-NonBayesian

Main methods:

- Penalization
- Cross validation

Penalization:

Minimize your favourite contrast function (MLE, LS, ..), but add a penalty for model complexity

Adaptation-NonBayesian

Main methods:

- Penalization
- Cross validation

Penalization:
Minimize your favourite contrast function (MLE, LS, ..), but add a penalty for model complexity

Cross validation:
Split the sample
Use first half to select best estimator for each model
Use second half to select best model

Adaptation-Penalization

Models	$\mathcal{P}_{n,\alpha}, \qquad \alpha \in A$
Estimator given model	$\hat{p}_{n,\alpha} = \operatorname*{argmin}_{p \in \mathcal{P}_{n,\alpha}} M_n(p)$

Adaptation-Penalization

Models $\mathcal{P}_{n,\alpha}, \qquad \alpha \in A$

Estimator given model $\hat{p}_{n,\alpha} = \underset{p \in \mathcal{P}_{n,\alpha}}{\operatorname{argmin}} M_n(p)$

Estimator model $\hat{\alpha}_n = \underset{\alpha \in A}{\operatorname{argmin}} \big(M_n(\hat{p}_{n,\alpha}) + \mathrm{pen}_n(\alpha) \big)$

Adaptation-Penalization

Models $\mathcal{P}_{n,\alpha}, \qquad \alpha \in A$

Estimator given model $\hat{p}_{n,\alpha} = \operatorname*{argmin}_{p \in \mathcal{P}_{n,\alpha}} M_n(p)$

Estimator model $\hat{\alpha}_n = \operatorname*{argmin}_{\alpha \in A} \big(M_n(\hat{p}_{n,\alpha}) + \mathrm{pen}_n(\alpha) \big)$

Final estimator $\hat{p}_n = \hat{p}_{n,\hat{\alpha}_n}$

Adaptation-Penalization

Models $\mathcal{P}_{n,\alpha}, \qquad \alpha \in A$

Estimator given model $\hat{p}_{n,\alpha} = \operatorname*{argmin}_{p \in \mathcal{P}_{n,\alpha}} M_n(p)$

Estimator model $\hat{\alpha}_n = \operatorname*{argmin}_{\alpha \in A} \big(M_n(\hat{p}_{n,\alpha}) + \mathrm{pen}_n(\alpha) \big)$

Final estimator $\hat{p}_n = \hat{p}_{n,\hat{\alpha}_n}$

If M_n is the log likelihood, then $\hat{p}_n$ is the posterior mode relative to prior $\pi_n(p, \alpha) \propto \exp\big(\mathrm{pen}_n(\alpha)\big)$

Adaptation-Bayesian

Models	$\mathcal{P}_{n,\alpha}$, $\qquad \alpha \in A$
Prior	$\Pi_{n,\alpha}$ on $\mathcal{P}_{n,\alpha}$
Prior	$(\lambda_{n,\alpha})_{\alpha \in A}$ on A
Overall Prior	$\Pi_n = \sum_{\alpha \in A} \lambda_{n,\alpha} \Pi_{n,\alpha}$

Adaptation-Bayesian

Models $\mathcal{P}_{n,\alpha}$, $\alpha \in A$

Prior $\Pi_{n,\alpha}$ on $\mathcal{P}_{n,\alpha}$

Prior $(\lambda_{n,\alpha})_{\alpha \in A}$ on A

Overall Prior $\Pi_n = \sum_{\alpha \in A} \lambda_{n,\alpha} \Pi_{n,\alpha}$

Posterior $B \mapsto \Pi_n(B|X_1, \ldots, X_n)$,

$$\Pi_n(B|X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(p)}{\int \prod_{i=1}^n p(X_i)\, d\Pi_n(p)}$$

$$= \frac{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}: p \in B} \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)}{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}} \prod_{i=1}^n p(X_i)\, d\Pi_{n,\alpha}(p)}.$$

Adaptation-Bayesian

Models	$\mathcal{P}_{n,\alpha}, \qquad \alpha \in A$
Prior	$\Pi_{n,\alpha}$ on $\mathcal{P}_{n,\alpha}$
Prior	$(\lambda_{n,\alpha})_{\alpha\in A}$ on A
Overall Prior	$\Pi_n = \sum_{\alpha\in A} \lambda_{n,\alpha}\Pi_{n,\alpha}$

Posterior $\quad B \mapsto \Pi_n(B|X_1,\ldots,X_n)$

Desired result:

If $p_0 \in \mathcal{P}_{n,\beta}$ (or is close) then

$E_{p_0}\Pi_n\big(p : d(p,p_0) \geq M_n\,\epsilon_{n,\beta}|X_1,\ldots,X_n\big) \to 0$ for every $M_n \to \infty$

Single Model

Model $\mathcal{P}_{n,\beta}$

Prior $\Pi_{n,\beta}$

THEOREM (GGvdV, 2000) If

$$\log N(\epsilon_{n,\beta}, \mathcal{P}_{n,\beta}, d) \leq En\epsilon_{n,\beta}^2 \qquad \text{entropy}$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big) \geq e^{-Fn\epsilon_{n,\beta}^2} \qquad \text{prior mass}$$

then the posterior rate of convergence is $\epsilon_{n,\beta}$

Single Model

Model $\mathcal{P}_{n,\beta}$

Prior $\Pi_{n,\beta}$

THEOREM (GGvdV, 2000) If

$$\log N(\epsilon_{n,\beta}, \mathcal{P}_{n,\beta}, d) \le E n \epsilon_{n,\beta}^2 \qquad \text{entropy}$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big) \ge e^{-Fn\epsilon_{n,\beta}^2} \qquad \text{prior mass}$$

then the posterior rate of convergence is $\epsilon_{n,\beta}$

$B_{n,\alpha}(\epsilon)$ is a Kullback-Leibler ball around p_0:

$$B_{n,\alpha}(\epsilon) = \left\{p \in \mathcal{P}_{n,\alpha} : -P_0 \log \frac{p}{p_0} \le \epsilon^2, P_0\left(\log \frac{p}{p_0}\right)^2 \le \epsilon^2\right\}$$

Covering Numbers

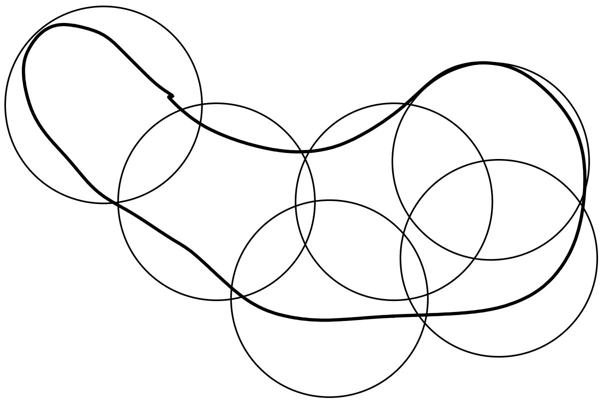
DEFINITION

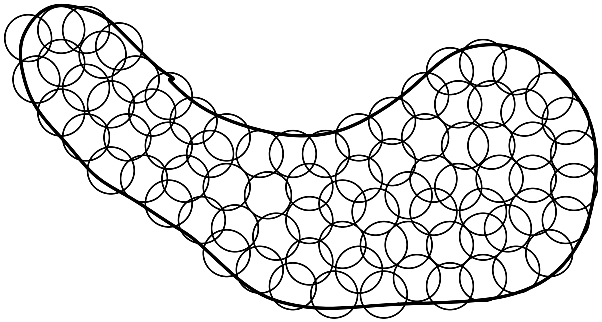
The covering number $N(\epsilon, \mathcal{P}, d)$ is the minimal number of balls of radius ϵ needed to cover the set $\mathcal{P}$.

Covering Numbers

DEFINITION

The covering number $N(\epsilon, \mathcal{P}, d)$ is the minimal number of balls of radius ϵ needed to cover the set $\mathcal{P}$.

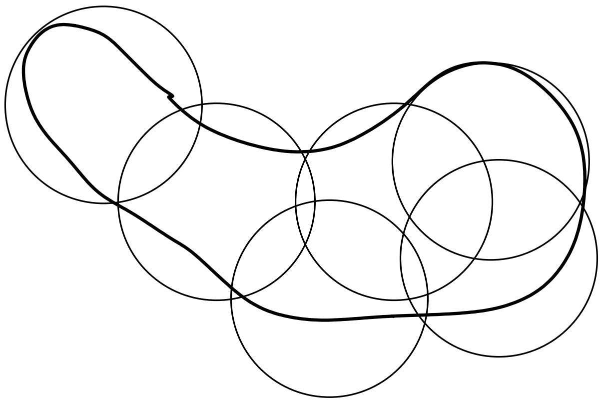


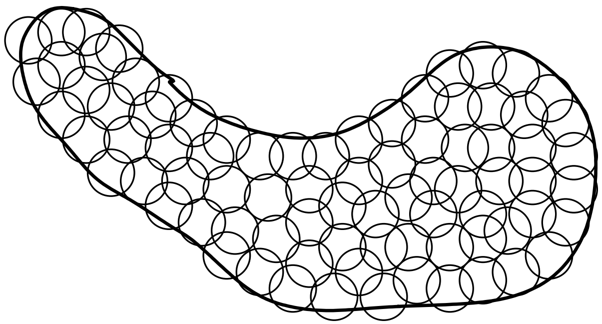


Covering Numbers

DEFINITION

The covering number $N(\epsilon, \mathcal{P}, d)$ is the minimal number of balls of radius ϵ needed to cover the set $\mathcal{P}$.





Rate at which $N(\epsilon, \mathcal{P}, d)$ increases if $\epsilon \downarrow 0$ determines size of model

Parametric model $(1/\epsilon)^d$

Nonparametric model $e^{(1/\epsilon)^{1/\alpha}}$ e.g. smoothness α

Motivation Entropy

Solution ϵ_n to

$$\log N(\epsilon, \mathcal{P}_n, d) \propto n\epsilon^2$$

gives optimal rate of convergence for model $\mathcal{P}_n$ in minimax sense

Le Cam (1975, 1986), Birgé (1983), Barron and Yang (1999)

Single Model

Model $\mathcal{P}_{n,\beta}$

Prior $\Pi_{n,\beta}$

THEOREM (GGvdV, 2000) If

$$\log N(\epsilon_{n,\beta}, \mathcal{P}_{n,\beta}, d) \leq E n \epsilon_{n,\beta}^2 \qquad \text{entropy}$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big) \geq e^{-F n \epsilon_{n,\beta}^2} \qquad \text{prior mass}$$

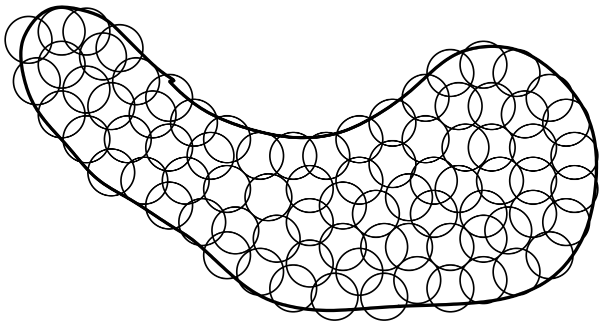
then the posterior rate of convergence is $\epsilon_{n,\beta}$

Motivation Prior Mass

$$\Pi_n\Big(B_n(\epsilon_n)\Big) \geq e^{-n\epsilon_n^2} \qquad \text{prior mass}$$

Motivation Prior Mass

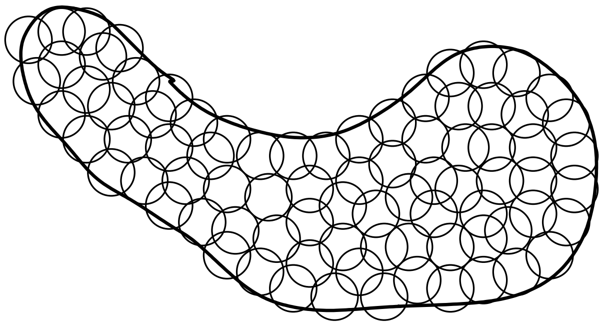
$$\Pi_n\big(B_n(\epsilon_n)\big) \geq e^{-n\epsilon_n^2}$$ prior mass



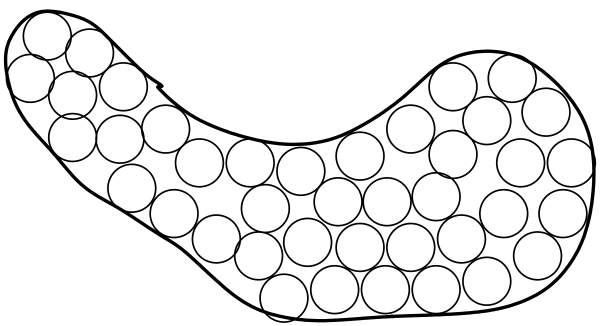
Need $N(\epsilon, \mathcal{P}, d) \approx \exp(n\epsilon_n^2)$ balls

Motivation Prior Mass

$$\Pi_n\Big(B_n(\epsilon_n)\Big) \geq e^{-n\epsilon_n^2}$$ prior mass



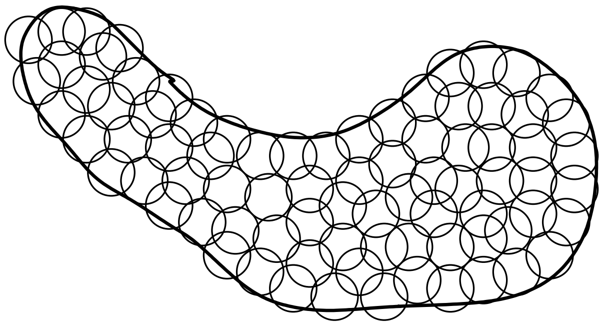
Need $N(\epsilon, \mathcal{P}, d) \approx \exp(n\epsilon_n^2)$ balls



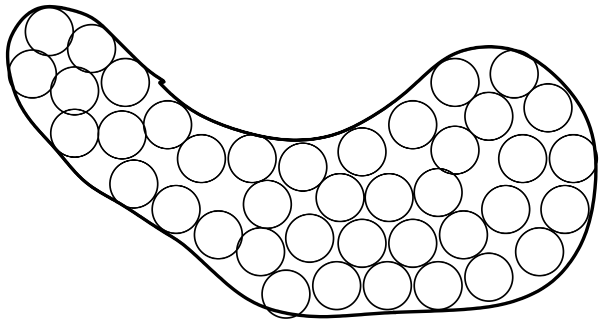
Can place $\exp(Cn\epsilon_n^2)$ balls

Motivation Prior Mass

$$\Pi_n\big(B_n(\epsilon_n)\big) \geq e^{-n\epsilon_n^2} \qquad \text{prior mass}$$



Need $N(\epsilon, \mathcal{P}, d) \approx \exp(n\epsilon_n^2)$ balls



Can place $\exp(Cn\epsilon_n^2)$ balls

If Π_n "uniform", then each ball receives mass $\exp(-Cn\epsilon_n^2)$

Equivalence KL and Hellinger

The prior mass condition uses Kullback-Leibler balls, whereas the entropy condition uses d-balls

These are typically (almost) equivalent

Equivalence KL and Hellinger

The prior mass condition uses Kullback-Leibler balls, whereas the entropy condition uses d-balls

These are typically (almost) equivalent

- If ratios p_0/p of densities are bounded, then fully equivalent

Equivalence KL and Hellinger

The prior mass condition uses Kullback-Leibler balls, whereas the entropy condition uses d-balls

These are typically (almost) equivalent

- If ratios p_0/p of densities are bounded, then fully equivalent

- If $P_0(p_0/p)^b$ is bounded, some $b > 0$, then equivalent up to logarithmic factors

Single Model

Model $\mathcal{P}_{n,\beta}$

Prior $\Pi_{n,\beta}$

THEOREM (GGvdV, 2000) If

$$\log N(\epsilon_{n,\beta}, \mathcal{P}_{n,\beta}, d) \leq En\epsilon_{n,\beta}^2 \qquad \text{entropy}$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big) \geq e^{-Fn\epsilon_{n,\beta}^2} \qquad \text{prior mass}$$

then the posterior rate of convergence is $\epsilon_{n,\beta}$

Single Model

Model $\mathcal{P}_{n,\beta}$

Prior $\Pi_{n,\beta}$

THEOREM (GGvdV, 2000) If

$$\log N(\epsilon_{n,\beta}, \mathcal{P}_{n,\beta}, d) \leq E n \epsilon_{n,\beta}^2 \qquad \text{entropy}$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big) \geq e^{-F n \epsilon_{n,\beta}^2} \qquad \text{prior mass}$$

then the posterior rate of convergence is $\epsilon_{n,\beta}$

Can actually replace entropy $\log N(\epsilon, \mathcal{P}_{n,\beta}, d)$ by Le Cam dimension $\sup_{\eta>\epsilon} \log N\big(\eta/2, C_{n,\beta}(\eta), d\big)$
Can also refine the prior mass condition

Adaptation-Bayesian

Models	$\mathcal{P}_{n,\alpha}$, $\qquad \alpha \in A$
Prior	$\Pi_{n,\alpha}$ on $\mathcal{P}_{n,\alpha}$
Prior	$(\lambda_{n,\alpha})_{\alpha\in A}$ on A
Overall Prior	$\sum_{\alpha\in A} \lambda_{n,\alpha}\Pi_{n,\alpha}$

Posterior $\qquad B \mapsto \Pi_n(B|X_1,\ldots,X_n)$

Desired result:
If $p_0 \in \mathcal{P}_{n,\beta}$ (or is close) then
$\mathrm{E}_{p_0}\Pi_n\bigl(p: d(p,p_0) \geq M\epsilon_{n,\beta_n}|X_1,\ldots,X_n\bigr) \to 0$ for every sufficiently large M.

Adaptation (1)

A finite, ordered $\qquad \epsilon_{n,\alpha} \ll \epsilon_{n,\beta}$ if $\alpha \geq \beta$

$n\epsilon^2_{n,\beta} \to \infty$

Adaptation (1)

A finite, ordered $\qquad \epsilon_{n,\alpha} \ll \epsilon_{n,\beta}$ if $\alpha \geq \beta$

$n\epsilon_{n,\beta}^2 \to \infty$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

Adaptation (1)

A finite, ordered $\epsilon_{n,\alpha} \ll \epsilon_{n,\beta}$ if $\alpha \geq \beta$

$n\epsilon_{n,\beta}^2 \to \infty$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

Small models get big weights

Adaptation (1)

A finite, ordered $\epsilon_{n,\alpha} \ll \epsilon_{n,\beta}$ if $\alpha \geq \beta$

$n\epsilon_{n,\beta}^2 \to \infty$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

THEOREM If

$$\log N(\epsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, d) \leq En\epsilon_{n,\alpha}^2 \qquad \text{entropy}, \forall \alpha.$$

$$\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big) \geq e^{-Fn\epsilon_{n,\beta}^2} \qquad \text{prior mass}$$

then posterior rate is $\epsilon_{n,\beta}$

Adaptation (2)

Extension to countable A possible in two ways:

- truncation of weights $\lambda_{n,\alpha}$ to subsets $A_n \uparrow A$
- additional entropy control

Adaptation (2)

Extension to countable A possible in two ways:

- truncation of weights $\lambda_{n,\alpha}$ to subsets $A_n \uparrow A$
- additional entropy control

Also replace β by β_n

Adaptation (2)

Extension to countable A possible in two ways:

- truncation of weights $\lambda_{n,\alpha}$ to subsets $A_n \uparrow A$
- additional entropy control

Also replace β by β_n

Always assume
$\sum_\alpha (\lambda_\alpha / \lambda_{\beta_n}) \exp(-C\epsilon_{n,\alpha}^2/4) = O(1)$

Adaptation (2a)-Truncation

$$A_n \uparrow A, \quad \beta_n \in A_n, \quad \log(\# A_n) \leq n\epsilon^2_{n,\beta_n}$$

$$n\epsilon^2_{n,\beta_n} \to \infty$$

Adaptation (2a)-Truncation

$A_n \uparrow A, \quad \beta_n \in A_n, \quad \log(\# A_n) \leq n\epsilon_{n,\beta_n}^2$

$n\epsilon_{n,\beta_n}^2 \longrightarrow \infty$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2} 1_{A_n}(\alpha)$

Adaptation (2a)-Truncation

$A_n \uparrow A$, $\quad \beta_n \in A_n$, $\quad \log(\# A_n) \le n\epsilon^2_{n,\beta_n}$

$n\epsilon^2_{n,\beta_n} \to \infty$

$$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon^2_{n,\alpha}} 1_{A_n}(\alpha)$$

$$\max_{\alpha \in A_n : \epsilon^2_{n,\alpha} \le H\epsilon^2_{n,\beta_n}} E_\alpha \frac{\epsilon^2_{n,\alpha}}{\epsilon^2_{n,\beta_n}} = O(1), \qquad H \gg 1$$

Adaptation (2a)-Truncation

$A_n \uparrow A$, $\quad \beta_n \in A_n$, $\quad \log(\# A_n) \leq n\epsilon_{n,\beta_n}^2$

$n\epsilon_{n,\beta_n}^2 \to \infty$

$$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2} 1_{A_n}(\alpha)$$

$$\max_{\alpha \in A_n : \epsilon_{n,\alpha}^2 \leq H\epsilon_{n,\beta_n}^2} E_\alpha \frac{\epsilon_{n,\alpha}^2}{\epsilon_{n,\beta_n}^2} = O(1), \qquad H \gg 1$$

THEOREM If

$$\log N(\epsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, d) \leq En\epsilon_{n,\alpha}^2 \qquad \text{entropy}, \forall \alpha.$$

$$\Pi_{n,\beta_n}\big(B_{n,\beta_n}(\epsilon_{n,\beta_n})\big) \geq e^{-Fn\epsilon_{n,\beta_n}^2} \qquad \text{prior mass}$$

then posterior rate is ϵ_{n,β_n}

Adaptation (2b)-Entropy control

A countable

$n\epsilon_{n,\beta}^2 \longrightarrow \infty$

Adaptation (2b)-Entropy control

A countable

$n\epsilon_{n,\beta}^2 \longrightarrow \infty$

$$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$$

Adaptation (2b)-Entropy control

A countable

$n\epsilon_{n,\beta}^2 \to \infty$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

THEOREM If $H \gg 1$ and

$$\log N\Big(\epsilon_{n,\beta_n}, \bigcup_{\alpha:\epsilon_{n,\alpha} \le H\epsilon_{n,\beta_n}} \mathcal{P}_{n,\alpha}, d\Big) \le En\epsilon_{n,\beta_n}^2, \qquad \text{entropy},$$

$$\Pi_{n,\beta_n}\big(B_{n,\beta_n}(\epsilon_{n,\beta_n})\big) \ge e^{-Fn\epsilon_{n,\beta_n}^2}, \qquad \text{prior mass}$$

then posterior rate is ϵ_{n,β_n}

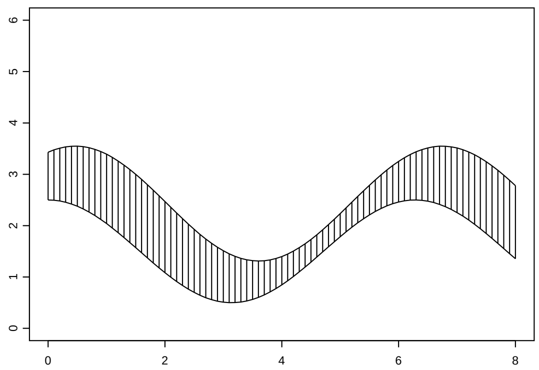
Discrete priors

Discrete priors that are uniform on
specially constructed approximating sets
are universal
in the sense that under abstract and mild conditions
they give the desired result

To avoid unnecessary logarithmic factors we need to
replace ordinary entropy by the slightly more restrictive
bracketing entropy

Bracketing Numbers

Given $l, u : \mathcal{X} \to \mathbb{R}$ the bracket $[l, u]$ is the set of $p : \mathcal{X} \to \mathbb{R}$ with $l \leq p \leq u$.



An ϵ-bracket relative to d is a bracket $[l, u]$ with $d(u, l) < \epsilon$.

DEFINITION
The bracketing number $N_{[\,]}(\epsilon, \mathcal{P}, d)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{P}$.

Discrete priors

$\mathcal{Q}_{n,\alpha}$ collection of nonnegative functions with

$$\log N_{]}(\epsilon_{n,\alpha}, \mathcal{Q}_{n,\alpha}, h) \leq E_\alpha n \epsilon_{n,\alpha}^2$$

$u_1, \ldots, u_N$ minimal set of $\epsilon_{n,\alpha}$-upper brackets
$\tilde{u}_1, \ldots, \tilde{u}_N$ normalized functions

Discrete priors

$\mathcal{Q}_{n,\alpha}$ collection of nonnegative functions with

$$\log N_{]}(\epsilon_{n,\alpha}, \mathcal{Q}_{n,\alpha}, h) \leq E_\alpha n\epsilon_{n,\alpha}^2$$

$u_1, \ldots, u_N$ minimal set of $\epsilon_{n,\alpha}$-upper brackets
$\tilde{u}_1, \ldots, \tilde{u}_N$ normalized functions

Prior $\Pi_{n,\alpha}$ uniform on $\tilde{u}_1, \ldots, \tilde{u}_N$
Model $\cup_{M>0} M\mathcal{Q}_{n,\alpha}$

Discrete priors

$\mathcal{Q}_{n,\alpha}$ collection of nonnegative functions with

$$\log N_{]}(\epsilon_{n,\alpha}, \mathcal{Q}_{n,\alpha}, h) \leq E_\alpha n \epsilon_{n,\alpha}^2$$

$u_1, \ldots, u_N$ minimal set of $\epsilon_{n,\alpha}$-upper brackets
$\tilde{u}_1, \ldots, \tilde{u}_N$ normalized functions

Prior $\Pi_{n,\alpha}$ uniform on $\tilde{u}_1, \ldots, \tilde{u}_N$
Model $\cup_{M>0} M \mathcal{Q}_{n,\alpha}$

THEOREM
If $\lambda_{n,\alpha}$ and $A_n \uparrow A$ are as before, and $p_0 \in M_0 \mathcal{Q}_{n,\beta}$
then posterior rate is $\epsilon_{n,\beta}$, relative to the Hellinger distance.

Smoothness Spaces

$\mathbb{B}_1^\alpha$ unit ball in a Banach $\mathbb{B}^\alpha$ of functions

$\log N_]\bigl(\epsilon_{n,\alpha}, \mathbb{B}_1^\alpha, \|\cdot\|_2\bigr) \le E_\alpha n \epsilon_{n,\alpha}^2$

Model $\sqrt{p} \in \mathbb{B}^\alpha$

Smoothness Spaces

$\mathbb{B}_1^\alpha$ unit ball in a Banach $\mathbb{B}^\alpha$ of functions

$\log N_{]}\big(\epsilon_{n,\alpha}, \mathbb{B}_1^\alpha, \|\cdot\|_2\big) \le E_\alpha n \epsilon_{n,\alpha}^2$

Model $\quad \sqrt{p} \in \mathbb{B}^\alpha$

THEOREM
There exists a prior such that the posterior rate is $\epsilon_{n,\beta}$ whenever $\sqrt{p_0} \in \mathbb{B}^\beta$ for some $\beta > 0$.

Smoothness Spaces

$\mathbb{B}_1^\alpha$ unit ball in a Banach $\mathbb{B}^\alpha$ of functions
$\log N_] \big(\epsilon_{n,\alpha}, \mathbb{B}_1^\alpha, \| \cdot \|_2\big) \le E_\alpha n \epsilon_{n,\alpha}^2$

Model $\quad \sqrt{p} \in \mathbb{B}^\alpha$

THEOREM
There exists a prior such that the posterior rate is $\epsilon_{n,\beta}$ whenever $\sqrt{p_0} \in \mathbb{B}^\beta$ for some $\beta > 0$.

EXAMPLE
- Hölder spaces and Sobolev spaces of α-smooth functions, with $\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$.
- Besov spaces (in progress)

Finite-Dimensional Models

Model $\mathcal{P}_J$ of dimension J

Finite-Dimensional Models

Model	$\mathcal{P}_J$ of dimension J
Bias	p_0 β-regular if $d(p_0, \mathcal{P}_J) \lesssim (1/J)^\beta$
Variance	Precision when estimating J parameters J/n

Finite-Dimensional Models

Model $\mathcal{P}_J$ of dimension J

Bias p_0 β-regular if $d(p_0, \mathcal{P}_J) \lesssim (1/J)^\beta$
Variance Precision when estimating J parameters J/n

Bias-variance trade-off $(1/J)^{2\beta} \sim J/n$

Optimal dimension $J \sim n^{1/(2\beta+1)}$
Rate $\epsilon_{n,J} \sim n^{-\beta/(2\beta+1)}$

Finite-Dimensional Models

Model $\mathcal{P}_J$ of dimension J

Bias p_0 β-regular if $d(p_0, \mathcal{P}_J) \lesssim (1/J)^\beta$
Variance Precision when estimating J parameters J/n

Bias-variance trade-off $(1/J)^{2\beta} \sim J/n$

Optimal dimension $J \sim n^{1/(2\beta+1)}$
Rate $\epsilon_{n,J} \sim n^{-\beta/(2\beta+1)}$ We want to adapt

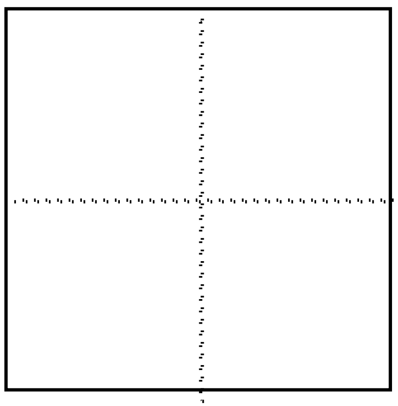
to β by putting weights on J

Finite-Dimensional Models

Model $\mathcal{P}_J$ of dimension J

Model dimension can be taken as Le Cam dimension

$$J \sim \sup_{\eta > \epsilon} \log N\Big(\eta/2, \{p \in \mathcal{P}_J : d(p, p_0) < \eta\}, d\Big)$$



dimension 2

Finite-Dimensional Models

Models $\mathcal{P}_{J,M}$ of Le Cam dimension $A_M J$, $J \in \mathbb{N}$, $M \in \mathcal{M}$,

Prior $\Pi_{J,M}\big(B_{J,M}(\epsilon)\big) \geq \big(B_J C_M \epsilon\big)^J$, $\epsilon > D_M\, d(p_0, \mathcal{P}_{J,M})$

Finite-Dimensional Models

Models $\mathcal{P}_{J,M}$ of Le Cam dimension $A_M J$, $J \in \mathbb{N}$, $M \in \mathcal{M}$,

Prior $\Pi_{J,M}\big(B_{J,M}(\epsilon)\big) \geq \big(B_J C_M \epsilon\big)^J$, $\epsilon > D_M\, d(p_0, \mathcal{P}_{J,M})$

This correspond to a smooth prior on the J-dimensional model

Finite-Dimensional Models

Models $\mathcal{P}_{J,M}$ of Le Cam dimension $A_M J$, $J \in \mathbb{N}$, $M \in \mathcal{M}$,

Prior $\Pi_{J,M}\big(B_{J,M}(\epsilon)\big) \geq \big(B_J C_M \epsilon\big)^J$, $\quad \epsilon > D_M\, d(p_0, \mathcal{P}_{J,M})$

Weights $\lambda_{n,J,M} \propto e^{-Cn\epsilon_{n,J,M}^2} 1_{\mathcal{J}_n \times \mathcal{M}_n}(J, M)$

Finite-Dimensional Models

Models $\mathcal{P}_{J,M}$ of Le Cam dimension $A_M J$, $J \in \mathbb{N}$, $M \in \mathcal{M}$,

Prior $\Pi_{J,M}\big(B_{J,M}(\epsilon)\big) \geq \big(B_J C_M \epsilon\big)^J$, $\quad \epsilon > D_M\, d(p_0, \mathcal{P}_{J,M})$

Weights $\lambda_{n,J,M} \propto e^{-Cn\epsilon_{n,J,M}^2} 1_{\mathcal{J}_n \times \mathcal{M}_n}(J, M)$

$(\log C_M) A_M \gg 1$, $B_J \gtrsim J^{-k}$, $\sum_{M \in \mathcal{M}} e^{-HA_M} < \infty$

$$\epsilon_{n,J,M} = \sqrt{\frac{J \log n}{n} A_M}$$

THEOREM

If there exist $J_n \in \mathcal{J}_n$ with $J_n \leq n$ and $d(p_0, \mathcal{P}_{n,J_n,M_0}) \lesssim \epsilon_{n,J_n,M_0}$, then posterior rate is ϵ_{n,J_n,M_0}

Finite-Dimensional Models: Examples

If $p_0 \in \mathcal{P}_{J_0,M_0}$ for some J_0, then rate $\sqrt{(\log n)/n}$.

Finite-Dimensional Models: Examples

If $p_0 \in \mathcal{P}_{J_0,M_0}$ for some J_0, then rate $\sqrt{(\log n)/n}$.

If $d(p_0, \mathcal{P}_{J,M_0}) \lesssim J^{-\beta}$ for every J, rate $(n/\log n)^{-\beta/(2\beta+1)}$.

Finite-Dimensional Models: Examples

If $p_0 \in \mathcal{P}_{J_0,M_0}$ for some J_0, then rate $\sqrt{(\log n)/n}$.

If $d(p_0, \mathcal{P}_{J,M_0}) \lesssim J^{-\beta}$ for every J, rate $(n/\log n)^{-\beta/(2\beta+1)}$.

If $d(p_0, \mathcal{P}_{J,M_0}) \lesssim e^{-J^\beta}$ for every J, then rate $(\log n)^{1/\beta+1/2}/\sqrt{n}$.

Finite-Dimensional Models: Examples

If $p_0 \in \mathcal{P}_{J_0,M_0}$ for some J_0, then rate $\sqrt{(\log n)/n}$.

If $d(p_0, \mathcal{P}_{J,M_0}) \lesssim J^{-\beta}$ for every J, rate $(n/\log n)^{-\beta/(2\beta+1)}$.

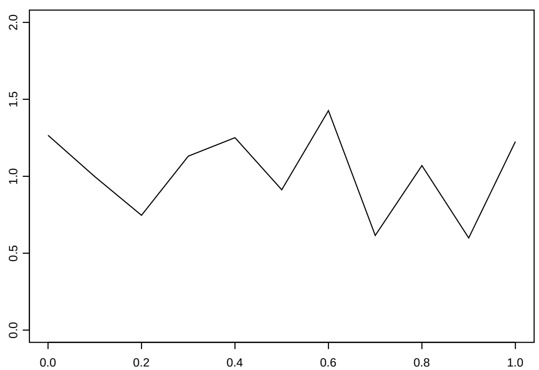
If $d(p_0, \mathcal{P}_{J,M_0}) \lesssim e^{-J^\beta}$ for every J, then rate $(\log n)^{1/\beta+1/2}/\sqrt{n}$.

Can logarithmic factors be avoided?
By using different weights and/or different model priors?

Splines

$$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$$

Spline of order q is continuous function $f : [0,1] \to \mathbb{R}$ with
• $q-2$ times differentiable on $[0,1)$
• restriction to every $\big[(k-1)/K, k/K\big)$ is a polynomial of degree $< q$.

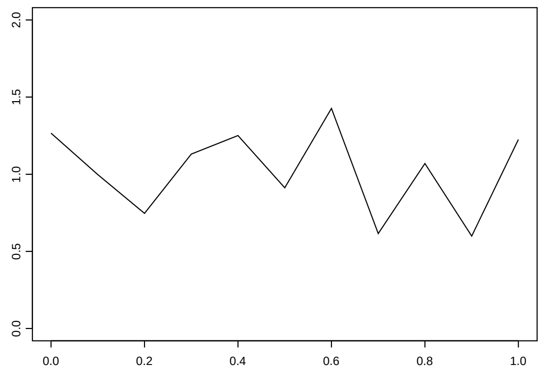


linear spine

Splines

$$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$$

Spline of order q is continuous function $f : [0,1] \to \mathbb{R}$ with
- $q-2$ times differentiable on $[0,1)$
- restriction to every $\big[(k-1)/K, k/K\big)$ is a polynomial of degree $< q$.



linear spine

Splines form a $J = q + K - 1$-dimensional vector space
Convenient basis B-splines $B_{J,1}, \ldots, B_{J,J}$

Splines-Properties

$$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$$

$$\theta^T B_J = \textstyle\sum_j \theta_j B_{J,j} \qquad \theta \in \mathbb{R}^J, \qquad J = K + q - 1$$

Approximation of smooth functions

If $q \geq \alpha > 0$ and f in $C^\alpha[0,1]$, then

$$\inf_{\theta \in \mathbb{R}^J} \|\theta^T B_J - f\|_\infty \leq C_{q,\alpha} \Big(\frac{1}{J}\Big)^\alpha \|f\|_\alpha$$

Equivalence of norms

For any $\theta \in \mathbb{R}^J$,

$$\|\theta\|_\infty \lesssim \|\theta^T B_J\|_\infty \leq \|\theta\|_\infty,$$

$$\|\theta\|_2 \lesssim \sqrt{J}\, \|\theta^T B_J\|_2 \lesssim \|\theta\|_2.$$

Log Spline Models

$$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$$
$$\theta^T B_J = \sum_j \theta_j B_{J,j}, \qquad J = K + q - 1$$

$$p_{J,\theta}(x) = e^{\theta^T B_J(x) - c_J(\theta)}, \qquad e^{c_J(\theta)} = \int_0^1 e^{\theta^T B_J(x)}\, dx.$$

Log Spline Models

$$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$$
$$\theta^T B_J = \sum_j \theta_j B_{J,j}, \qquad J = K + q - 1$$

$$p_{J,\theta}(x) = e^{\theta^T B_J(x) - c_J(\theta)}, \qquad e^{c_J(\theta)} = \int_0^1 e^{\theta^T B_J(x)}\, dx.$$

prior on θ induces prior on $p_{J,\theta}$ for fixed J
prior on J give model weights $\lambda_{n,J}$

Log Spline Models

$$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$$
$$\theta^T B_J = \textstyle\sum_j \theta_j B_{J,j}, \qquad J = K + q - 1$$

$$p_{J,\theta}(x) = e^{\theta^T B_J(x) - c_J(\theta)}, \qquad e^{c_J(\theta)} = \int_0^1 e^{\theta^T B_J(x)}\, dx.$$

prior on θ induces prior on $p_{J,\theta}$ for fixed J
prior on J give model weights $\lambda_{n,J}$

flat prior on θ and model weights $\lambda_{n,J}$ as before gives adaptation to smoothness classes
up to logarithmic factor

Log Spline Models

$[0,1) = \cup_{k=1}^{K} \big[(k-1)/K, k/K\big)$
$\theta^T B_J = \sum_j \theta_j B_{J,j}, \qquad J = K + q - 1$

$$p_{J,\theta}(x) = e^{\theta^T B_J(x) - c_J(\theta)}, \qquad e^{c_J(\theta)} = \int_0^1 e^{\theta^T B_J(x)}\, dx.$$

prior on θ induces prior on $p_{J,\theta}$ for fixed J
prior on J give model weights $\lambda_{n,J}$

flat prior on θ and model weights $\lambda_{n,J}$ as before gives adaptation to smoothness classes
up to logarithmic factor

Can do better?

Adaptation (3)

A finite, ordered $\qquad \epsilon_{n,\alpha} < \epsilon_{n,\beta}$ if $\alpha > \beta$

$n\epsilon_{n,\alpha}^2 \to \infty$ for every α

THEOREM If

$$\sup_{\epsilon \geq \epsilon_{n,\alpha}} \log N\big(\epsilon/2, C_{n,\alpha}(\epsilon), d\big) \leq E n\epsilon_{n,\alpha}^2, \qquad \alpha \in A,$$

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}} \frac{\Pi_{n,\alpha}\big(C_{n,\alpha}(B\epsilon_{n,\alpha})\big)}{\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big)} = o\big(e^{-2n\epsilon_{n,\beta}^2}\big), \qquad \alpha < \beta,$$

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta}} \frac{\Pi_{n,\alpha}\big(C_{n,\alpha}(i\epsilon_{n,\alpha})\big)}{\Pi_{n,\beta}\big(B_{n,\beta}(\epsilon_{n,\beta})\big)} \leq e^{i^2 n(\epsilon_{n,\alpha}^2 \vee \epsilon_{n,\beta}^2)},$$

then posterior rate is $\epsilon_{n,\beta}$

$B_{n,\alpha}(\epsilon)$ and $C_{n,\alpha}(\epsilon)$ are KL-ball and d-ball in $\mathcal{P}_{n,\alpha}$ around p_0

Log Spline Models

Consider four combinations of
priors $\bar{\Pi}_{n,\alpha}$ on θ
weights $\lambda_{n,\alpha}$ on $J_{n,\alpha}$
to adapt to smoothness classes

$$J_{n,\alpha} \sim n^{1/(2\alpha+1)}$$
$$\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$$

Assume p_0 is β-smooth and sufficiently regular

Flat prior, uniform weights

$\bar{\Pi}_{n,\alpha}$ "uniform" on $[-M, M]^{J_{n,\alpha}}$, M large

Uniform weights $\lambda_{n,\alpha} = \lambda_\alpha$

Flat prior, uniform weights

$\bar{\Pi}_{n,\alpha}$ "uniform" on $[-M, M]^{J_{n,\alpha}}$, M large

Uniform weights $\lambda_{n,\alpha} = \lambda_\alpha$

THEOREM Posterior rate is $\epsilon_{n,\beta}\sqrt{\log n}$

Flat prior, decreasing weights

$\bar{\Pi}_{n,\alpha}$ "uniform" on $[-M, M]^{J_{n,\alpha}}$, M large

$\lambda_{n,\alpha} \propto \prod_{\gamma<\alpha}(C\epsilon_{n,\gamma})^{J_{n,\gamma}}$, $C > 1$

THEOREM Posterior rate is $\epsilon_{n,\beta}$

Flat prior, decreasing weights

$\bar{\Pi}_{n,\alpha}$ "uniform" on $[-M, M]^{J_{n,\alpha}}$, $\qquad M$ large

$\lambda_{n,\alpha} \propto \prod_{\gamma<\alpha} (C\epsilon_{n,\gamma})^{J_{n,\gamma}}$, $\qquad C > 1$

THEOREM Posterior rate is $\epsilon_{n,\beta}$

Small models get small weight!

Discrete priors, increasing weights

$\bar{\Pi}_{n,\alpha}$ discrete on $\mathbb{R}^J$ with minimal number of support points to obtain approximation error $\epsilon_{n,\alpha}$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

THEOREM   Posterior rate is $\epsilon_{n,\beta}$

Discrete priors, increasing weights

$\bar{\Pi}_{n,\alpha}$ discrete on $\mathbb{R}^J$ with minimal number of support points to obtain approximation error $\epsilon_{n,\alpha}$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

THEOREM Posterior rate is $\epsilon_{n,\beta}$

Small models get big weight!

Discrete priors, increasing weights

$\bar{\Pi}_{n,\alpha}$ discrete on $\mathbb{R}^J$ with minimal number of support points to obtain approximation error $\epsilon_{n,\alpha}$

$\lambda_{n,\alpha} \propto \lambda_\alpha e^{-Cn\epsilon_{n,\alpha}^2}$

THEOREM   Posterior rate is $\epsilon_{n,\beta}$

Splines of dimension $J_{n,\alpha}$ give approximation error $\epsilon_{n,\alpha}$.
A uniform grid on coefficients in dimension $J_{n,\alpha}$ that gives approximation error $\epsilon_{n,\alpha}$ is too large. Need sparse subset.
Similarly a smooth prior on coefficients in dimension $J_{n,\alpha}$ is too rich.

Special smooth prior, increasing weights

$\bar{\Pi}_{n,\alpha}$ continuous and uniform on minimal subset of $\mathbb{R}^J$ that allows approximation with error $\epsilon_{n,\alpha}$

Special, increasing weights $\lambda_{n,\alpha}$

THEOREM (Huang, 2002) Posterior rate is $\epsilon_{n,\beta}$

Huang obtains this result for the full scale of regularity spaces in a general finite-dimensional setting

Conclusion

There is a range of weights $\lambda_{n,\alpha}$ that works

Which weights $\lambda_{n,\alpha}$ work depends on the fine properties of the priors on the models $\mathcal{P}_{n,\alpha}$

Gaussian mixtures

Model $\quad p_{F,\sigma}(x) = \int \phi_\sigma(x-z)\,dF(z)$

Prior $\quad F \sim \mathrm{Dirichlet}(\alpha)$, $\sigma \sim \pi_n$, independent

(α Gaussian, π_n smooth)

Gaussian mixtures

Model $p_{F,\sigma}(x) = \int \phi_\sigma(x - z)\, dF(z)$

Prior $F \sim \mathrm{Dirichlet}(\alpha)$, $\sigma \sim \pi_n$, independent

(α Gaussian, π_n smooth)

CASE ss: π_n fixed

CASE s : π_n shrinks at rate $n^{-1/5}$

Gaussian mixtures

Model $p_{F,\sigma}(x) = \int \phi_\sigma(x-z)\,dF(z)$

Prior $F \sim \mathrm{Dirichlet}(\alpha)$, $\sigma \sim \pi_n$, independent

(α Gaussian, π_n smooth)

CASE ss: π_n fixed

CASE s : π_n shrinks at rate $n^{-1/5}$

THEOREM (Ghosal,vdV) Rate of convergence relative to (truncated) Hellinger distance is

- CASE ss: if $p_0 = p_{\sigma_0,F_0}$, then $(\log n)^k/\sqrt{n}$
- CASE s: if p_0 is 2-smooth, then $n^{-2/5}(\log n)^2$

Assume p_0 subGaussian

Gaussian mixtures

Model $\qquad p_{F,\sigma}(x) = \int \phi_\sigma(x-z)\,dF(z)$

Prior $\qquad F \sim \mathrm{Dirichlet}(\alpha)$, $\sigma \sim \pi_n$, independent

(α Gaussian, π_n smooth)

CASE ss: π_n fixed

CASE s : π_n shrinks at rate $n^{-1/5}$

THEOREM (Ghosal,vdV) Rate of convergence relative to (truncated) Hellinger distance is

- CASE ss: if $p_0 = p_{\sigma_0,F_0}$, then $(\log n)^k/\sqrt{n}$
- CASE s: if p_0 is 2-smooth, then $n^{-2/5}(\log n)^2$

Can we adapt to the two cases?

Gaussian mixtures

Weights $\lambda_{n,s}$ et $\lambda_{n,ss}$

Gaussian mixtures

Weights $\lambda_{n,s}$ et $\lambda_{n,ss}$

THEOREM
Adaptation up to logarithmic factors if

$$\exp\big(c(\log n)^k\big) < \frac{\lambda_{n,ss}}{\lambda_{n,s}} < \exp\big(Cn^{1/5}(\log n)^k\big)$$

Gaussian mixtures

Weights $\lambda_{n,s}$ et $\lambda_{n,ss}$

THEOREM
Adaptation up to logarithmic factors if

$$\exp\bigl(c(\log n)^k\bigr) < \frac{\lambda_{n,ss}}{\lambda_{n,s}} < \exp\bigl(Cn^{1/5}(\log n)^k\bigr)$$

We believe this works already if

$$\exp\bigl(-c(\log n)^k\bigr) < \frac{\lambda_{n,ss}}{\lambda_{n,s}} < \exp\bigl(Cn^{1/5}(\log n)^k\bigr)$$

In particular: equal weights.

Conclusion

There is a range of weights $\lambda_{n,\alpha}$ that works

Which weights $\lambda_{n,\alpha}$ work depends on the fine properties of the priors on the models $\mathcal{P}_{n,\alpha}$

Conclusion

There is a range of weights $\lambda_{n,\alpha}$ that works

Which weights $\lambda_{n,\alpha}$ work depends on the fine properties of the priors on the models $\mathcal{P}_{n,\alpha}$

This interaction makes comparison with penalized minimum contrast estimation difficult

Need refined asymptotics and numerical implementation for further understanding

Conclusion

There is a range of weights $\lambda_{n,\alpha}$ that works

Which weights $\lambda_{n,\alpha}$ work depends on the fine properties of the priors on the models $\mathcal{P}_{n,\alpha}$

This interaction makes comparison with penalized minimum contrast estimation difficult

Need refined asymptotics and numerical implementation for further understanding

Bayesian density estimation is 10 years behind?