Nonparametric Bayesian Methods - Lecture III

Harry van Zanten

Korteweg-de Vries Institute for Mathematics

University of Amsterdam

CRiSM Masterclass, April 4-6, 2016

Overview of Lecture III

- Recall: general rate of contraction results
- Gaussian process priors
  - reproducing kernel Hilbert space
  - small ball probabilities
  - concentration of measure
  - general theorem
- Rate results for concrete GP priors

General rate of contraction results

General contraction rate theorem - regression case - 1

Observations: pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$, where

$$Y_i = f(x_i) + e_i,$$

with fixed $x_1, \ldots, x_n \in \mathcal{X}$ and e_i i.i.d. $N(0, \sigma^2)$.

Object of interest: regression function $f : \mathcal{X} \to \mathbb{R}$. Element of class $\mathcal{F}$ of all such functions.

Norm on regression functions:

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

General contraction rate theorem - regression case - 2

Theorem.
If there exist $\mathcal{F}_n \subset \mathcal{F}$ and positive numbers ε_n such that $n\varepsilon_n^2 \to \infty$ and, for some $c > 0$,

$$\Pi(f : \|f - f_0\|_n \le \varepsilon_n)) \ge e^{-cn\varepsilon_n^2},$$

$$\Pi(\mathcal{F}_n^c) \le e^{-(c+8)n\varepsilon_n^2},$$

$$\log N(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_n) \le n{\varepsilon_n}^2,$$

then for $M > 0$ large enough

$$\Pi(f : \|f - f_0\|_n > M\varepsilon_n \mid Y_1, \dots, Y_n) \overset{P_{f_0}}{\to} 0.$$

Rates for Gaussian process priors

If $W = (W_x)_{x \in \mathcal{X}}$ is a stochastic process indexed by $\mathcal{X}$, we can use the law, or distribution of W as prior Π: for a set of functions $B \subset \mathcal{F}$,

$$\Pi(B) = \mathbb{P}(W \in B).$$

Convenient/popular choice: take W a Gaussian process (GP): Gaussian process prior (GP prior).

Examples:
If $\mathcal{X} = [0, 1]$: Brownian motion, (multiply) integrated Brownian motion, ...
If $\mathcal{X} = [0, 1]^d$: Matérn process, squared exponential GP, ...

Q: how do we verify the three conditions for GP's?

Gaussian process priors

What do we want to know about the GP?

For simplicity: take $\mathcal{X} = [0,1]$, bound $\|f\|_n$ by $\|f\|_\infty$,

$$\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|.$$

For a GP $W = (W_t : t \in [0,1])$, want to

- lower bound probabilities of the form $\mathbb{P}(\|W - f_0\|_\infty < \varepsilon)$,
- find sets of functions $\mathcal{F}$ such that
  - $\mathbb{P}(W \in \mathcal{F}^c)$ is exponentially small
  - $N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ is "small"

Gaussian processes

$\rightarrow$ reproducing kernel Hilbert space

GP's - linear functionals

Let $W = (W_t : t \in [0,1])$ be a *centered, continuous* Gaussian process with covariance function $r(s,t) = \mathbb{E}W_s W_t$. Say defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

(Can view W as a random element in $\mathbb{B} = C[0,1]$.)

Space $\mathcal{L}$ of linear functionals of W:

- $\mathcal{L}_0$: all finite linear combinations of the form $\sum c_i W_{t_i}$
- $\mathcal{L}$: closure of $\mathcal{L}_0$ in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

$\mathcal{L}$ is a separable Hilbert space with norm $\|L\| = \sqrt{\mathbb{E}L^2}$.

GP's - RKHS - 1

The reproducing kernel Hilbert space (RKHS) of W:

$$\mathbb{H} = \{t \mapsto \mathbb{E} W_t L : L \in \mathcal{L}\}.$$

Inner product:

$$\langle t \mapsto \mathbb{E} W_t L_1, t \mapsto \mathbb{E} W_t L_2 \rangle_{\mathbb{H}} = \mathbb{E} L_1 L_2.$$

We have isometry

$$\mathcal{L} \ni L \leftrightarrow (t \mapsto \mathbb{E} W_t L) \in \mathbb{H}.$$

In particular, $\mathbb{H}$ is a separable Hilbert space of functions.

GP's - RKHS - 1

The reproducing kernel Hilbert space (RKHS) of W:

$$\mathbb{H} = \{t \mapsto \mathbb{E} W_t L : L \in \mathcal{L}\}.$$

Inner product:

$$\langle t \mapsto \mathbb{E} W_t L_1, t \mapsto \mathbb{E} W_t L_2 \rangle_{\mathbb{H}} = \mathbb{E} L_1 L_2.$$

We have isometry

$$\mathcal{L} \ni L \leftrightarrow (t \mapsto \mathbb{E} W_t L) \in \mathbb{H}.$$

In particular, $\mathbb{H}$ is a separable Hilbert space of functions.

GP's - RKHS - 2

For fixed s, have that $t \mapsto r(s,t) = \mathbb{E} W_t W_s \in \mathbb{H}$ and for $h(t) = \mathbb{E} W_t L$, $L \in \mathcal{L}$,

$$\langle h, r(s,\cdot) \rangle_{\mathbb{H}} = \mathbb{E} L W_s = h(s).$$

This is the reproducing property.

We have $\mathbb{H} \subset \mathbb{B}$ and for $L \in \mathcal{L}$ and $h(t) = \mathbb{E} W_t L$,

$$\|h\|_\infty \leq \sup_{t\in[0,1]} \sqrt{\mathbb{E} W_t^2} \sqrt{\mathbb{E} L^2} = \sigma(W) \|h\|_{\mathbb{H}},$$

for $\sigma(W) = \sup_{t\in[0,1]} \sqrt{\mathbb{E} W_t^2}$.

Example: RKHS of Brownian motion

Let W be Brownian motion, with $r(s,t) = s \wedge t$. Then for all $s \in [0,1]$,

$$t \mapsto \int_0^t 1_{[0,s]}(u)\, du = t \wedge s = \mathbb{E} W_t W_s \in \mathbb{H}$$

and

$$\left\langle \int_0^\cdot 1_{[0,s_1]}(u)\, du, \int_0^\cdot 1_{[0,s_2]}(u)\, du \right\rangle_{\mathbb{H}} = s_1 \wedge s_2 = \left\langle 1_{[0,s_1]}, 1_{[0,s_2]} \right\rangle_{L^2}.$$

Hence

$$\mathbb{H} = \left\{ \int_0^\cdot f(u)\, du : f \in L^2 \right\},$$

$$\left\langle \int_0^\cdot f(u)\, du, \int_0^\cdot g(u)\, du \right\rangle_{\mathbb{H}} = \langle f, g \rangle_{L^2}.$$

(Cameron-Martin space)

Gaussian processes

$\rightarrow$ small ball probabilities

Important tool: Cameron-Martin formula

Let $U : \mathbb{H} \to \mathcal{L}$ be the isometry defined by $U(t \mapsto \mathbb{E} W_t L) = L$.

Let P^W be the law of W on $\mathbb{B} = C[0,1]$, i.e. $P^W(B) = \mathbb{P}(W \in B)$.

Theorem. [Cameron-Martin (1944)]
If $h \in \mathbb{H}$, then P^W and P^{W+h} are equivalent and

$$\frac{dP^{W+h}}{dP^W}(W) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}.$$

(If $h \notin \mathbb{H}$, then P^W and P^{W+h} are orthogonal.)

(For BM, compare with Girsanov)

Support of a GP

Support of W:

- smallest closed subset $\mathbb{B}_0 \subseteq \mathbb{B}$ such that $\mathbb{P}(W \in \mathbb{B}_0) = 1$.
- f_0 in support iff $\forall \varepsilon > 0$, $\mathbb{P}(\|W - f_0\|_\infty < \varepsilon) > 0$.

Theorem. [Kallianpur (1971)]

The support is the closure of $\mathbb{H}$ in $\mathbb{B}$.

Proof.

- Elementary arguments: $\mathbb{P}(\|W\|_\infty < \varepsilon) > 0$ for all $\varepsilon > 0$.
- Cameron-Martin: $\mathbb{H} \subseteq \mathbb{B}_0$.
- Closing off: $\overline{\mathbb{H}} \subseteq \mathbb{B}_0$.
- Hahn-Banach: $\overline{\mathbb{H}} = \mathbb{B}_0$.

□

Example: support of BM is $\{f \in C[0,1] : f(0) = 0\}$.

Non-centered small ball probabilities - 1

For $h \in \mathbb{H}$, by Cameron-Martin,

$$\mathbb{P}(\|W - h\|_\infty < \varepsilon) = \mathbb{E}\frac{dP^{W-h}}{dP^W}(W)1_{\|W\|_\infty < \varepsilon}$$
$$= \mathbb{E}e^{-Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1_{\|W\|_\infty < \varepsilon}.$$

Since $W \stackrel{\mathrm{d}}{=} -W$, also

$$\mathbb{P}(\|W - h\|_\infty < \varepsilon) = \mathbb{E}\frac{dP^{W+h}}{dP^W}(W)1_{\|W\|_\infty < \varepsilon}$$
$$= \mathbb{E}e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1_{\|W\|_\infty < \varepsilon}.$$

Together,

$$\mathbb{P}(\|W - h\|_\infty < \varepsilon) = e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \mathbb{E}1_{\|W\|_\infty < \varepsilon} \frac{1}{2}(e^{Uh} + e^{-Uh})$$
$$\geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Non-centered small ball probabilities - 2

Concentration function: for $f_0 \in \mathbb{B}$ and $\varepsilon > 0$:

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\| < \varepsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\| < \varepsilon).$$

Theorem.
for $f_0 \in \mathbb{B}$ and $\varepsilon > 0$

$$\varphi_{f_0}(\varepsilon) \leq -\log \mathbb{P}(\|W - f_0\| < \varepsilon) \leq \varphi_{f_0}(\varepsilon/2).$$

Centered small ball probabilities - 1

Let $h_1, \ldots, h_N$ be elements of $\mathbb{H}_1$ that are 2ε-separated in $\mathbb{B}$. Then

$$\begin{aligned}\sqrt{e} &\geq \sqrt{e}\sum_{j=1}^{N}\mathbb{P}(\|W - h_j\|_\infty < \varepsilon)\\ &\geq \sqrt{e}\sum_{j=1}^{N} e^{-\frac{1}{2}\|h_j\|_{\mathbb{H}}^2}\mathbb{P}(\|W\|_\infty < \varepsilon)\\ &\geq N\mathbb{P}(\|W\|_\infty < \varepsilon).\end{aligned}$$

Hence:

large metric entropy of $\mathbb{H}_1 \sim$ small probability $\mathbb{P}(\|W\|_\infty < \varepsilon)$.

Centered small ball probabilities - 2

More careful (much more) analysis gives (Kuelbs and Li (1993), Li and Linde (1999)):

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \varepsilon^{-\frac{2\alpha}{2+\alpha}} \iff -\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \varepsilon^{-\alpha},$$

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|_\infty) \asymp \log^\gamma \frac{1}{\varepsilon} \iff -\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \log^\gamma \frac{1}{\varepsilon}.$$

Gaussian processes

$\rightarrow$ concentration of measure

Concentration of measure - 1

Finite-dimensional situation:

Let $X \sim N_d(0, \Sigma)$, Σ invertible. Then $\mathbb{H} = \mathbb{R}^d$ and

$$\langle x, y \rangle_{\mathbb{H}} = x^T \Sigma^{-1} y.$$

Hence, the balls in $\mathbb{H}$ (ellipsoids in the ordinary metric) are precisely the level sets of the density of X.

In other words: the RKHS unit ball describes the "geometry" of the support of the distribution of X.

Concentration of measure - 2

$\mathbb{B}_1$: unit ball of $\mathbb{B}$, $\mathbb{H}_1$: unit ball of $\mathbb{H}$.

Theorem. [Borell (1975), Sudakov-Tsirelson (1974)]
For $\varepsilon > 0$ and $M \geq 0$,

$$\mathbb{P}(W \notin \varepsilon\mathbb{B}_1 + M\mathbb{H}_1) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(W \in \varepsilon\mathbb{B}_1)) + M).$$

Let $M(W)$ be the median of $\|W\|_\infty$. Take $\varepsilon = M(W)$ and $M = x/\sigma(W)$, use $\mathbb{H}_1 \subset \sigma(W)\mathbb{B}_1$.

Corollary.
For $x > 0$,

$$\mathbb{P}(\|W\|_\infty - M(W) > x) \leq 1 - \Phi(x/\sigma(W)).$$

Concentration of measure - 2

$\mathbb{B}_1$: unit ball of $\mathbb{B}$, $\mathbb{H}_1$: unit ball of $\mathbb{H}$.

Theorem. [Borell (1975), Sudakov-Tsirelson (1974)]
For $\varepsilon > 0$ and $M \geq 0$,

$$\mathbb{P}(W \notin \varepsilon\mathbb{B}_1 + M\mathbb{H}_1) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(W \in \varepsilon\mathbb{B}_1)) + M).$$

Let $M(W)$ be the median of $\|W\|_\infty$. Take $\varepsilon = M(W)$ and $M = x/\sigma(W)$, use $\mathbb{H}_1 \subset \sigma(W)\mathbb{B}_1$.

Corollary.
For $x > 0$,

$$\mathbb{P}(\|W\|_\infty - M(W) > x) \leq 1 - \Phi(x/\sigma(W)).$$

Gaussian processes

$\rightarrow$ general theorem

Recall what we want

For a continuous GP $W = (W_t)_{t \in [0,1]}$ and function f_0, want to find smallest possible $\varepsilon_n \downarrow 0$ for which there exist $\mathcal{F}_n \subset C[0,1]$ and $c > 0$ such that

1. $$\mathbb{P}(\|W - f_0\|_\infty < \varepsilon_n) \geq e^{-cn\varepsilon_n^2},$$
2. $$\mathbb{P}(W \notin \mathcal{F}_n) \leq e^{-(c+8)n\varepsilon_n^2},$$
3. $$\log N(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \leq n\varepsilon_n^2.$$

Putting things together - 1

Prior mass condition 1. is fulfilled if $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$, where

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Borell-Sudakov suggests to take sieve $\mathcal{F}_n$ of the form $\mathcal{F}_n = M_n \mathbb{H}_1 + \varepsilon_n \mathbb{B}_1$. Have

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W\|_\infty \leq \varepsilon_n)) + M_n).$$

But if $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$, then

$$\mathbb{P}(\|W\|_\infty \leq \varepsilon_n) \geq e^{-\varphi_{w_0}(\varepsilon_n)} \geq e^{-n\varepsilon_n^2},$$

Putting things together - 1

Prior mass condition 1. is fulfilled if $\varphi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$, where

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Borell-Sudakov suggests to take sieve $\mathcal{F}_n$ of the form $\mathcal{F}_n = M_n \mathbb{H}_1 + \varepsilon_n \mathbb{B}_1$. Have

$$\mathbb{P}(W \notin \mathcal{F}_n) \le 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W\|_\infty \le \varepsilon_n)) + M_n).$$

But if $\varphi_{w_0}(\varepsilon_n) \le n\varepsilon_n^2$, then

$$\mathbb{P}(\|W\|_\infty \le \varepsilon_n) \ge e^{-\varphi_{w_0}(\varepsilon_n)} \ge e^{-n\varepsilon_n^2},$$

Putting things together - 2

so

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2}) + M_n).$$

Now take $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$ for some $C > 1$. Then

$$\begin{aligned}\mathbb{P}(W \notin \mathcal{F}_n) &\leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2}) + M_n) \\ &\leq 1 - \Phi(-\Phi^{-1}(e^{-Cn\varepsilon_n^2})) \\ &= e^{-Cn\varepsilon_n^2}.\end{aligned}$$

So for this sieve $\mathcal{F}_n$, the remaining mass condition 2. holds, provided we choose C large enough.

How about the entropy condition?

Putting things together - 2

so

$$\mathbb{P}(W \notin \mathcal{F}_n) \leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2}) + M_n).$$

Now take $M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2})$ for some $C > 1$. Then

$$\begin{aligned}\mathbb{P}(W \notin \mathcal{F}_n) &\leq 1 - \Phi(\Phi^{-1}(e^{-n\varepsilon_n^2}) + M_n) \\ &\leq 1 - \Phi(-\Phi^{-1}(e^{-Cn\varepsilon_n^2})) \\ &= e^{-Cn\varepsilon_n^2}.\end{aligned}$$

So for this sieve $\mathcal{F}_n$, the remaining mass condition 2. holds, provided we choose C large enough.

How about the entropy condition?

Putting things together - 3

The condition $\varphi(\varepsilon_n) \leq n\varepsilon_n^2$ implies a lower bound for the centered small ball probability $\mathbb{P}(\|W\|_\infty < \varepsilon_n)$.

This gives an upper bound on $N(2\varepsilon_n, \mathbb{H}_1, \|\cdot\|_\infty)$, hence also on $N(2\varepsilon_n, M_n\mathbb{H}_1, \|\cdot\|_\infty)$!

Since $N(3\varepsilon_n, M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1, \|\cdot\|_\infty) \leq N(2\varepsilon_n, M_n\mathbb{H}_1, \|\cdot\|_\infty)$, we obtain a bound for $N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty)$.

It turns out that with our choice of M_n, we get

$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \leq 6Cn\varepsilon_n^2.$$

This takes care of the entropy condition 3!

General theorem for Gaussian process priors

Let $W = (W_t)_{t\in[0,1]}$ be a centered, continuous GP, with RKHS $\mathbb{H}$.

Define, for a function f_0,

$$\varphi_{f_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Theorem. [Van der Vaart and vZ. (2008)]
If $\varepsilon_n > 0$ is such that $n\varepsilon_n^2 \to \infty$ and $\varphi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$, then $\forall C > 1$, there exist $\mathcal{F}_n \subset C[0,1]$ s.t.

$$\mathbb{P}(\|W - f_0\|_\infty < 2\varepsilon_n) \ge e^{-n\varepsilon_n^2},$$

$$\mathbb{P}(W \notin \mathcal{F}_n) \le e^{-Cn\varepsilon_n^2}$$

$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \le 6Cn\varepsilon_n^2.$$

General theorem for Gaussian process priors

Let $W = (W_t)_{t\in[0,1]}$ be a centered, continuous GP, with RKHS $\mathbb{H}$.

Define, for a function f_0,

$$\varphi_{f_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_\infty < \varepsilon).$$

Theorem. [Van der Vaart and vZ. (2008)]
If $\varepsilon_n > 0$ is such that $n\varepsilon_n^2 \to \infty$ and $\varphi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$, then $\forall C > 1$, there exist $\mathcal{F}_n \subset C[0,1]$ s.t.

$$\mathbb{P}(\|W - f_0\|_\infty < 2\varepsilon_n) \ge e^{-n\varepsilon_n^2},$$

$$\mathbb{P}(W \notin \mathcal{F}_n) \le e^{-Cn\varepsilon_n^2}$$

$$\log N(3\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \le 6Cn\varepsilon_n^2.$$

Rate results for concrete GP priors

Splitting up the problem

To get a rate ε_n solving $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ we need two things:

1. $$-\log \mathbb{P}(\|W\|_\infty < \varepsilon_n) \leq n\varepsilon_n^2,$$

2. $$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2.$$

Problem 1. only depends on the GP. Widely studied in the "small deviations" or "small balls" community (Lifshits (2015): 329 refs). Result only depends on the prior.

For problem 2. we have to study the approximation of the function f_0 by elements of the RKHS of the GP. Result depends on the relation between f_0 and the prior.

Splitting up the problem

To get a rate ε_n solving $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ we need two things:

1. $$-\log \mathbb{P}(\|W\|_\infty < \varepsilon_n) \leq n\varepsilon_n^2,$$

2. $$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2.$$

Problem 1. only depends on the GP. Widely studied in the "small deviations" or "small balls" community (Lifshits (2015): 329 refs). Result only depends on the prior.

For problem 2. we have to study the approximation of the function f_0 by elements of the RKHS of the GP. Result depends on the relation between f_0 and the prior.

Example: Brownian motion

Brownian motion W:

$$\mathbb{P}(\|W - f_0\|_\infty < \varepsilon) \leq \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp e^{-c(1/\varepsilon)^2}.$$

Can only have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ for $\varepsilon_n \geq n^{-1/4}$.

Hence, can not get faster rate than $\varepsilon_n \asymp n^{-1/4}$ with BM prior. (See Castillo (2008)).

Question: under which conditions on f_0 do we achieve the rate $n^{-1/4}$?

Example: Brownian motion

Brownian motion W:

$$\mathbb{P}(\|W - f_0\|_\infty < \varepsilon) \leq \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp e^{-c(1/\varepsilon)^2}.$$

Can only have $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ for $\varepsilon_n \geq n^{-1/4}$.

Hence, can not get faster rate than $\varepsilon_n \asymp n^{-1/4}$ with BM prior. (See Castillo (2008)).

Question: under which conditions on f_0 do we achieve the rate $n^{-1/4}$?

Example: Brownian motion

Lemma.
If $f_0 \in C^\beta[0,1]$, $\beta \in (0,1]$, then

$$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{-(2-2\beta)/\beta}.$$

Proof.
Approximate f_0 by convolutions. □

Hence for $f_0 \in C^\beta[0,1]$ the prior mass condition $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ holds for

$$\varepsilon_n \asymp \begin{cases} n^{-1/4} & \text{if } \beta \geq 1/2 \\ n^{-\beta/2} & \text{if } \beta \leq 1/2. \end{cases}$$

Example: Brownian motion

Lemma.
If $f_0 \in C^\beta[0,1]$, $\beta \in (0,1]$, then

$$\inf_{h \in \mathbb{H}: \|h - f_0\|_\infty < \varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{-(2-2\beta)/\beta}.$$

Proof.
Approximate f_0 by convolutions. □

Hence for $f_0 \in C^\beta[0,1]$ the prior mass condition $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ holds for

$$\varepsilon_n \asymp \begin{cases} n^{-1/4} & \text{if } \beta \geq 1/2 \\ n^{-\beta/2} & \text{if } \beta \leq 1/2. \end{cases}$$

Univariate nonparametric regression with a BM prior - 1

Observations: $Y_1, \ldots, Y_n$ satisfying

$$Y_i = f_0(i/n) + e_i,$$

with e_i i.i.d. $N(0, \sigma^2)$.

Prior on f: law Π of Brownian motion (with standard normal initial distribution).

Theorem.
Suppose $f_0 \in C^\beta[0,1]$ for $\beta > 0$. Then the posterior contracts around f_0 at the rate

$$\varepsilon_n \asymp \begin{cases} n^{-1/4} & \text{if } \beta \geq 1/2 \\ n^{-\beta/2} & \text{if } \beta \leq 1/2. \end{cases}$$

Univariate nonparametric regression with a BM prior - 2

Note: rate equals $n^{-\beta/(1+2\beta)}$ only for $\beta = 1/2$.

In other words:

The BM prior is rate-optimal if the smoothness of the true regression function equals the smoothness of the BM paths.

Suppose that $f_0 \in C^\beta[0,1]$ for $\beta > 0$.

Which GP prior leads to the optimal rate $n^{-\beta/(1+2\beta)}$?

Univariate nonparametric regression with a BM prior - 2

Note: rate equals $n^{-\beta/(1+2\beta)}$ only for $\beta = 1/2$.

In other words:

The BM prior is rate-optimal if the smoothness of the true regression function equals the smoothness of the BM paths.

Suppose that $f_0 \in C^\beta[0,1]$ for $\beta > 0$.

Which GP prior leads to the optimal rate $n^{-\beta/(1+2\beta)}$?

Priors for other smoothness levels - 1

Candidate: Riemann-Liouville process

$$W_t = \int_0^t (t-s)^{\beta-1/2}\, dB_s.$$

For $\beta - 1/2$ integer: W is $(\beta - 1/2)$-fold repeated integral of B.
(For other β: use fractional calculus.)

Idea: should be good prior model for β-smooth functions.

Priors for other smoothness levels - 2

Known results for the RL-process:

Li and Linde (1998):

$$-\log \mathbb{P}(\|W\|_\infty < \varepsilon) \asymp \varepsilon^{-1/\beta}$$

RKHS is $I_{0+}^{\beta+1/2}(L^2)$, with norm

$$\|I_{0+}^{\beta+1/2} f\|_{\mathbb{H}} = \frac{\|f\|_{L^2}}{\Gamma(\beta + 1/2)}.$$

Priors for other smoothness levels - 3

Modified RL-process with parameter $\beta > 0$:

$$W_t = \sum_{k=0}^{\underline{\beta}+1} Z_k t^k + \int_0^t (t-s)^{\beta-1/2}\, dB_s.$$

Theorem.
The support of the process W is $C[0,1]$. For $f_0 \in C^\beta[0,1]$ we have $\varphi_{f_0}(\varepsilon) = O(\varepsilon^{-1/\beta})$ as $\varepsilon \to 0$.

Univariate nonparametric regression with a RL prior

Observations: $Y_1, \dots, Y_n$ satisfying

$$Y_i = f_0(i/n) + e_i,$$

with e_i i.i.d. $N(0, \sigma^2)$.

Prior on f: law Π of a modified RL-process with parameter $\beta > 0$.

Theorem.
Suppose $f_0 \in C^\beta[0,1]$ for $\beta > 0$. Then the posterior contracts around f_0 at the rate $\varepsilon_n \asymp n^{-\beta/(1+2\beta)}$.

Possible extensions

Other Gaussian priors that can be handled:

- (multiply integrated) fractional BM
- series priors
- Matérn process
- squared exponential process
- ...

Can also handle other statistical settings: density estimation, extracting a signal in white noise, nonparametric classification, drift estimation for diffusions, ...

Concluding remarks

Take home from Lecture III

- Gaussian process theory provides powerful tools for studying asymptotic behaviour of procedures using Gaussian priors.
- Have a powerful general theorem for GP's that matches with general contraction rate theorems.
- For many important nonparametric problems, rate-optimal GP priors can be exhibited.
- Performance depends very much on the fine properties of the GP that is used.
- To get rate-optimal performance, a Gaussian prior has to be carefully tuned to the true parameter to avoid over- or undersmoothing.

Q: How can we get optimal rates without knowledge of the true regularity?

Take home from Lecture III

- Gaussian process theory provides powerful tools for studying asymptotic behaviour of procedures using Gaussian priors.
- Have a powerful general theorem for GP's that matches with general contraction rate theorems.
- For many important nonparametric problems, rate-optimal GP priors can be exhibited.
- Performance depends very much on the fine properties of the GP that is used.
- To get rate-optimal performance, a Gaussian prior has to be carefully tuned to the true parameter to avoid over- or undersmoothing.

Q: How can we get optimal rates without knowledge of the true regularity?

Main references for Lecture III

Review of Gaussian process theory useful for Bayesian nonparametrics:

- van der Vaart, A.W., van Zanten, J.H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In: Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh, Inst. Math. Stat. Collect., 3, pp. 200–222, Beachwood, OH.

General theorem for GP priors, and RL and other examples:

- van der Vaart, A.W., van Zanten, J.H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Statist. 36, no. 3, 1435–1463.