Frequentist properties of Bayesian procedures for infinite-dimensional parameters

Aad van der Vaart
Vrije Universiteit Amsterdam

Forum Lectures
European Meeting of Statisticians
Toulouse, 2009
LECTURE II: GAUSSIAN PROCESS PRIORS

Recap: frequentist Bayesian theory
Examples
Rescaling
Adaptation
General formulation of rates
Examples of settings
Reproducing kernel Hilbert space
Proof ingredients
Harry van Zanten
Recap: frequentist Bayesian theory
Frequentist Bayesian

Given a collection of densities \( \{p_w : w \in \mathcal{W}\} \) indexed by a parameter \( w \), and a prior \( \Pi \) on \( \mathcal{W} \), the posterior is defined by

\[
d\Pi(w|X) \propto p_w(X) \, d\Pi(w).
\]
Frequentist Bayesian

Given a collection of densities \( \{p_w : w \in \mathcal{W}\} \) indexed by a parameter \( w \), and a prior \( \Pi \) on \( \mathcal{W} \), the posterior is defined by

\[
d\Pi(w|X) \propto p_w(X) d\Pi(w).
\]

Assume that the data \( X \) is generated according to a given parameter \( w_0 \) and consider the posterior \( \Pi(w \in \cdot | X) \) as a random measure on the parameter set \( \mathcal{W} \).

We like the posterior to put “most” of its mass near \( w_0 \) for “most” \( X \).
Frequentist Bayesian

Given a collection of densities \( \{p_w : w \in \mathcal{W} \} \) indexed by a parameter \( w \), and a prior \( \Pi \) on \( \mathcal{W} \), the posterior is defined by

\[
d\Pi(w | X) \propto p_w(X) \, d\Pi(w).
\]

Assume that the data \( X \) is generated according to a given parameter \( w_0 \) and consider the posterior \( \Pi(w \in \cdot | X) \) as a random measure on the parameter set \( \mathcal{W} \).

We like the posterior to put “most” of its mass near \( w_0 \) for “most” \( X \).

Asymptotic setting: data \( X^n \) where the information increases as \( n \to \infty \). Three desirable properties:

- Contraction to \( \{w_0\} \) at a fast rate
- Adaptation
- (Distributional convergence)
Rate of contraction

Assume $X^n$ is generated according to a given parameter $w_0$ where the information increases as $n \to \infty$.

- Posterior is **consistent** if $\mathbb{E}_{w_0} \Pi_n (w: d(w, w_0) < \varepsilon | X^n) \to 1$ for every $\varepsilon > 0$.

- Posterior **contracts at rate at least** $\varepsilon_n$ if $\mathbb{E}_{w_0} \Pi_n (w: d(w, w_0) < \varepsilon_n | X^n) \to 1$. 
To a given class of parameters is attached an optimal rate of convergence defined by the minimax criterion.

We like the posterior to contract at this rate.

Given a scale of regularity classes, indexed by a parameter $\alpha$, we like the posterior to adapt: if the true parameter has regularity $\alpha$, then we like the contraction rate to be the minimax rate for the $\alpha$-class.
Adaptation

To a given class of parameters is attached an optimal rate of convergence defined by the minimax criterion.

We like the posterior to contract at this rate.

Given a scale of regularity classes, indexed by a parameter $\alpha$, we like the posterior to adapt: if the true parameter has regularity $\alpha$, then we like the contraction rate to be the minimax rate for the $\alpha$-class.

For instance, in typical examples $n^{-\alpha/(2\alpha+d)}$ if $w_0$ is a function of $d$ arguments with partial derivatives of order $\alpha$ bounded by a constant.
General findings

If $w$ is infinite-dimensional the prior is important.

- The posterior may be inconsistent.
- The rate of contraction often depends on the prior.
- For estimating a functional the prior is less critical, but still plays a role.

The prior does not (completely) wash out as $n \to \infty$. 
Examples
Gaussian process

The law of a stochastic process \((W_t: t \in T)\) is a prior distribution on the space of functions \(w: T \rightarrow \mathbb{R}\).

**Gaussian processes** have been found useful, because

- they offer great variety;
- they have a general index set \(T\);
- they are easy (?) to understand through their covariance function
  \[
  (s, t) \mapsto \mathbb{E}W_s W_t;
  \]
- they can be computationally attractive.
Brownian density estimation

For $W$ Brownian motion use as **prior on a density** $p$ on $[0, 1]$:

$$x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y} \, dy}.$$  

[Leonard, Lenk, Tokdar & Ghosh]
Brownian density estimation

For $W$ Brownian motion use as prior on a density $p$ on $[0, 1]$:

$$x \mapsto \frac{e^{Wx}}{\int_0^1 e^{Wy} \, dy}.$$
Brownian density estimation

Let $X_1, \ldots, X_n$ be iid $p_0$ on $[0, 1]$ and let $W$ Brownian motion. Let the prior be

$$x \mapsto \frac{e^{Wx}}{\int_0^1 e^{Wy} \, dy}$$

**THEOREM**

If $w_0 := \log p_0 \in C^\alpha[0, 1]$, then $L_2$-rate is: $n^{-1/4}$ if $\alpha \geq 1/2$; $n^{-\alpha/2}$ if $\alpha \leq 1/2$. 
Let $X_1, \ldots, X_n$ be iid $p_0$ on $[0, 1]$ and let $W$ Brownian motion. Let the prior be

$$x \mapsto \frac{e^{Wx}}{\int_0^1 e^{Wy} \, dy}$$

**THEOREM**

If $w_0 := \log p_0 \in C^\alpha[0, 1]$, then $L_2$-rate is:

- $n^{-1/4}$ if $\alpha \geq 1/2$;
- $n^{-\alpha/2}$ if $\alpha \leq 1/2$.

- This is optimal if and only if $\alpha = 1/2$.
- Rate does not improve if $\alpha$ increases from $1/2$.
- Consistency for any $\alpha > 0$.

(The same result is true for $w_0$ a regression or classification function.) [vZanten, Castillo (2008)].
Brownian motion $t \mapsto W_t$ — Prior density $t \mapsto c \exp(W_t)$
Integrated Brownian motion

0, 1, 2, 3 and 4 times integrated Brownian motion
Integrated Brownian motion: Riemann-Liouville process

$(\alpha - 1/2)$-times integrated Brownian motion, released at 0

$$W_t = \int_0^t (t - s)^{\alpha - 1/2} dB_s + \sum_{k=0}^{[\alpha]+1} Z_k t^k.$$ 

[Brownian motion, $\alpha > 0$, $(Z_k)$ iid $\mathcal{N}(0, 1)$, “fractional integral”]

**THEOREM**

IBM gives appropriate model for $\alpha$-smooth functions: consistency for any true smoothness $\beta > 0$, but the optimal $n^{-\beta/(2\beta+1)}$ if and only if $\alpha = \beta$. 
Consider nonparametric regression $Y_i = w(x_i) + e_i$ with Gaussian errors, and prior

$$W_t = \sqrt{b} \int_0^t (t - s)^k dB_s + \sqrt{a} \sum_{j=0}^k Z_j t^j.$$

**THEOREM** [Kimeldorf & Wahba (1970s)]
If $a \to \infty$ and $b, n$ are fixed, then the posterior mean tends to the minimizer of

$$w \mapsto \frac{1}{n} \sum_{i=1}^n (Y_i - w(x_i))^2 + \frac{1}{nb} \int_0^1 w^{(k)}(t)^2 \, dt.$$

If $w_0 \in C^k[0, 1]$ and $b \sim n^{-1/(2k+1)}$, then the penalized least squares estimator is rate optimal.
Brownian sheet

Brownian sheet \((W_t: t \in [0, 1]^d)\) has covariance function

\[
\text{cov}(W_s, W_t) = (s_1 \land t_1) \cdots (s_d \land t_d).
\]

BS gives rates of the order

\[
n^{-1/4} (\log n)^{(2d-1)/4}
\]

for sufficiently smooth \(w_0\) \((\alpha \geq d/2)\).
Fractional Brownian motion

$W$ zero-mean Gaussian with (Hurst index $0 < \alpha < 1$)

$$\text{cov}(W_s, W_t) = s^{2\alpha} + t^{2\alpha} - |t - s|^{2\alpha}.$$  

fBM is appropriate model for $\alpha$-smooth functions. Integrate to cover $\alpha > 1$. 

![Graphs showing time series for alpha=0.8 and alpha=0.2]
Given a basis $e_1, e_2, \ldots$ put a Gaussian prior on the coefficients $(\theta_1, \theta_2, \ldots)$ in an expansion

$$\theta = \sum_i \theta_i e_i.$$ 

For instance: $\theta_1, \theta_2, \ldots$ independent with $\theta_i \sim N(0, \sigma_i^2)$.

Appropriate decay of $\sigma_i$ gives proper model for $\alpha$-smooth functions.
For a wavelet basis \((\psi_{j,k})\) with good approximation properties for \(B^\beta_{\infty,\infty}[0, 1]^d\), and \(Z_{j,k}\) iid standard normal variables, \(W = \sum_{j=1}^{J\alpha} \sum_{k=1}^{2^j} 2^{-jc} 2^{j/2} Z_{j,k} \psi_{j,k}\), \(2^{J\alpha} = n^{d/(2\alpha+d)}\).

**THEOREM**
If \(w_0 \in B^\beta_{\infty,\infty}[0, 1]^d\), the rate is
\[
\varepsilon_n = \begin{cases} 
  n^{-\beta/(2\alpha+d)} \log n & \text{if } c \leq \beta \leq \alpha, \\
  n^{-\alpha/(2\alpha+d)} \log n & \text{if } c \leq \alpha \leq \beta, \\
  n^{-c/(2c+d)} (\log n)^{d/(2c+d)} & \text{if } \alpha \leq c \leq \beta, \\
  n^{-\beta/(2c+d)} (\log n)^{d/(2c+d)} & \text{if } \alpha \leq \beta \leq c.
\end{cases}
\]

In particular, equal prior weight to all levels \((c = 0)\) gives the optimal weight if \(\beta = \alpha\) (\(c = \beta\) is better).
A stationary Gaussian field \((W_t: t \in \mathbb{R}^d)\) is characterized through a spectral measure \(\mu\), by

\[
\text{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)} \, d\mu(\lambda).
\]

Smoothness of \(t \mapsto W_t\) is controlled by the tails of \(\mu\). For instance, exponentially small tails give infinitely smooth sample paths; Matérn gives \(\alpha\)-regular functions.
A stationary Gaussian field \((W_t: t \in \mathbb{R}^d)\) is characterized through a spectral measure \(\mu\), by

\[
\text{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)} \, d\mu(\lambda).
\]

Smoothness of \(t \mapsto W_t\) is controlled by the tails of \(\mu\). For instance, exponentially small tails give infinitely smooth sample paths; Matérn gives \(\alpha\)-regular functions.

**THEOREM**  If \(\int e^{\|\lambda\|} \left| \hat{w}_0(\lambda) \right|^2 \, d\lambda < \infty\), then the Gaussian spectral measure gives a near \(1/\sqrt{n}\)-rate of contraction; it gives consistency but suboptimal rates for Hölder smooth functions.

**Conjecture:** Matérn gives good results for Sobolev spaces.
Rescaling
Stretching or shrinking

Sample paths can be smoothed by stretching.
Stretching or shrinking

Sample paths can be *smoothed* by stretching

and *roughened* by shrinking
Rescaled Brownian motion

\[ W_t = \frac{B_t}{c_n} \] for \( B \) Brownian motion, and \( c_n \sim n^{(2\alpha-1)/(2\alpha+1)} \)

- \( \alpha < 1/2 \): \( c_n \to 0 \) (shrink).
- \( \alpha \in (1/2, 1] \): \( c_n \to \infty \) (stretch).

**THEOREM**
The prior \( W_t = \frac{B_t}{c_n} \) gives optimal rate for \( w_0 \in C^\alpha[0, 1], \alpha \in (0, 1] \).

Surprising? (Brownian motion is self-similar!.)
Rescaled Brownian motion

\[ W_t = B_{t/c_n} \text{ for } B \text{ Brownian motion, and } c_n \sim n^{(2\alpha-1)/(2\alpha+1)} \]

- \( \alpha < 1/2: \) \( c_n \to 0 \) (shrink).
- \( \alpha \in (1/2, 1]: \) \( c_n \to \infty \) (stretch).

**THEOREM**

The prior \( W_t = B_{t/c_n} \) gives optimal rate for \( w_0 \in C^\alpha[0, 1], \alpha \in (0, 1] \).

Surprising? (Brownian motion is self-similar!)

Appropriate rescaling of \( k \) times integrated Brownian motion gives optimal prior for every \( \alpha \in (0, k + 1] \).
Rescaled Brownian motion

\[ W_t = B_t/c_n \] for \( B \) Brownian motion, and \( c_n \sim n^{(2\alpha-1)/(2\alpha+1)} \)

- \( \alpha < 1/2 \): \( c_n \to 0 \) (shrink).
- \( \alpha \in (1/2, 1] \): \( c_n \to \infty \) (stretch).

**THEOREM**
The prior \( W_t = B_t/c_n \) gives optimal rate for \( w_0 \in C^\alpha[0, 1], \alpha \in (0, 1] \).

Surprising? (Brownian motion is self-similar!)

Appropriate rescaling of \( k \) times integrated Brownian motion gives optimal prior for every \( \alpha \in (0, k + 1] \).

For \( \alpha = k \) we find the optimal bandwidth for penalized regression as in Kimeldorf and Wahba.
A Gaussian field with infinitely-smooth sample paths is obtained with

\[ \int e^{\|\lambda\|} \hat{\psi}(\lambda) \, d\lambda < \infty. \]

**THEOREM**

The prior \( W_t = G_{t/c_n} \) for \( c_n \sim n^{-1/(2\alpha+d)} \) gives nearly optimal rate for \( w_0 \in C^\alpha[0, 1] \), any \( \alpha > 0 \).
Scaling changes the properties of the prior and hence hyperparameters are important.

A smooth prior process can be scaled to achieve any desired level of “prior roughness”, but a rough process cannot be smoothed much and will necessarily impose its roughness on the data.
Adaptation
Hierarchical priors

For each $\alpha > 0$ there are several priors $\Pi_{\alpha}$ (Riemann-Liouville, Fractional, Series, Matern, rescaled processes,...) that are appropriate for estimating $\alpha$-smooth functions.

We can combine them into a mixture prior:

- Put a prior weight $d\rho(\alpha)$ on $\alpha$.
- Given $\alpha$ use an optimal prior $\Pi_{\alpha}$ for that $\alpha$.

This works (nearly), provided $\rho$ is chosen with some (but not much) care.

The weights $d\rho(\alpha) \propto e^{-n\varepsilon_{n,\alpha}^2} d\alpha$ always work.

[Lember, Szabo]
Adaptation by rescaling

- Choose $A^d$ from a Gamma distribution.
- Choose $(G_t: t > 0)$ centered Gaussian with $EG_s G_t = \exp(-\|s - t\|^2)$.
- Set $W_t \sim G_{At}$.

**THEOREM**

- if $w_0 \in C^\alpha[0, 1]^d$, then the rate of contraction is nearly $n^{-\alpha/(2\alpha+d)}$.
- if $w_0$ is supersmooth, then the rate is nearly $n^{-1/2}$.

Reverend Thomas solved the bandwidth problem!?
Gaussian regression with Brownian motion rescaled by a Gamma variable.

Conjecture: this (nearly) gives the optimal rate $n^{-\alpha/(2\alpha+1)}$ if true regression function is in $C^\alpha[0, 1]$ for $\alpha \in (0, 1]$. Integrating BM extends this to higher $\alpha$. 
General formulation of rates
Two ingredients:

- RKHS
- Small ball exponent
Think of the Gaussian process as a random element in a Banach space \((\mathcal{B}, \| \cdot \|)\).

To every such Gaussian random element is attached a certain Hilbert space \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), called the RKHS.

\(\| \cdot \|_\mathcal{H}\) is stronger than \(\| \cdot \|\) and hence can consider \(\mathcal{H} \subset \mathcal{B}\).
Think of the Gaussian process as a random element in a Banach space \((\mathcal{B}, \| \cdot \|)\).

To every such Gaussian random element is attached a certain Hilbert space \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), called the RKHS.

\(\| \cdot \|_\mathcal{H}\) is stronger than \(\| \cdot \|\) and hence can consider \(\mathcal{H} \subset \mathcal{B}\).

**EXAMPLE**

If \(W\) is multivariate normal \(N_d(0, \Sigma)\), then the RKHS is \(\mathbb{R}^d\) with norm

\[
\| h \|_\mathcal{H} = \sqrt{h^t \Sigma^{-1} h}
\]
Think of the Gaussian process as a random element in a Banach space \((\mathcal{B}, \| \cdot \|)\).

To every such Gaussian random element is attached a certain Hilbert space \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), called the RKHS.

\(\| \cdot \|_\mathcal{H}\) is stronger than \(\| \cdot \|\) and hence can consider \(\mathcal{H} \subset \mathcal{B}\).

EXAMPLE
Brownian motion is a random element in \(C[0, 1]\).
Its RKHS is \(\mathcal{H} = \{ h : \int h'(t)^2 \, dt < \infty \}\) with norm \(\| h \|_\mathcal{H} = \| h' \|_2\).
The small ball probability of a Gaussian random element $W$ in $(\mathbb{B}, \| \cdot \|)$ is

$$P(\|W\| < \varepsilon),$$

and the small ball exponent is

$$\phi_0(\varepsilon) = - \log P(\|W\| < \varepsilon).$$
The **small ball probability** of a Gaussian random element $W$ in $(\mathbb{B}, \| \cdot \|)$ is

$$P(\|W\| < \varepsilon),$$

and the **small ball exponent** is

$$\phi_0(\varepsilon) = -\log P(\|W\| < \varepsilon).$$

**EXAMPLE**

For Brownian motion $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$ as $\varepsilon \downarrow 0$. 
Small ball probabilities can be computed either by probabilistic arguments, or analytically from the RKHS.
Small ball probabilities can be computed either by probabilistic arguments, or analytically from the RKHS.

$$N(\varepsilon, B, d) = \# \varepsilon\text{-balls}$$

**THEOREM**  [Kuelbs & Li 93]

For $\mathbb{H}_1$ the unit ball of the RKHS (up to constants),

$$\phi_0(\varepsilon) \asymp \log N\left(\frac{\varepsilon}{\sqrt{\phi_0(\varepsilon)}}, \mathbb{H}_1, \| \cdot \| \right).$$

There is a big literature on small ball probabilities. (In July 2009 243 entries in database maintained by Michael Lifshits.)
Prior $W$ is Gaussian map in $(\mathcal{B}, \| \cdot \|)$ with RKHS $(\mathcal{H}, \| \cdot \|_\mathcal{H})$ and small ball exponent $\phi_0(\varepsilon) = -\log P(\|W\| < \varepsilon)$.

**THEOREM**
If statistical distances on the model combine appropriately with the norm $\| \cdot \|$ of $\mathcal{B}$, then the posterior rate is $\varepsilon_n$ if

$$\phi_0(\varepsilon_n) \leq n\varepsilon_n^2 \quad \text{AND} \quad \inf_{h \in \mathcal{H} : \|h - w_0\| < \varepsilon_n} \|h\|_\mathcal{H}^2 \leq n\varepsilon_n^2.$$

- Both inequalities give lower bound on $\varepsilon_n$.
- The first depends on $W$ and not on $w_0$.
- If $w_0 \in \mathcal{H}$, then second inequality is satisfied.
Example — Brownian motion

\(W\) one-dimensional Brownian motion on \([0, 1]\).

- RKHS \(\mathbb{H} = \{h: \int h'(t)^2 \, dt < \infty\}\), \(|h|_\mathbb{H} = |h'|_2\).

- Small ball exponent \(\phi_0(\varepsilon) \lesssim (1/\varepsilon)^2\).

**LEMMA**

If \(w_0 \in C^\alpha[0, 1]\) for \(0 < \alpha < 1\), then

\[
\inf_{h \in \mathbb{H}: \|h - w_0\|_\infty < \varepsilon} \|h'|_2^2 \lesssim \left(\frac{1}{\varepsilon}\right)^{(2-2\alpha)/\alpha}.
\]
Example — Brownian motion

$W$ one-dimensional Brownian motion on $[0, 1]$.

- RKHS $\mathbb{H} = \{ h: \int h'(t)^2 \, dt < \infty \}, \| h \|_\mathbb{H} = \| h' \|_2$.
- Small ball exponent $\phi_0(\varepsilon) \lesssim (1/\varepsilon)^2$.

**LEMMA**

If $w_0 \in C^\alpha[0, 1]$ for $0 < \alpha < 1$, then

$$\inf_{h \in \mathbb{H}: \| h-w_0 \|_\infty < \varepsilon} \| h' \|_2^2 \lesssim \left( \frac{1}{\varepsilon} \right)^{(2-2\alpha)/\alpha}.$$  

**CONSEQUENCE:**

Rate is $\varepsilon_n$ if $(1/\varepsilon_n)^2 \leq n\varepsilon_n^2$ AND $(1/\varepsilon_n)^{(2-2\alpha)/\alpha} \leq n\varepsilon_n^2$.

- First implies $\varepsilon_n \geq n^{-1/4}$ for any $w_0$.
- Second implies $\varepsilon_n \geq n^{-\alpha/2}$ for $w_0 \in C^\alpha[0, 1]$. 
Examples of settings
Prior $W$ is Gaussian map in $(B, \| \cdot \|)$ with RKHS $(H, \| \cdot \|_H)$ and small ball exponent $\phi_0(\varepsilon)$.

**THEOREM**

If statistical distances on the model combine appropriately with the norm $\| \cdot \|$ of $B$, then the posterior rate is $\varepsilon_n$ if

$$\phi_0(\varepsilon_n) \leq n\varepsilon_n^2 \quad \text{AND} \quad \inf_{h \in H : \| h - w_0 \| < \varepsilon_n} \| h \|_H^2 \leq n\varepsilon_n^2.$$
Density estimation

Data $X_1, \ldots, X_n$ iid from density on $[0, 1]$,

$$p_w(x) = \frac{e^{wx}}{\int_0^1 e^{wt} dt}.$$ 

- Distance on parameter: Hellinger on $p_w$.
- Norm on $W$: uniform.
Density estimation

Data \( X_1, \ldots, X_n \) iid from density on \([0, 1]\),

\[
p_w(x) = \frac{e^{w_x}}{\int_0^1 e^{w_t} \, dt}.
\]

- Distance on parameter: Hellinger on \( p_w \).
- Norm on \( W \): uniform.

**LEMMA** \( \forall v, w \)

- \( h(p_v, p_w) \leq \|v - w\|_\infty e^{\|v-w\|_\infty/2} \).
- \( K(p_v, p_w) \lesssim \|v - w\|_\infty^2 e^{\|v-w\|_\infty} (1 + \|v - w\|_\infty) \).
- \( V(p_v, p_w) \lesssim \|v - w\|_\infty^2 e^{\|v-w\|_\infty} (1 + \|v - w\|_\infty)^2 \).
Data \((X_1, Y_1), \ldots, (X_n, Y_n)\) iid in \([0, 1] \times \{0, 1\}\)

\[
P_w(Y = 1|X = x) = \Psi(w_x),
\]

for \(\Psi\) the logistic or probit link function.

- **Distance on parameter**: \(L_2\)-norm on \(\Psi(w)\).
- **Norm on \(W\) for logistic**: \(L_2(G)\), \(G\) marginal of \(X_i\).
  
  **Norm on \(W\) for probit**: combination of \(L_2(G)\) and \(L_4(G)\).
Regression

Data $Y_1, \ldots, Y_n$, fixed design points $x_1, \ldots, x_n$

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

for $e_1, \ldots, e_n$ iid Gaussian mean-zero errors.

- Distance on parameter: empirical $L_2$-distance on $w$.
- Norm on $W$: uniform.
Data \((X_t: t \in [0, n])\)

\[ dX_t = w(X_t) \, dt + \sigma(X_t) \, dB_t. \]

Ergodic, recurrent on \(\mathbb{R}\), stationary measure \(\mu_0\), “usual” conditions.

- **Distance on parameter:** random Hellinger \(h_n\).
- **Norm on \(W\):** \(L_2(\mu_0)\).

\[
h^2_n(w_1, w_2) = \int_0^n \left( \frac{w_1(X_t) - w_2(X_t)}{\sigma(X_t)} \right)^2 dt \approx \| (w_1 - w_2)/\sigma \|^2_{\mu_0,2}.
\]

[ van der Meulen & vZ & vdV, Panzar & vZ]
Reproducing kernel Hilbert space
For a zero-mean Gaussian $W$ in Banach space $(\mathcal{B}, \| \cdot \|)$, define $S: \mathcal{B}^* \to \mathcal{B}$ by

$$Sb^* = EWb^*(W).$$

**DEFINITION**

The RKHS $(\mathcal{H}, \| \cdot \|_\mathcal{H})$ is the completion of $S\mathcal{B}^*$ under

$$\langle Sb_1^*, Sb_2^* \rangle_\mathcal{H} = Eb_1^*(W)b_2^*(W).$$
Let \( W = (W_x : x \in \mathcal{X}) \) be a Gaussian process with bounded sample paths and covariance function

\[
K(x, y) = \mathbb{E}W_x W_y.
\]

**DEFINITION**
The RKHS is the completion of the set of functions

\[
x \mapsto \sum_i \alpha_i K(y_i, x),
\]

relative to inner product

\[
\left\langle \sum_i \alpha_i K(y_i, \cdot), \sum_j \beta_j K(z_j, \cdot) \right\rangle_{\mathcal{H}} = \sum_i \sum_j \alpha_i \beta_j K(y_i, z_j).
\]
Definition (3)

Any Gaussian random element in a separable Banach space can be represented as

\[ W = \sum_{i=1}^{\infty} \mu_i Z_i e_i, \]

for

- \( \mu_i \downarrow 0 \)
- \( Z_1, Z_2, \ldots \) iid \( N(0, 1) \)
- \( \|e_1\| = \|e_2\| = \cdots = 1 \)

The RKHS consists of all elements \( h := \sum_i h_i e_i \) with

\[ \|h\|_{H}^2 := \sum_i \frac{h_i^2}{\mu_i} < \infty. \]
THEOREM
The RKHS of $TW$ for a 1-1 operator $T : \mathcal{B} \rightarrow \mathcal{B}'$ between Banach spaces is $T\mathcal{H}$, and $T : \mathcal{H} \rightarrow \mathcal{H}'$ is an isometry.

EXAMPLE
The integration operator

$$T_\alpha w(t) = \int_0^t (t - s)^{\alpha - 1} w(s) \, ds$$

applied to Brownian motion gives the $(\alpha + 1/2)$-Riemann-Liouville process $T_\alpha W$. Its RKHS is $\mathcal{H} = T_{\alpha+1}(L_2[0,1])$ with norm

$$\| T_{\alpha+1} h \|_\mathcal{H} = \| h \|_2.$$
Useful properties

**THEOREM**
The RKHS of $TW$ for a 1-1 operator $T: \mathcal{B} \to \mathcal{B}'$ between Banach spaces is $TH$, and $T: \mathcal{H} \to \mathcal{H}'$ is an isometry.

**THEOREM**
The RKHS of the sum $V + W$ of independent Gaussian variables is $\mathcal{H}^V + \mathcal{H}^W$ with norm

$$\| h^V + h^W \|_{\mathcal{H}^V + \mathcal{H}^W}^2 = \| h^V \|_{\mathcal{H}^V}^2 + \| h^W \|_{\mathcal{H}^W}^2,$$

whenever the supports of $V$ and $W$ have trivial intersection (and are complemented).
Example — stationary processes

A stationary Gaussian process \((W_t: t \in \mathbb{R}^d)\) is characterized through a spectral measure \(\mu\), by

\[
\text{cov}(W_s, W_t) = \int e^{i\lambda^T(s-t)} \, d\mu(\lambda).
\]

**LEMMA**

The RKHS of \((W_t: t \in T)\) is the set of real parts of the functions

\[
t \mapsto \int e^{i\lambda^T t} \psi(\lambda) \, d\mu(\lambda), \quad \psi \in L_2(\mu),
\]

with RKHS-norm equal to the infimum of \(\|\psi\|_2\) over all \(\psi\). If \(T\) has nonempty interior and \(\int e^{\|\lambda\|} \mu(d\lambda) < \infty\), then \(\psi\) is unique.
Example — stationary processes

A stationary Gaussian process \( (W_t: t \in \mathbb{R}^d) \) is characterized through a spectral measure \( \mu \), by

\[
\text{cov}(W_s, W_t) = \int e^{i\lambda T(s-t)} \, d\mu(\lambda).
\]

**LEMMA**
The RKHS of \( (W_t: t \in T) \) is the set of real parts of the functions

\[
t \mapsto \int e^{i\lambda T t} \psi(\lambda) \, d\mu(\lambda), \quad \psi \in L_2(\mu),
\]

with RKHS-norm equal to the infimum of \( \|\psi\|_2 \) over all \( \psi \). If \( T \) has nonempty interior and \( \int e^{\|\lambda\|} \mu(d\lambda) < \infty \), then \( \psi \) is unique.

To compute rate must approximate \( w_0 \) by an element of RKHS. If \( d\mu(\lambda) = m(\lambda)d\lambda \), then

\[
w_0(t) = \int e^{itT \lambda} \hat{w}_0(\lambda) \, d\lambda = \int e^{itT \lambda} \hat{w}_0(\lambda) \frac{1}{m(\lambda)} \, d\mu(\lambda).
\]
Proof ingredients
Proof

Given that the relevant statistical distances translate into the Banach space norm, it follows from general results that the posterior rate is $\varepsilon_n$ if there exist sets $\mathbb{B}_n$ such that

1. $\log N(\varepsilon_n, \mathbb{B}_n, d) \leq n\varepsilon_n^2$ and $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$.  

2. $\Pi_n\left(w: \|w - w_0\| < \varepsilon_n\right) \geq e^{-n\varepsilon_n^2}$. 

The second condition actually implies the first.
$W$ a Gaussian map in $(B, \| \cdot \|)$ with RKHS $(H, \| \cdot \|_H)$ and small ball exponent $\phi_0(\varepsilon)$.

$$\phi_{w_0}(\varepsilon) := \phi_0(\varepsilon) + \inf_{h \in H : \| h - w_0 \| < \varepsilon} \| h \|_H^2.$$
Prior mass

$W$ a Gaussian map in $(\mathbb{B}, \| \cdot \|)$ with RKHS $(\mathbb{H}, \| \cdot \|_{\mathbb{H}})$ and small ball exponent $\phi_0(\varepsilon)$.

$$\phi_{w_0}(\varepsilon): = \phi_0(\varepsilon) + \inf_{h \in \mathbb{H}: \| h - w_0 \| < \varepsilon} \| h \|^2_{\mathbb{H}}.$$

**THEOREM** [Kuelbs & Li 93)]

Concentration function measures concentration around $w_0$:

$$P(\| W - w_0 \| < \varepsilon) \asymp e^{-\phi_{w_0}(\varepsilon)}.$$

up to factors 2
RKHS gives the “geometry of the support of $W$”.

**THEOREM**
The closure of $\mathbb{H}$ in $\mathbb{B}$ is support of the Gaussian measure (and hence posterior inconsistent if $\|w_0 - \mathbb{H}\| > 0$).

**THEOREM** [Borell 75]
For $\mathbb{H}_1$ and $\mathbb{B}_1$ the unit balls of RKHS and $\mathbb{B}$

$$P(W \notin M\mathbb{H}_1 + \varepsilon\mathbb{B}_1) \leq 1 - \Phi\left(\Phi^{-1}(e^{-\phi_0(\varepsilon)}) + M\right).$$
Proof

Given that the relevant statistical distances translate into the Banach space norm, it follows from general results that the posterior rate is $\varepsilon_n$ if there exist sets $\mathcal{B}_n$ such that

1. $\log N(\varepsilon_n, \mathcal{B}_n, d) \leq n\varepsilon_n^2$ and $\Pi_n(\mathcal{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$. 

2. $\Pi_n(w: \|w - w_0\| < \varepsilon_n) \geq e^{-n\varepsilon_n^2}$

Take $\mathcal{B}_n = M_n\mathbb{H}_1 + \varepsilon_n\mathcal{B}_1$ for appropriate $M_n$. 

entropy. 

prior mass
Conclusion
Conclusion

Bayesian inference with Gaussian processes is flexible and elegant. However, priors must be chosen with some care: eye-balling pictures of sample paths does not reveal the fine properties that matter for posterior performance.