The Bayesian paradigm is an elegant and unified approach.
The Bayesian Choice

The Bayesian paradigm is an elegant and unified approach.

According to the complete class theorem (e.g. Le Cam, 1964) the set of Bayes procedures is sufficiently rich to dominate every statistical procedure.
The Bayesian paradigm is an elegant and unified approach.

According to the complete class theorem (e.g. Le Cam, 1964) the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.
The Bayesian paradigm is an elegant and unified approach.

According to the complete class theorem (e.g. Le Cam, 1964) the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.

**Does it work in nonparametrics?**
The Bayesian Choice

The Bayesian paradigm is an elegant and unified approach.

According to the complete class theorem (e.g. Le Cam, 1964) the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.

Does it work in nonparametrics?

Frequentist study:
We use the Bayesian paradigm to define a random measure (the posterior) and see if this contracts to the distribution that generated the data, as the information in the data increases indefinitely, and at what rate.
The Bayesian Choice

The Bayesian paradigm is an elegant and unified approach.

According to the complete class theorem (e.g. Le Cam, 1964) the set of all limits of Bayes procedures is sufficiently rich to dominate every statistical procedure.

Does it work in nonparametrics?

Frequentist study:
We use the Bayesian paradigm to define a random measure (the posterior) and see if this contracts to the distribution that generated the data, as the information in the data increases indefinitely, and at what rate.

In nonparametrics the prior matters.
There are good ones and bad ones.
PART 1: Generalities
PART 2: Gaussian process priors
PART 3: Adaptation
Generalities
Setting

For $n = 1, 2, \ldots$

- $(\mathcal{X}^{(n)}, A^{(n)}, P^{(n)}_\theta : \theta \in \Theta_n)$ experiment
- $(\Theta_n, d_n)$ metric space
- $X^{(n)}$ observation, law $P^{(n)}_{\theta_0}$

Given prior $\Pi_n$ on $\Theta_n$ form posterior

$$
\Pi_n(B|X^{(n)}) = \frac{\int_B p^{(n)}_\theta(X^{(n)}) \, d\Pi_n(\theta)}{\int_{\Theta_n} p^{(n)}_\theta(X^{(n)}) \, d\Pi_n(\theta)}
$$
For $n = 1, 2, \ldots$

- $(X^{(n)}, A^{(n)}, P^{(n)}_{\theta} : \theta \in \Theta_n)$ experiment
- $(\Theta_n, d_n)$ metric space
- $X^{(n)}$ observation, law $P^{(n)}_{\theta_0}$

Given prior $\Pi_n$ on $\Theta_n$ form posterior

$$\Pi_n(B|X^{(n)}) = \frac{\int_B p^{(n)}_{\theta}(X^{(n)}) d\Pi_n(\theta)}{\int_{\Theta_n} p^{(n)}_{\theta}(X^{(n)}) d\Pi_n(\theta)}$$

Rate of contraction is at least $\varepsilon_n$ if $\forall M_n \to \infty$

$$P^{(n)}_{\theta_0} \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \geq M_n \varepsilon_n|X^{(n)}) \to 0$$
Toy problem

- $X_1, \ldots, X_n$ i.i.d. density $p_0$ on $[0, 1]$
- $(W_x : x \in [0, 1])$ Brownian motion

Form prior on $p$:

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

- $X_1, \ldots, X_n$ i.i.d. density $p_0$ on $[0, 1]$
- $(W_x : x \in [0, 1])$ Brownian motion

Form prior on $p$:

\[
x \mapsto \frac{e^{W_x}}{\int_0^1 e^{W_y} \, dy}
\]

Find rate if $\log p_0 \in C^\alpha [0, 1]$
Brownian motion—density

Motivation
Generalities
Setting
Toy problem
Brownian motion—density
Integrated Brownian motion—density
Setting
Setting
Entropy
Rate theorem
Rate theorem-refined
Examples of settings
Examples of priors
Gaussian priors
Gaussian priors-settings
Gaussian priors-proof
Gaussian priors-examples
Adaptation
Integrating Brownian Motion—density

Motivation

Generalities

Setting

Toy problem

Brownian motion—density

Integrated Brownian motion—density

Setting

Setting

Entropy

Rate theorem

Rate theorem-refined

Examples of settings

Examples of priors

Gaussian priors

Gaussian priors-settings

Gaussian priors-proof

Gaussian priors-examples

Adaptation
### Setting

For $n = 1, 2, \ldots$

- $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)}: \theta \in \Theta_n)$ experiment
- $(\Theta_n, d_n)$ metric space
- $X^{(n)}$ observation, law $P_{\theta_0}^{(n)}$

Given prior $\Pi_n$ on $\Theta_n$ form posterior

$$\Pi_n(B|X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)}) d\Pi_n(\theta)}{\int_{\Theta_n} p_\theta^{(n)}(X^{(n)}) d\Pi_n(\theta)}$$

**Rate of contraction is at least** $\varepsilon_n$ **if** $\forall M_n \to \infty$

$$P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n: d_n(\theta, \theta_0) \geq M_n \varepsilon_n|X^{(n)}) \to 0$$
Setting

For \( n = 1, 2, \ldots \)

- \((\mathcal{X}^{(n)}, A^{(n)}, P^{(n)}_\theta : \theta \in \Theta_n)\) experiment
- \((\Theta_n, d_n)\) metric space
- \(X^{(n)}\) observation, law \(P^{(n)}_{\theta_0}\)

Assume \( \exists \xi > 0 \) such that \( \forall n \) \( \exists \) metric \( \bar{d}_n \geq d_n \) such that \( \forall \varepsilon > 0 \):

\( \forall \theta_1 \in \Theta_n \) with \( d_n(\theta_1, \theta_0) > \varepsilon \) \( \exists \) test \( \phi_n \) with

\[
P^{(n)}_{\theta_0} \phi_n \leq e^{-n\varepsilon^2}, \quad \sup_{\theta \in \Theta_n \colon \bar{d}_n(\theta, \theta_1) < \varepsilon \xi} P^{(n)}_\theta (1 - \phi_n) \leq e^{-n\varepsilon^2}
\]
Entropy

\[ N(\varepsilon, \Theta, d) = \text{smallest number of balls of radius } \varepsilon \text{ needed to cover } \Theta \]

Le Cam 73,75,86, Birgé 83, 06:
\[ \exists \text{ estimators } \hat{\theta}_n \text{ with } d_n(\hat{\theta}_n, \theta_0) = O_P(\varepsilon_n) \text{ if } \]
\[ \sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon \xi, \{ \theta \in \Theta_n : d_n(\theta, \theta_0) \leq \varepsilon \}, \bar{d}_n) \leq n\varepsilon_n^2 \]
Entropy

\[ N(\varepsilon, \Theta, d) = \text{smallest number of balls of radius } \varepsilon \text{ needed to cover } \Theta \]

Le Cam 73,75,86, Birgé 83, 06:
\[ \exists \text{ estimators } \hat{\theta}_n \text{ with } d_n(\hat{\theta}_n, \theta_0) = O_P(\varepsilon_n) \text{ if} \]
\[ \log N(\varepsilon_n \xi, \Theta_n, \bar{d}_n) \leq n\varepsilon_n^2 \]

If many balls are needed, then rate \( \varepsilon_n \) is slow
THEOREM (Ghosal & vdV, 2006)
For $\varepsilon_n \to 0$, $\varepsilon_n \gg 1/\sqrt{n}$, assume $\exists \tilde{\Theta}_n \subset \Theta_n$:

1. $\log N(\varepsilon_n, \tilde{\Theta}_n, \bar{d}_n) \leq n\varepsilon_n^2$ \hspace{1cm} \text{entropy}
2. $\Pi_n(\tilde{\Theta}_n - \Theta_n) = o(e^{-3n\varepsilon_n^2})$
3. $\Pi_n(B_n(\theta_0, \varepsilon_n; k)) \geq e^{-n\varepsilon_n^2}$ \hspace{1cm} \text{prior mass}

Then $P_{\theta_0}^{(n)}\Pi_n(\theta \in \Theta_n: d_n(\theta, \theta_0) \geq M_n\varepsilon_n|X^{(n)}) \to 0$

$$B_n(\theta_0, \varepsilon; k) =$$
$$\left\{ \theta \in \Theta_n: K(p_{\theta_0}^{(n)}, p^{(n)}_\theta) \leq n\varepsilon^2, V_k(p_{\theta_0}^{(n)}, p^{(n)}_\theta) \leq n^{k/2}\varepsilon^k \right\}$$
(Kullback-Leibler neighborhood)

$$K(p, q) = P \log(p/q) \hspace{1cm} V_k(p, q) = P \left| \log(p/q) - K(p, q) \right|^k$$
THEOREM (Ghosal & vdV, 2006)

For $\varepsilon_n \to 0$, assume $\exists \tilde{\Theta}_n \subset \Theta_n$:

1. $\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon \xi, \{\theta \in \tilde{\Theta}_n: d_n(\theta, \theta_0) < \varepsilon\}, \bar{d}_n) \leq n\varepsilon_n^2$
2. $\frac{\Pi_n(\tilde{\Theta}_n - \Theta_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n; k))} = o(e^{-2n\varepsilon_n^2})$
3. $\frac{\Pi_n(\theta \in \Theta_n: d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n; k))} \leq e^{Kn\varepsilon_n^2j^2/2} \quad \forall j$

Then $P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n: d_n(\theta, \theta_0) \geq M_n\varepsilon_n|X^{(n)}) \to 0$
Examples of settings

Motivation
Generalities
Examples of settings
I.i.d. observations
Independent observations
Markov chains
Gaussian white noise model
Gaussian time series
Ergodic diffusions
Examples of priors
Gaussian priors
Gaussian priors-settings
Gaussian priors-proof
Gaussian priors-examples
Adaptation
I.i.d. observations

Data $X_1, \ldots, X_n$, i.i.d. with density $p_\theta$

MAIN RESULT HOLDS WITH

- $d_n$ Hellinger distance $h$ (or $L_1$ or $L_2$)
- $B_n(\theta_0, \varepsilon; 2) = \{\theta: K(\theta_0, \theta) \leq \varepsilon^2, V_2(\theta_0, \theta) \leq \varepsilon^2\}$

\[
\begin{align*}
    h(\theta, \theta')^2 &= \int (\sqrt{p_\theta} - \sqrt{p_{\theta'}})^2 \, d\mu \\
    K(\theta, \theta') &= P_\theta \log(p_\theta / p_{\theta'}) \\
    V_2(\theta, \theta') &= P_\theta (\log(p_\theta / p_{\theta'}))^2
\end{align*}
\]
Independent observations

Data $X_1, \ldots, X_n$, independent with $X_i \sim p_{\theta,i}$

MAIN RESULT HOLDS WITH

- $d_n^2(\theta, \theta') = \frac{1}{n} \sum_{i=1}^{n} h_i(\theta, \theta')^2$
- $B_n(\theta_0, \varepsilon; 2) = \{ \theta : \frac{1}{n} \sum_{i=1}^{n} K_i(\theta_0, \theta) \vee \frac{1}{n} \sum_{i=1}^{n} V_{2,i}(\theta_0, \theta) \leq \varepsilon^2 \}$

$h_i$, $K_i$ and $V_{2,i}$ computed for $i$th observation
Markov chains

Data \((X_0, X_1, \ldots, X_n)\) for \(\cdots, X_0, X_1, X_2, \cdots\) stationary
Markov chain with initial density \(q_\theta\) and transition density \(p_\theta(\cdot|\cdot)\).

Assume \(\exists\) integrable \(r\), constants \(0 < c < C\) and \(k > 2\):

1. \(c r(y) \leq p_\theta(y|x) \leq C r(y),\)
2. \(\alpha\)-mixing, \(\sum_{h=0}^{\infty} \alpha_h^{1-1/k} < \infty\)

**MAIN RESULT HOLDS WITH**

\[
\begin{align*}
\left|\frac{1}{n} \sum_{i=1}^{n} \left[ \log \frac{p_{\theta_0}}{p_{\theta}}(X_i|X_{i-1}) \right] - \log \frac{p_{\theta_0}}{p_{\theta}}(X_1|X_0) \right|^k &\leq \varepsilon^k \\
\end{align*}
\]
Gaussian white noise model

Data $(X_t^{(n)}: 0 \leq t \leq 1)$ for $dX_t^{(n)} = \theta(t) \, dt + n^{-1/2} \, dB_t$, where $B$ is Brownian motion

**MAIN RESULT HOLDS WITH**

- $d_n$: $L_2$-norm
- $B_n(\theta_0, \varepsilon; 2)$: $L_2$-ball
Gaussian time series

Data \((X_0, X_1, \ldots, X_n)\) for \(\cdots, X_0, X_1, X_2, \cdots\) stationary mean zero Gaussian process with spectral density \(\theta \in \Theta\)

Assume

1. \(\sup_{\theta \in \Theta} \| \log \theta \|_\infty < \infty\)
2. \(\sup_{\theta \in \Theta} \sum_{h=-\infty}^{\infty} |h|(E_\theta X_h X_0)^2 < \infty\)

MAIN RESULT HOLDS WITH

- \(d_n: L_2\)-norm, \(\bar{d}_n: \text{supremum-norm}\)
- \(B_n(\theta_0, \varepsilon; 2): L_2\)-ball
Ergodic diffusions

Data \((X_t: 0 \leq t \leq n)\) for \(X\) solution to
\[dX_t = \theta(X_t) \, dt + \sigma(X_t) \, dB_t,\]
where \(B\) is Brownian motion \(B\)

Assume

1. stationary ergodic, state space \(I\),
2. stationary measure \(\mu_{\theta_0}\)

MAIN RESULT HOLDS WITH

\[d(\theta, \theta') = \| (\theta - \theta') \|_{\mu_{\theta_0}, 2} \quad J \subset I\]
\[e(\theta, \theta') = \| (\theta - \theta') / \sigma \|_{\mu_{\theta_0}, 2}\]
\[B(\theta_0, \varepsilon; 2) \| \cdot / \sigma \|_{\mu_{\theta_0}, 2} \text{-ball}\]
Examples of priors
Uniform priors on $\varepsilon_n$-nets
Uniform priors on $\varepsilon_n$-nets

Smooth Euclidean prior on the parameters in a finite-dimensional approximation (e.g. series approximation, finite mixture density)
Priors

Uniform priors on $\varepsilon_n$-nets

Smooth Euclidean prior on the parameters in a finite-dimensional approximation (e.g. series approximation, finite mixture density)

Random measures, such as Ferguson’s Dirichlet or Polya trees
Uniform priors on $\varepsilon_n$-nets

Smooth Euclidean prior on the parameters in a finite-dimensional approximation (e.g. series approximation, finite mixture density)

Random measures, such as Ferguson’s Dirichlet or Polya trees

A stochastic process as model for a function, e.g. a Gaussian process or a Lévy process
Uniform priors on $\varepsilon_n$-nets

Smooth Euclidean prior on the parameters in a finite-dimensional approximation (e.g. series approximation, finite mixture density)

Random measures, such as Ferguson’s Dirichlet or Polya trees

A stochastic process as model for a function, e.g. a Gaussian process or a Lévy process

Combination of the previous as building blocks, e.g. mixtures
Gaussian priors
Setting

Data $X^{(n)}$ follows density $p_{w_0}^{(n)}$ indexed by a function $w_0: T \rightarrow \mathbb{R}$

Prior $\Pi_n$ for $w$ is law of Gaussian process $(W_t: t \in T)$

Form posterior as before

$$\Pi_n(B|X^{(n)}): = \frac{\int_B p_{w}^{(n)}(X^{(n)}) d\Pi_n(w)}{\int p_{w}^{(n)}(X^{(n)}) d\Pi_n(w)}$$
Data $X^{(n)}$ follows density $p_{w_0}^{(n)}$ indexed by a function $w_0: T \rightarrow \mathbb{R}$

Prior $\Pi_n$ for $w$ is law of Gaussian process $(W_t: t \in T)$

Form posterior as before

$$\Pi_n(B|X^{(n)}) = \frac{\int_B p_w^{(n)}(X^{(n)}) d\Pi_n(w)}{\int p_w^{(n)}(X^{(n)}) d\Pi_n(w)}$$

Rate of contraction is at least $\varepsilon_n$ if $\forall M_n \rightarrow \infty$

$$P_{w_0}^{(n)} \Pi_n(w: d_n(w, w_0) \geq M_n \varepsilon_n | X^{(n)}) \rightarrow 0$$
Reproducing kernel Hilbert space

To every Gaussian random element with values in a Banach space \((\mathcal{B}, \| \cdot \|)\) is attached a certain Hilbert space \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), called the RKHS.
To every Gaussian random element with values in a Banach space \((B, \| \cdot \|)\) is attached a certain Hilbert space \((H, \| \cdot \|_H)\), called the RKHS

\[ \| \cdot \|_H \text{ is stronger than } \| \cdot \| \text{ and can view } H \subset B \]
Reproducing kernel Hilbert space

To every Gaussian random element with values in a Banach space \((\mathcal{B}, \| \cdot \|)\) is attached a certain Hilbert space \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), called the RKHS

\[ \| \cdot \|_\mathcal{H} \text{ is stronger than } \| \cdot \| \text{ and can view } \mathcal{H} \subset \mathcal{B} \]

**EXAMPLE**

The RKHS of Brownian motion as map in \(C[0,1]\) is

\[ \mathcal{H} = \{ h : \int h'(t)^2 \, dt < \infty \} \text{ with norm } \| h \|_\mathcal{H} = \| h' \|_2 \]
Small ball probability

$W$ Gaussian map in $(\mathcal{B}, \| \cdot \|)$

$\mathbb{P}(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$
Small ball probability

$W$ Gaussian map in $(\mathcal{B}, \| \cdot \|)$

$$P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$$

Small ball probability can be computed for many examples, either by probabilistic arguments, or by using:

**THEOREM** (Kuelbs and Li, 1993)

$$\phi_0(\varepsilon) \asymp \log N(\varepsilon/\sqrt{\phi_0(\varepsilon)}, \mathbb{H}_1, \| \cdot \|)$$

for $\mathbb{H}_1$ the unit ball of the RKHS

up to factors of 2 and regularity
Concentration function

\[ W \text{ Gaussian map in } (\mathbb{B}, \| \cdot \|) \text{ with RKHS } (\mathbb{H}, \| \cdot \|_{\mathbb{H}}) \]
\[ P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)} \]

\[ \phi_{w_0}(\varepsilon) = \phi_0(\varepsilon) + \inf_{h \in \mathbb{H} : \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2 \]
**Concentration function**

$W$ Gaussian map in $(\mathbb{B}, \| \cdot \|)$ with RKHS $(\mathbb{H}, \| \cdot \|_\mathbb{H})$

$$P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$$

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

**THEOREM** (Kuelbs and Li, 1993)

Concentration function measures concentration around $w_0$:

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

up to factors 2
Main result

\( W \) Gaussian map in \((\mathcal{B}, \| \cdot \|), \text{RKHS} (\mathcal{H}, \| \cdot \|_{\mathcal{H}})\)

\[ P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)} \]

Assume that various distances on the model combine “appropriately” with the norm \( \| \cdot \| \) on \( W \) (see below) and that \( \varepsilon_n \gg 1/\sqrt{n} \)

**THEOREM**

Posterior rate is \( \varepsilon_n \) if \( \phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2 \), i.e. if

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

First depends on \( W \) and not on \( w_0 \)
Toy problem-Brownian motion

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

**Intuition**

Support is full space *(if started at random)*

Sample paths are $1/2$-smooth

So BM is appropriate prior if $w_0 \in C^\alpha[0, 1]$ for $\alpha = 1/2$

If $w_0$ smoother than $1/2$: BM too spread out

If $w_0$ coarser than $1/2$: BM too smooth to be close

In fact: 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 $\alpha = 1/2$
Toy problem-Brownian motion

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

**Mathematics**

Small ball probability $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$

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

**LEMMA**

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

$$\inf_{h \in \mathbb{H}} \| h - w_0 \|_\infty < \varepsilon \quad \| h \|_{\mathbb{H}}^2 \asymp (1/\varepsilon)^{(2-2\alpha)/\alpha}$$
Toy problem-Brownian motion

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

**Mathematics**

Small ball probability $\phi_0(\varepsilon) \asymp (1/\varepsilon)^2$

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

**Lemma**

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

$$\inf_{h \in \mathbb{H}}: \|h - w_0\|_\infty < \varepsilon \|h\|_\mathbb{H}^2 \asymp (1/\varepsilon)^{(2-2\alpha)/\alpha}$$

**Consequence:**

Rate is $\varepsilon_n$ if

$$(1/\varepsilon_n)^2 \leq n\varepsilon_n^2 \text{ AND } (1/\varepsilon_n)^{(2-2\alpha)/\alpha} \leq n\varepsilon_n^2$$

First implies $\varepsilon_n \geq n^{-1/4}$ for any $w_0$.

Second implies $\varepsilon_n \geq n^{-\alpha/2}$ for $w_0 \in C^\alpha[0, 1]$
Gaussian priors-settings

Main
result-remember
Density estimation
Classification
Regression
Gaussian white noise

Gaussian priors-proof
Gaussian priors-examples

Adaptation
Main result-remember

\[ W \text{ Gaussian map in } (\mathcal{B}, \| \cdot \|), \text{ RKHS } (\mathcal{H}, \| \cdot \|_\mathcal{H}) \]
\[ P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)} \]

Assume that various distances on the model combine
“appropriately” with the norm \( \| \cdot \| \) on \( W \) (see below) and that
\( \varepsilon_n \gg 1/\sqrt{n} \)

THEOREM
Posterior rate is \( \varepsilon_n \) if \( \phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2 \), i.e.
\[ \inf_{h \in \mathcal{H}}: \| h - w_0 \| < \varepsilon_n \| h \|_{\mathcal{H}}^2 \leq n\varepsilon_n^2 \quad \text{AND} \quad \phi_0(\varepsilon_n) \leq n\varepsilon_n^2 \]
Data $X_1, \ldots, X_n$ i.i.d. from density on $[0, 1]$

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

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

Data $X_1, \ldots, X_n$ i.i.d. from density on $[0, 1]$

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

- Distance on parameter: Hellinger distance on $p_w$
- Norm on $W$: uniform

**LEMMA**

$\forall v, w$

- $h(p_v, p_w) \leq \|v - w\|_{\infty} e^{\|v-w\|_{\infty}} / 2$
- $K(p_v, p_w) \lesssim \|v - w\|_{\infty}^2 e^{\|v-w\|_{\infty}} (1 + \|v - w\|_{\infty})$
- $V(p_v, p_w) \lesssim \|v - w\|_{\infty}^2 e^{\|v-w\|_{\infty}} (1 + \|v - w\|_{\infty})^2$
Classification

Data \((X_1, Y_1), \ldots, (X_n, Y_n)\) i.i.d. in \([0, 1] \times \{0, 1\}\)

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

E.g. \(\Psi\) 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 $Y_1, \ldots, Y_n$

$Y_i = w_0(x_i) + e_i$

$x_1, \ldots, x_n$ fixed design points

$e_1, \ldots, e_n$ i.i.d. Gaussian mean-zero errors

- Distance on parameter: empirical $L_2$-distance on $w$
- Norm on $W$: uniform

Can use posterior for Gaussian errors also if errors have only mean zero? (Kleijn & vdV, 2006)
**Gaussian white noise**

**Data** \( (X_t: t \in [0, 1]) \)

\[
dX_t = w_t + n^{-1/2} dB_t
\]

- **Distance on parameter:** \( L_2 \)
- **Norm on \( W \):** \( L_2 \)
Gaussian priors-proof
$W$ Gaussian map in $(\mathbb{B}, \| \cdot \|)$, RKHS $(\mathbb{H}, \| \cdot \|_\mathbb{H})$

$P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)}$

Assume that various distances on the model combine “appropriately” with the norm $\| \cdot \|$ on $W$ (see below) and that $\varepsilon_n \gg 1/\sqrt{n}$

**THEOREM**

Posterior rate is $\varepsilon_n$ if $\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$, i.e.

$$\inf_{h \in \mathbb{H}}: \|h - w_0\| < \varepsilon_n \|h\|_\mathbb{H}^2 \leq n\varepsilon_n^2 \quad \text{AND} \quad \phi_0(\varepsilon_n) \leq n\varepsilon_n^2$$
$W$ zero-mean Gaussian in $(\mathcal{B}, \| \cdot \|)$

$S: \mathcal{B}^* \to \mathcal{B}, \quad Sb^* = EWb^*(W)$

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

$$\langle Sb_1^*, Sb_2^* \rangle_\mathcal{H} = Eb_1^*(W)b_2^*(W)$$
\[ W = (W_x : x \in \mathcal{X}) \] Gaussian stochastic process which can be seen as tight, Borel measurable map in \[ \ell^\infty(\mathcal{X}) = \{ f : \mathcal{X} \to \mathbb{R} : \sup_x |f(x)| < \infty \} \]

Covariance function \[ K(x, y) = \mathbb{E}W_xW_y \]

Then RKHS is completion of the set of functions \[ x \mapsto \sum_i \alpha_i K(y_i, x) \] relative to inner product \[ \langle \sum_i \alpha_i K(y_i, \cdot), \sum_j \beta_j K(z_j, \cdot) \rangle_{\mathcal{H}} = \sum_i \sum_j \alpha_i \beta_j K(y_i, z_j) \]
RKHS gives the “geometry” of the support of $W$

**THEOREM**

Norm closure of $\mathcal{H}$ in $\mathcal{B}$ is smallest closed set with probability one under Gaussian measure
RKHS gives the “geometry” of the support of $W$

**THEOREM**
Norm closure of $\mathcal{H}$ in $\mathcal{B}$ is smallest closed set with probability one under Gaussian measure

**CONSEQUENCE:** posterior inconsistent if $\|w_0 - \mathcal{H}\| > 0$
RKHS gives the “geometry” of the support of $W$

THEOREM
Norm closure of $\mathcal{H}$ in $\mathcal{B}$ is smallest closed set with probability one under Gaussian measure

THEOREM (Kuelbs & Li, 1993)
For $N(\varepsilon, \mathbb{H}_1, \| \cdot \|)$ minimal number of balls needed to cover unit ball of RKHS:

$$\phi_0(\varepsilon) \approx \log N(\varepsilon / \sqrt{\phi_0(\varepsilon)}, \mathbb{H}_1, \| \cdot \|)$$
**Geometry**

RKHS gives the “geometry” of the support of $W$

**THEOREM**
Norm closure of $\mathbb{H}$ in $\mathbb{B}$ is smallest closed set with probability one under Gaussian measure

**THEOREM (Kuelbs & Li, 1993)**
For $N(\varepsilon, \mathbb{H}_1, \| \cdot \|)$ minimal number of balls needed to cover unit ball of RKHS:

$$\phi_0(\varepsilon) \asymp \log N(\varepsilon/\sqrt{\phi_0(\varepsilon)}, \mathbb{H}_1, \| \cdot \|)$$

**THEOREM (Borell, 1975)**

$$P(W \notin \varepsilon \mathbb{B}_1 + M \mathbb{H}_1) \leq 1 - \Phi^{-1}(e^{-\phi_0(\varepsilon)}) + M$$
Sufficient for posterior rate of $\varepsilon_n$ is existence of sets $\mathbb{B}_n$ with

- $\log N(\varepsilon_n, \mathbb{B}_n, d_n) \leq n\varepsilon_n^2$ \hspace{1cm} \text{entropy}
- $\Pi_n(\mathbb{B}_n) = 1 - o(e^{-3n\varepsilon_n^2})$ \hspace{1cm} \text{prior mass}
- $\Pi_n(B_n(w_0, \varepsilon_n)) \geq e^{-n\varepsilon_n^2}$

$B_n(w_0, \varepsilon)$ is Kullback-Leibler neighborhood of $P_{w_0}^{(n)}$ 
(Ghosal & vdV, 2000, 2006)
Proof (2)

\[ W \text{ Gaussian map in } (\mathcal{B}, \| \cdot \|), \text{ RKHS } (\mathbb{H}, \| \cdot \|_{\mathbb{H}}) \]
\[ P(\|W\| < \varepsilon) = e^{-\phi_0(\varepsilon)} \]

**THEOREM**
\[ \forall w_0 \in \mathbb{H} \text{ and } \varepsilon_n > 0 \text{ with } \]
\[ \inf_{h \in \mathbb{H}} \| h - w_0 \| < \varepsilon_n \| h \|^2_{\mathbb{H}} + \phi_0(\varepsilon_n) \leq n\varepsilon_n^2 \]
\[ \exists \mathcal{B}_n \subset \mathcal{B} \text{ with } \]
\[ - \log N(\varepsilon_n, \mathcal{B}_n, \| \cdot \|) \preceq n\varepsilon_n^2 \]
\[ P(W \notin \mathcal{B}_n) \preceq e^{-4n\varepsilon_n^2} \]
\[ P(\|W - w_0\| < \varepsilon_n) \succeq e^{-n\varepsilon_n^2} \]

**PROOF**
Take \( \mathcal{B}_n = M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1 \) for appropriate \( M_n \)

Use Borell’s inequality
Gaussian priors-examples
Brownian motion

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

BM is appropriate prior if truth is $1/2$-smooth

If truth smoother than $1/2$: BM too spread out
If truth coarser than $1/2$: BM too smooth to be close

In both cases can obtain better rates with other priors
Let $I_{0+}^k$ denote $k$ times integration from 0 and

$$W_t = (I_{0+}^k B)_t + \sum_{j=0}^{k} Z_j t^j$$

[B Brownian motion, $(Z_j)$ iid $N(0, 1)$]

Gives appropriate model for $k + 1/2$-smooth functions
Spline smoothing in regression

\[ W_t = \sqrt{b} (I_{0+}^k B)_t + \sqrt{a} \sum_{j=0}^{k} Z_j t^j \]

If \( a \to \infty \) and \( b, n \) are fixed, then the posterior mean tends to the minimizer of

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

(Kimeldorf and Wahba, 1970, Wahba, 1978)

If \( w_0 \in H^k[0, 1] \) and \( \sigma^2/nb \sim n^{-2k/(2k+1)} \), i.e. \( b \sim n^{-1/(2k+1)} \), the penalized least squares estimator is rate optimal.
Riemann-Liouville process

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

\([B \text{ Brownian motion}, \alpha > 0, (Z_k) \text{ iid } N(0, 1)]\)

“Fractional integral”

Gives appropriate models for \(\alpha\)-smooth functions
Fractional Brownian motion

$W$ zero-mean Gaussian with
\[
\text{cov}(W_s, W_t) = s^{2\alpha} + t^{2\alpha} - |t - s|^{2\alpha}
\]

[Hurst index $0 < \alpha < 1$]

Gives appropriate model for $\alpha$-smooth functions

Can integrate this to cover $\alpha > 1$
Expansions

Infinite series
\[ W_t = \sum_{i=1}^{\infty} \lambda_i Z_i e_i(t) \]
[(\(e_i\) basis, (\(Z_i\)) i.i.d. \(N(0, 1)\), \(\lambda_i \to 0\)]
[example: eigen expansion]
RKHS \(\{\sum w_i e_i : \sum_i w_i / \lambda_i^2 < \infty\}\)

Truncated series
\[ W_t = \sum_{i=1}^{N} \mu_i Z_i e_i(t) \]
[(\(e_i\) basis, (\(Z_i\)) i.i.d. \(N(0, 1)\), \(\mu_i \to 0\)]

Appropriate \((\lambda_i)\) or \(N \to \infty\) and \((\mu_i)\) give proper models for \(\alpha\)-smooth functions
Rescaled Brownian motion

\[ W_t = \frac{B_t}{c} \] for \( B \) Brownian motion, \( t \in [0, 1] \) and
\[ c \sim n^{(2\alpha-1)/(2\alpha+1)} \]

\( \alpha < 1/2: \ 1/c \rightarrow \infty \) (shrink)
\( \alpha \in (1/2, 1]: \ 1/c \rightarrow 0 \) (stretch)

Gives optimal rate for \( w_0 \in C^\alpha[0, 1] \), any \( \alpha \in (0, 1] \)

Surprising? (Brownian motion is self-similar!)
“On infinitesimal intervals BM looks like a function in its RKHS”
Rescaled integrated Brownian motion

\[ W_t = (I_{0+}^k B)_{t/c} + \sum_{j=0}^{k} Z_j t^j, \ t \in [0, 1] \ \text{and} \ c \sim \mathcal{N}(\alpha-k-1/2)/(2\alpha+1)(k+1/2) \]

Gives optimal rate for \( w_0 \in C^\alpha[0, 1], \ \text{any} \ \alpha \in (0, k + 1] \)
Rescaled stationary process

\[ W_t = G_t/c_n \] for a centered Gaussian process \( G \) with

\[ \mathbb{E}G_sG_t = \phi(s - t), \quad c_n = n^{1/(2\alpha+1)} \]

if \( \int e^{\gamma t} F \phi(t) \, dt < \infty \) for some \( \gamma > 0 \), then prior gives optimal rate for \( w_0 \in C^\alpha[0, 1] \) up to a \( \log n \)-factor, \( \alpha > 0 \)
Adaptation
Adaptation

For every level of regularity $\alpha$ there is an optimal prior.

Construct an overall prior in two steps:

- Sample a regularity level $\alpha$ from a prior on $\left(0, \infty\right)$
- Given $\alpha$, choose $w$ from the prior that is optimal for $\alpha$

The Bayesian machine will make the data choose the $\alpha$ that is appropriate for the data?
For $n = 1, 2, \ldots$ and every $\alpha$ in a arbitrary countable set $A_n$ let $\Pi_{n,\alpha}$ a prior on a model $P_{n,\alpha}$ and let $\varepsilon_{n,\alpha}$ a rate such that
\[
\log N(\varepsilon_{n,\alpha}, P_{n,\alpha}, d_n) \lesssim n\varepsilon_{n,\alpha}^2
\]

Let $\lambda_n$ a probability measure on $A_n$ such that
\[
\lambda_n\{\alpha\} \propto \mu_\alpha e^{-Cn\varepsilon_{n,\alpha}^2}
\]

**THEOREM (Lember, vdV, 2005, 2007)**

If $\sum_\alpha \sqrt{\mu_\alpha} < \infty$ and $\sum_\alpha (\mu_\alpha / \mu_\beta) e^{-Cn\varepsilon_{n,\alpha}^2 / 4} = O(1)$, then the posterior rate is at least $\varepsilon_{n,\beta}$ for any $\beta$ such that
\[
\Pi_{n,\beta}(B_n(\varepsilon_{n,\beta})) \geq e^{-Fn\varepsilon_{n,\beta}^2}.
\]
Adaptation works for more general weight functions $\lambda_n$ under more complicated conditions.

However the choice of weights $\lambda_n$ on $A_n$ and priors $\Pi_{n,\alpha}$ “interact”

For Gaussian process priors it appears that the choice of weights $\lambda_n$ is inessential (v Zanten, 2008)
With the rescaled processes we put the hyper prior on the scale:

- Choose $c$ from some prior on $(0, \infty)$
- Given $c$ choose $W \sim G/c$

Does it work????