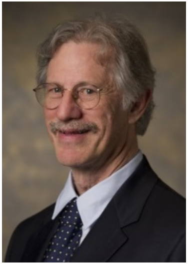
Part III

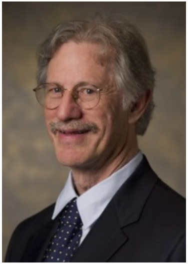


Average Case ϵ-Complexity in Computer Science: A Bayesian View

J. B. KADANE, Carnegie-Mellon University; G. W. WASILKOWSKI, Columbia University/

In Bayesian Statistics 2, Proceedings of the Second Valencia International Meeting (pp. 361–374), 1985.

> *Relations between average case ϵ-complexity and Bayesian Statistics are discussed. An algorithm corresponds to a decision function and the choice of information to the choice of an experiment. [...] We hope that the relation reported here can lead to further fruitful results for both fields.*



Average Case ϵ-Complexity in Computer Science: A Bayesian View

J. B. KADANE, Carnegie-Mellon University; G. W. WASILKOWSKI, Columbia University/

In Bayesian Statistics 2, Proceedings of the Second Valencia International Meeting (pp. 361–374), 1985.

> *Relations between average case ϵ-complexity and Bayesian Statistics are discussed. An algorithm corresponds to a decision function and the choice of information to the choice of an experiment. [...] We hope that the relation reported here can lead to further fruitful results for both fields.*

The *Bayesian* approach, popularised in Stuart (2010), can be used:

- a *prior* measure P_x is placed on $\mathcal{X}$
- a *posterior* measure $P_{x|a}$ is defined as the "restriction of P_x to those functions $x \in \mathcal{X}$ for which

$$A(x) = a \qquad \text{e.g.} \quad A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_n) \end{bmatrix} = a$$

is satisfied" (to be formalised).

$\implies$ Principled and general uncertainty quantification for numerical methods.

The *Bayesian* approach, popularised in Stuart (2010), can be used:

- a *prior* measure P_x is placed on $\mathcal{X}$
- a *posterior* measure $P_{x|a}$ is defined as the "restriction of P_x to those functions $x \in \mathcal{X}$ for which

$$A(x) = a \qquad \text{e.g.} \quad A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_n) \end{bmatrix} = a$$

is satisfied" (to be formalised).

$\implies$ Principled and general uncertainty quantification for numerical methods.

The *Bayesian* approach, popularised in Stuart (2010), can be used:

- a *prior* measure P_x is placed on $\mathcal{X}$
- a *posterior* measure $P_{x|a}$ is defined as the "restriction of P_x to those functions $x \in \mathcal{X}$ for which

$$A(x) = a \qquad \text{e.g.} \quad A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_n) \end{bmatrix} = a$$

is satisfied" (to be formalised).

$\Longrightarrow$ Principled and general uncertainty quantification for numerical methods.

The *Bayesian* approach, popularised in Stuart (2010), can be used:

- a *prior* measure P_x is placed on $\mathcal{X}$
- a *posterior* measure $P_{x|a}$ is defined as the "restriction of P_x to those functions $x \in \mathcal{X}$ for which

$$A(x) = a \qquad \text{e.g.} \quad A(x) = \begin{bmatrix} -\Delta x(t_1) \\ \vdots \\ -\Delta x(t_n) \end{bmatrix} = a$$

is satisfied" (to be formalised).

$\implies$ Principled and general uncertainty quantification for numerical methods.

Sixth Job: Analysis of the Gaussian Case

Gaussian process

$\downarrow$

covariance function k $\leftrightarrow$ kernel function k

$\uparrow$

Reproducing kernel Hilbert space

Let $\mathcal{X}$ be a Hilbert space (i.e. a complete inner product space) of real-valued functions on D. Let $L_t : x \mapsto x(t)$ denote the evaluation functional at a point $t \in D$. Then $\mathcal{X}$ is a *reproducing kernel Hilbert space* (RKHS) if there exists C such that

$$|L_t x| \leq C \|x\|_{\mathcal{X}}$$

for all $x \in \mathcal{X}$.

Riesz Representation Theorem

Let $\mathcal{X}^*$ denote the dual of $\mathcal{X}$ (i.e. the space of continuous linear functionals on $\mathcal{X}$). Then $x \mapsto \langle \cdot, x \rangle_{\mathcal{X}}$ is an isometric isomorphism from $\mathcal{X}$ to $\mathcal{X}^*$.

Since L_t is an element of $\mathcal{X}^*$, there exists an element k_t of $\mathcal{X}$ such that $L_t = \langle \cdot, k_t \rangle_{\mathcal{X}}$. This allows us to define the kernel $k(t, t') = \langle k_t, k_{t'} \rangle_{\mathcal{X}}$.

It can be shown that k *characterises* $\mathcal{X}$. The relation $x(t) = \langle x, k(\cdot, t) \rangle_{\mathcal{X}}$ is called the reproducing property.

Riesz Representation Theorem

Let $\mathcal{X}^*$ denote the dual of $\mathcal{X}$ (i.e. the space of continuous linear functionals on $\mathcal{X}$). Then $x \mapsto \langle \cdot, x \rangle_{\mathcal{X}}$ is an isometric isomorphism from $\mathcal{X}$ to $\mathcal{X}^*$.

Since L_t is an element of $\mathcal{X}^*$, there exists an element k_t of $\mathcal{X}$ such that $L_t = \langle \cdot, k_t \rangle_{\mathcal{X}}$. This allows us to define the kernel $k(t, t') = \langle k_t, k_{t'} \rangle_{\mathcal{X}}$.

It can be shown that k *characterises* $\mathcal{X}$. The relation $x(t) = \langle x, k(\cdot, t) \rangle_{\mathcal{X}}$ is called the reproducing property.

Reproducing Kernel Hilbert Spaces

Riesz Representation Theorem

Let $\mathcal{X}^*$ denote the dual of $\mathcal{X}$ (i.e. the space of continuous linear functionals on $\mathcal{X}$). Then $x \mapsto \langle \cdot, x \rangle_{\mathcal{X}}$ is an isometric isomorphism from $\mathcal{X}$ to $\mathcal{X}^*$.

Since L_t is an element of $\mathcal{X}^*$, there exists an element k_t of $\mathcal{X}$ such that $L_t = \langle \cdot, k_t \rangle_{\mathcal{X}}$. This allows us to define the kernel $k(t, t') = \langle k_t, k_{t'} \rangle_{\mathcal{X}}$.

It can be shown that k *characterises* $\mathcal{X}$. The relation $x(t) = \langle x, k(\cdot, t) \rangle_{\mathcal{X}}$ is called the reproducing property.

The native space of an RKHS is

$$\{x : D \to \mathbb{R} \ : \ \|x\|_{\mathcal{X}} < \infty\}.$$

How is the native space related to the kernel?

Recall, from Mercer's theorem if $\int \sqrt{k(t,t)} \mathrm{d}\nu(t) < \infty$, then

$$k(t,t') = \sum_{i=1}^{\infty} \lambda_i \psi_i(t) \psi_i(t').$$

Then the native space of the RKHS associated to k is:

$$\left\{ x = \sum_{i=1}^{\infty} c_i \lambda_i^{\frac{1}{2}} \psi_i \ : \ \|x\|_{\mathcal{X}}^2 = \sum_{i=1}^{\infty} c_i^2 < \infty \right\}$$

The native space of an RKHS is

$$\{x : D \to \mathbb{R} \ : \ \|x\|_{\mathcal{X}} < \infty\}.$$

How is the native space related to the kernel?

Recall, from Mercer's theorem if $\int \sqrt{k(t,t)}\mathrm{d}\nu(t) < \infty$, then

$$k(t,t') = \sum_{i=1}^{\infty} \lambda_i \psi_i(t) \psi_i(t').$$

Then the native space of the RKHS associated to k is:

$$\left\{ x = \sum_{i=1}^{\infty} c_i \lambda_i^{\frac{1}{2}} \psi_i \ : \ \|x\|_{\mathcal{X}}^2 = \sum_{i=1}^{\infty} c_i^2 < \infty \right\}$$

The native space of an RKHS is

$$\{x : D \to \mathbb{R} \ : \ \|x\|_{\mathcal{X}} < \infty\}.$$

How is the native space related to the kernel?

Recall, from Mercer's theorem if $\int \sqrt{k(t,t)} \mathrm{d}\nu(t) < \infty$, then

$$k(t, t') = \sum_{i=1}^{\infty} \lambda_i \psi_i(t) \psi_i(t').$$

Then the native space of the RKHS associated to k is:

$$\left\{ x = \sum_{i=1}^{\infty} c_i \lambda_i^{\frac{1}{2}} \psi_i \ : \ \|x\|_{\mathcal{X}}^2 = \sum_{i=1}^{\infty} c_i^2 < \infty \right\}$$

Approximation in RKHS

Examples of native spaces (notation: $z_+^k := (\max(0,z))^k$):

Kernel $k(t,t')$	Native Space
$\exp(-\|t-t'\|^2)$	$\cap_{m\in\mathbb{N}} H^m(D)$
$(c^2+\|t-t'\|^2)^{-\beta}$, $\beta > \frac{d}{2}$	$H^{\beta-\frac{d}{2}}(D)$
$(1-\|t-t'\|)_+^2$	$H^{\frac{d}{2}+\frac{1}{2}}(D)$
$(1-\|t-t'\|)_+^4(4\|t-t'\|+1)$	$H^{\frac{d}{2}+\frac{3}{2}}(D)$

Here

$$H^m(D) = \left\{ x : D \to \mathbb{R} \quad \text{s.t.} \quad \|x\|_{H^m(D)}^2 = \sum_{|\alpha|\leq m} \|D^\alpha x\|_{L^2(D)}^2 < \infty \right\}$$

is the Sobolev space of order $m \in \mathbb{N}$. It is well-defined for $m > \frac{d}{2}$.

Notation: $|\alpha| = \alpha_1 + \cdots + \alpha_d$, $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_m^{\alpha_m}}$, $\|x\|_{L^2(D)}^2 := \int x(t)^2 \mathrm{d}t$.

Approximation in RKHS

Examples of native spaces (notation: $z_+^k := (\max(0,z))^k$):

Kernel $k(t,t')$	Native Space
$\exp(-\|t-t'\|^2)$	$\cap_{m\in\mathbb{N}} H^m(D)$
$(c^2+\|t-t'\|^2)^{-\beta}$, $\beta > \frac{d}{2}$	$H^{\beta-\frac{d}{2}}(D)$
$(1-\|t-t'\|)_+^2$	$H^{\frac{d}{2}+\frac{1}{2}}(D)$
$(1-\|t-t'\|)_+^4(4\|t-t'\|+1)$	$H^{\frac{d}{2}+\frac{3}{2}}(D)$

Here

$$H^m(D) = \left\{ x : D \to \mathbb{R} \quad \text{s.t.} \quad \|x\|_{H^m(D)}^2 = \sum_{|\alpha|\leq m} \|D^\alpha x\|_{L^2(D)}^2 < \infty \right\}$$

is the Sobolev space of order $m \in \mathbb{N}$. It is well-defined for $m > \frac{d}{2}$.

Notation: $|\alpha| = \alpha_1 + \cdots + \alpha_d$, $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_m^{\alpha_m}}$, $\|x\|_{L^2(D)}^2 := \int x(t)^2 \mathrm{d}t$.

Examples of native spaces (notation: $z_+^k := (\max(0,z))^k$):

Kernel $k(t,t')$	Native Space
$\exp(-\|t-t'\|^2)$	$\cap_{m\in\mathbb{N}} H^m(D)$
$(c^2+\|t-t'\|^2)^{-\beta}$, $\beta > \frac{d}{2}$	$H^{\beta-\frac{d}{2}}(D)$
$(1-\|t-t'\|)_+^2$	$H^{\frac{d}{2}+\frac{1}{2}}(D)$
$(1-\|t-t'\|)_+^4(4\|t-t'\|+1)$	$H^{\frac{d}{2}+\frac{3}{2}}(D)$

Here

$$H^m(D) = \left\{ x : D \to \mathbb{R} \quad \text{s.t.} \quad \|x\|_{H^m(D)}^2 = \sum_{|\alpha|\leq m} \|D^\alpha x\|_{L^2(D)}^2 < \infty \right\}$$

is the Sobolev space of order $m \in \mathbb{N}$. It is well-defined for $m > \frac{d}{2}$.

Notation: $|\alpha| = \alpha_1 + \cdots + \alpha_d$, $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_m^{\alpha_m}}$, $\|x\|_{L^2(D)}^2 := \int x(t)^2 \mathrm{d}t$.

Approximation in RKHS

Consider the task of estimation of $x \in \mathcal{X}$ based on the information that

$$\begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n. \end{aligned}$$

This is clearly ill-posed if $\dim(\mathcal{X}) > n$.

Consider instead the regularised problem:

$$\hat{x} \quad := \quad \arg\inf_{x \in \mathcal{X}} \|x\|_{\mathcal{X}} \quad \text{s.t.} \quad \begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n \end{aligned} \quad .$$

What is the relevance of the *interpolant* $\hat{x}$? It is the posterior mean under the Gaussian process prior $P_x = \mathcal{GP}(0, k)$ combined with the data $\{(t_i, x(t_i))\}_{i=1}^n$.

Approximation in RKHS

Consider the task of estimation of $x \in \mathcal{X}$ based on the information that

$$\begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n. \end{aligned}$$

This is clearly ill-posed if $\dim(\mathcal{X}) > n$.

Consider instead the regularised problem:

$$\hat{x} \quad := \quad \arg\inf_{x \in \mathcal{X}} \|x\|_{\mathcal{X}} \quad \text{s.t.} \quad \begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n \end{aligned} .$$

What is the relevance of the *interpolant* $\hat{x}$? It is the posterior mean under the Gaussian process prior $P_x = \mathcal{GP}(0, k)$ combined with the data $\{(t_i, x(t_i))\}_{i=1}^n$.

Consider the task of estimation of $x \in \mathcal{X}$ based on the information that

$$\begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n. \end{aligned}$$

This is clearly ill-posed if $\dim(\mathcal{X}) > n$.

Consider instead the regularised problem:

$$\hat{x} := \arg\inf_{x \in \mathcal{X}} \|x\|_{\mathcal{X}} \quad \text{s.t.} \quad \begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n \end{aligned}.$$

What is the relevance of the *interpolant* $\hat{x}$? It is the posterior mean under the Gaussian process prior $P_x = \mathcal{GP}(0, k)$ combined with the data $\{(t_i, x(t_i))\}_{i=1}^n$.

Consider the task of estimation of $x \in \mathcal{X}$ based on the information that

$$\begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n. \end{aligned}$$

This is clearly ill-posed if $\dim(\mathcal{X}) > n$.

Consider instead the regularised problem:

$$\hat{x} \quad := \quad \arg\inf_{x \in \mathcal{X}} \|x\|_{\mathcal{X}} \quad \text{s.t.} \quad \begin{array}{rcl} x(t_1) & = & c_1 \\ & \vdots & \\ x(t_n) & = & c_n \end{array}.$$

What is the relevance of the *interpolant* $\hat{x}$? It is the posterior mean under the Gaussian process prior $P_x = \mathcal{GP}(0, k)$ combined with the data $\{(t_i, x(t_i))\}_{i=1}^n$.

Consider the task of estimation of $x \in \mathcal{X}$ based on the information that

$$\begin{aligned} x(t_1) &= c_1 \\ &\vdots \\ x(t_n) &= c_n. \end{aligned}$$

This is clearly ill-posed if $\dim(\mathcal{X}) > n$.

Consider instead the regularised problem:

$$\hat{x} \quad := \quad \underset{x \in \mathcal{X}}{\arg\inf} \|x\|_{\mathcal{X}} \quad \text{s.t.} \quad \begin{array}{rcl} x(t_1) & = & c_1 \\ & \vdots & \\ x(t_n) & = & c_n \end{array} .$$

What is the relevance of the *interpolant* $\hat{x}$? It is the posterior mean under the Gaussian process prior $P_x = \mathcal{GP}(0, k)$ combined with the data $\{(t_i, x(t_i))\}_{i=1}^n$.

Approximation in RKHS

In general $|\hat{x}(t) - x(t)| \leq p_{\mathcal{X}}(t_1, \ldots, t_n)\|x\|_{\mathcal{X}}$ where $p_{\mathcal{X}}$ is the power function associated to $\mathcal{X}$. Our aim now is to understand more about $p_{\mathcal{X}}$.

Consider the kernel matrix

$$K = \begin{bmatrix} k(t_1, t_1) & \ldots & k(t_1, t_n) \\ \vdots & & \vdots \\ k(t_n, t_1) & \ldots & k(t_n, t_n) \end{bmatrix}.$$

If K^{-1} exists then, from linear algebra, there exist functions such that

$$\varphi_i(t_j) = \delta_{ij}, \quad \varphi_i \in \text{span}\{k(\cdot, t_j), \ j = 1, \ldots, n\}.$$

Representer Theorem

The regularised estimate $\hat{x}$ is given by $\hat{x} = \sum_{i=1}^{n} c_i \varphi_i = \sum_{i=1}^{n} x(t_i) \varphi_i$.

Approximation in RKHS

In general $|\hat{x}(t) - x(t)| \leq p_{\mathcal{X}}(t_1, \dots, t_n)\|x\|_{\mathcal{X}}$ where $p_{\mathcal{X}}$ is the power function associated to $\mathcal{X}$. Our aim now is to understand more about $p_{\mathcal{X}}$.

Consider the kernel matrix

$$\boldsymbol{K} = \begin{bmatrix} k(t_1, t_1) & \dots & k(t_1, t_n) \\ \vdots & & \vdots \\ k(t_n, t_1) & \dots & k(t_n, t_n) \end{bmatrix}.$$

If $\boldsymbol{K}^{-1}$ exists then, from linear algebra, there exist functions such that

$$\varphi_i(t_j) = \delta_{ij}, \quad \varphi_i \in \operatorname{span}\{k(\cdot, t_j),\ j = 1, \dots, n\}.$$

Representer Theorem

The regularised estimate $\hat{x}$ is given by $\hat{x} = \sum_{i=1}^{n} c_i \varphi_i = \sum_{i=1}^{n} x(t_i)\varphi_i$.

Approximation in RKHS

In general $|\hat{x}(t) - x(t)| \leq p_{\mathcal{X}}(t_1, \ldots, t_n)\|x\|_{\mathcal{X}}$ where $p_{\mathcal{X}}$ is the power function associated to $\mathcal{X}$. Our aim now is to understand more about $p_{\mathcal{X}}$.

Consider the kernel matrix

$$\boldsymbol{K} = \begin{bmatrix} k(t_1, t_1) & \ldots & k(t_1, t_n) \\ \vdots & & \vdots \\ k(t_n, t_1) & \ldots & k(t_n, t_n) \end{bmatrix}.$$

If $\boldsymbol{K}^{-1}$ exists then, from linear algebra, there exist functions such that

$$\varphi_i(t_j) = \delta_{ij}, \quad \varphi_i \in \mathrm{span}\{k(\cdot, t_j),\ j = 1, \ldots, n\}.$$

Representer Theorem

The regularised estimate $\hat{x}$ is given by $\hat{x} = \sum_{i=1}^{n} c_i \varphi_i = \sum_{i=1}^{n} x(t_i)\varphi_i$.

Derivation of the power function:

$$
\begin{aligned}
|x(t) - \hat{x}(t)| &= \left| x(t) - \sum_{i=1}^{n} x(t_i)\varphi_i(t) \right| \\
&= \left| \left\langle x, k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\rangle_{\mathcal{X}} \right| \quad \text{(reproducing property)} \\
&\leq \|x\|_{\mathcal{X}} \underbrace{\left\| k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\|_{\mathcal{X}}}_{p_{\mathcal{X}}(t_1, \dots, t_n)} \quad \text{(Cauchy-Schwarz)}
\end{aligned}
$$

To study $\hat{x}$ (the posterior mean in a Gaussian process regression) we need to consider the mathematical properties of

$$
p_{\mathcal{X}}(t_1, \dots, t_n) = \left\| k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\|_{\mathcal{X}}.
$$

Approximation in RKHS

Derivation of the power function:

$$\begin{aligned} |x(t) - \hat{x}(t)| &= \left| x(t) - \sum_{i=1}^{n} x(t_i)\varphi_i(t) \right| \\ &= \left| \left\langle x, k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\rangle_{\mathcal{X}} \right| \quad \text{(reproducing property)} \\ &\leq \|x\|_{\mathcal{X}} \underbrace{\left\| k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\|_{\mathcal{X}}}_{p_{\mathcal{X}}(t_1, \ldots, t_n)} \quad \text{(Cauchy-Schwarz)} \end{aligned}$$

To study $\hat{x}$ (the posterior mean in a Gaussian process regression) we need to consider the mathematical properties of

$$p_{\mathcal{X}}(t_1, \ldots, t_n) = \left\| k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\|_{\mathcal{X}}.$$

Derivation of the power function:

$$
\begin{aligned}
|x(t) - \hat{x}(t)| &= \left| x(t) - \sum_{i=1}^{n} x(t_i)\varphi_i(t) \right| \\
&= \left| \left\langle x, k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\rangle_{\mathcal{X}} \right| \quad \text{(reproducing property)} \\
&\leq \|x\|_{\mathcal{X}} \underbrace{\left\| k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\|_{\mathcal{X}}}_{p_{\mathcal{X}}(t_1,\ldots,t_n)} \quad \text{(Cauchy-Schwarz)}
\end{aligned}
$$

To study $\hat{x}$ (the posterior mean in a Gaussian process regression) we need to consider the mathematical properties of

$$
p_{\mathcal{X}}(t_1, \ldots, t_n) = \left\| k(\cdot, t) - \sum_{i=1}^{n} \varphi_i(t) k(\cdot, t_i) \right\|_{\mathcal{X}}.
$$

Derivation of the power function:

$$\begin{aligned}|x(t)-\hat{x}(t)| &= \left|x(t)-\sum_{i=1}^{n}x(t_i)\varphi_i(t)\right| \\ &= \left|\left\langle x, k(\cdot,t)-\sum_{i=1}^{n}\varphi_i(t)k(\cdot,t_i)\right\rangle_{\mathcal{X}}\right| \quad \text{(reproducing property)} \\ &\leq \|x\|_{\mathcal{X}} \underbrace{\left\|k(\cdot,t)-\sum_{i=1}^{n}\varphi_i(t)k(\cdot,t_i)\right\|_{\mathcal{X}}}_{p_{\mathcal{X}}(t_1,\ldots,t_n)} \quad \text{(Cauchy-Schwarz)}\end{aligned}$$

To study $\hat{x}$ (the posterior mean in a Gaussian process regression) we need to consider the mathematical properties of

$$p_{\mathcal{X}}(t_1,\ldots,t_n) = \left\|k(\cdot,t)-\sum_{i=1}^{n}\varphi_i(t)k(\cdot,t_i)\right\|_{\mathcal{X}}.$$

Approximation in RKHS

Equip $D \subset \mathbb{R}^d$ with the Euclidean norm $\|\cdot\|$.

Let $h = \sup_{t \in D} \min_{i=1,\ldots,n} \|t - t_i\|$ denote the fill distance of the points $t_1, \ldots, t_n$ in D.

Then bounds of the form $p_{\mathcal{X}}(t_1, \ldots, t_n) \leq F(h)$ can be obtained (e.g. see Sec. 11.3 of Wendland [2004]):

Kernel $k(t,t')$	Native Space	$F(h)$
$\exp(-\|t-t'\|^2)$	$\cap_{m\in\mathbb{N}} H^m(D)$	$\exp(-c\lvert\log(h)\rvert/h)$
$(c^2+\|t-t'\|^2)^{-\beta}$, $\beta > \frac{d}{2}$	$H^{\beta-\frac{d}{2}}(D)$	$\exp(-c/h)$
$(1-\|t-t'\|)_+^2$	$H^{\frac{d}{2}+\frac{1}{2}}(D)$	$h^{\frac{1}{2}}$
$(1-\|t-t'\|)_+^4(4\|t-t'\|+1)$	$H^{\frac{d}{2}+\frac{3}{2}}(D)$	$h^{\frac{3}{2}}$

... and that's enough theoretical background!

Equip $D \subset \mathbb{R}^d$ with the Euclidean norm $\|\cdot\|$.

Let $h = \sup_{t \in D} \min_{i=1,\dots,n} \|t - t_i\|$ denote the fill distance of the points $t_1, \dots, t_n$ in D.

Then bounds of the form $p_{\mathcal{X}}(t_1, \dots, t_n) \leq F(h)$ can be obtained (e.g. see Sec. 11.3 of Wendland [2004]):

Kernel $k(t, t')$	Native Space	$F(h)$
$\exp(-\|t - t'\|^2)$	$\cap_{m \in \mathbb{N}} H^m(D)$	$\exp(-c\lvert\log(h)\rvert/h)$
$(c^2 + \|t - t'\|^2)^{-\beta}$, $\beta > \frac{d}{2}$	$H^{\beta - \frac{d}{2}}(D)$	$\exp(-c/h)$
$(1 - \|t - t'\|)_+^2$	$H^{\frac{d}{2} + \frac{1}{2}}(D)$	$h^{\frac{1}{2}}$
$(1 - \|t - t'\|)_+^4 (4\|t - t'\| + 1)$	$H^{\frac{d}{2} + \frac{3}{2}}(D)$	$h^{\frac{3}{2}}$

... and that's enough theoretical background!

Approximation in RKHS

Equip $D \subset \mathbb{R}^d$ with the Euclidean norm $\|\cdot\|$.

Let $h = \sup_{t \in D} \min_{i=1,\dots,n} \|t - t_i\|$ denote the fill distance of the points $t_1, \dots, t_n$ in D.

Then bounds of the form $p_{\mathcal{X}}(t_1, \dots, t_n) \leq F(h)$ can be obtained (e.g. see Sec. 11.3 of Wendland [2004]):

Kernel $k(t, t')$	Native Space	$F(h)$
$\exp(-\|t - t'\|^2)$	$\cap_{m \in \mathbb{N}} H^m(D)$	$\exp(-c\|\log(h)\|/h)$
$(c^2 + \|t - t'\|^2)^{-\beta}$, $\beta > \frac{d}{2}$	$H^{\beta - \frac{d}{2}}(D)$	$\exp(-c/h)$
$(1 - \|t - t'\|)_+^2$	$H^{\frac{d}{2} + \frac{1}{2}}(D)$	$h^{\frac{1}{2}}$
$(1 - \|t - t'\|)_+^4 (4\|t - t'\| + 1)$	$H^{\frac{d}{2} + \frac{3}{2}}(D)$	$h^{\frac{3}{2}}$

... and that's enough theoretical background!

Seventh Job: Solution of Integrals, in Detail

Solution of Integrals

Consider estimation of the Quantity of Interest

$$Q(x) = \int x(t)\mathrm{d}\nu(t)$$

where x is an integrand of interest and ν is a measure on $D \subseteq \mathbb{R}^d$.

In the Bayesian approach to Probabilistic Numerics, we must select an information operator

$$A(x) = \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_n) \end{bmatrix}.$$

I.e. we must select points $\{t_i\}_{i=1}^n$ at which to evaluate the integrand. But how?

Consider estimation of the Quantity of Interest

$$Q(x) = \int x(t) \mathrm{d}\nu(t)$$

where x is an integrand of interest and ν is a measure on $D \subseteq \mathbb{R}^d$.

In the Bayesian approach to Probabilistic Numerics, we must select an information operator

$$A(x) = \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_n) \end{bmatrix}.$$

I.e. we must select points $\{t_i\}_{i=1}^n$ at which to evaluate the integrand. But how?

Consider estimation of the Quantity of Interest

$$Q(x) = \int x(t)\mathrm{d}\nu(t)$$

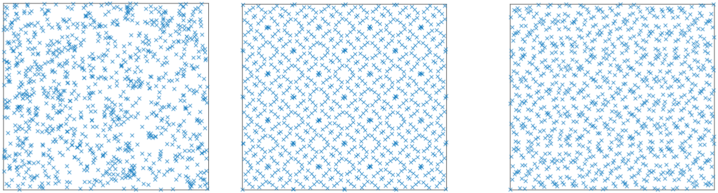
where x is an integrand of interest and ν is a measure on $D \subseteq \mathbb{R}^d$.

In the Bayesian approach to Probabilistic Numerics, we must select an information operator

$$A(x) = \begin{bmatrix} x(t_1) \\ \vdots \\ x(t_n) \end{bmatrix}.$$

I.e. we must select points $\{t_i\}_{i=1}^n$ at which to evaluate the integrand. But how?

Monte Carlo & Quasi-Monte Carlo Points



Monte Carlo | Sobol Sequence | Higher-Order Digital Net

	Monte Carlo	Sobol Sequence	Higher-Order Digital Net
$\mathcal{F} =$	$L^2(D)$	$H^1(D)$	$H^\beta_{\text{mix}} := H^\beta_1(D) \times \cdot \times H^\beta_1(D)$
$e_{\text{WCE}}(M) =$	$O_P(n^{-1/2})$	$O(n^{-1})$	$O(n^{-\beta})$

Here we show worst case error $e_{\text{WCE}}(M)$ for the method $M = (A, b)$ where $b(a) = \frac{1}{n} \sum_{i=1}^n a_i$. i.e. an un-weighted average of function evaluations at the points $\{t_i\}_{i=1}^n$.

Bayesian Quadrature

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_\# P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) = \int \hat{x}(t) \mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_\infty \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \dots, t_n)\|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_\# P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) = \int \hat{x}(t) \mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_\infty \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \ldots, t_n) \|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_{\#} P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) = \int \hat{x}(t) \mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_{\infty} \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \dots, t_n) \|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_\# P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) = \int \hat{x}(t) \mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_\infty \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \dots, t_n) \|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_\# P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) \;=\; \int \hat{x}(t)\mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_\infty \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \ldots, t_n)\|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Bayesian Quadrature

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_\# P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) = \int \hat{x}(t) \mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_\infty \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \dots, t_n) \|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Bayesian Quadrature

Bayesian Quadrature is a Bayesian probabilistic numerical method based on a Gaussian prior $P_x : x \sim \mathcal{GP}(0, k)$.

The mean of the posterior $Q_{\#} P_{x|a}$ is denoted $b(a)$. It satisfies

$$b(a) = \int \hat{x}(t) \mathrm{d}\nu(t)$$

where $\hat{x}$ is the RKHS interpolant based on the information $A(x) = a$.

The performance of the posterior mean b, viewed as a classical numerical method, can be studied with our established results on RKHS interpolants:

Suppose D is a bounded subset of $\mathcal{X}$. Then:

$$\begin{aligned} |b(A(x)) - Q(x)| &\leq \|\hat{x} - x\|_{L^2(\nu)} \quad \text{(regression bound)} \\ &\leq \|\hat{x} - x\|_{\infty} \quad \text{(sup bound)} \\ &\leq p_{\mathcal{X}}(t_1, \dots, t_n) \|x\|_{\mathcal{X}} \quad \text{(RKHS fill-distance bound)} \end{aligned}$$

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \ldots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\text{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \ldots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\text{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \ldots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\text{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \ldots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\text{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \ldots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\mathrm{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \ldots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\mathrm{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

Thus, the analysis of Bayesian Quadrature can be reduced to analysis of how the power function $p_{\mathcal{X}}(t_1, \dots, t_n)$ depends on the choice of the points $\{t_i\}_{i=1}^n$.

Let $\mathcal{X} = H^1(D)$ be the standard Sobolev space, with an appropriate choice of kernel k. Let $\mathcal{X} = [0,1]^d$, ν be uniform on D and let the points $\{t_i\}_{i=1}^n$ be quasi-uniform over D (i.e. $h = O(\frac{1}{n})$). Then $\exists C$ s.t. whenever $\alpha > \frac{d}{2}$:

$$e_{\text{WCE}}(M) \quad = \quad O(n^{-1/d})$$

for all $\epsilon > 0$.

- Recall that $\hat{b}$ is the trapezoidal rule - so this matches known results.
- Optimal rate for the WCE of a deterministic method for integration of functions in the space $H^1(D)$.
- The method of proof can be extended to other domains/measures/point sets.

The mean $b(a)$ of the posterior $Q_{\#}P_{x|a}$ can be considered as a classical numerical method and we can ask about *optimal information* for b, either in the sense of worst-case or average-case optimal. But why would this be relevant?

The variance of the posterior $Q_{\#}P_{x|a}$ is equal to $e_{\mathrm{WCE}}(M)^2$ where $M = (A, b)$.

(This is a special case of the fact from Bayesian decision theory that (for equaliser rules) minimax $\leftrightarrow$ Bayes.)

For the $\mathcal{X} = H^1(D)$ example, with $D = [0, 1]$ the kernel $k(t, t') = \min(t, t')$, we will prove later that optimal information (i.e. the points $\{t_i\}_{i=1}^n$ that minimise the posterior variance) are a uniform grid over $[0, 1]$.

The mean $b(a)$ of the posterior $Q_{\#}P_{x|a}$ can be considered as a classical numerical method and we can ask about *optimal information* for b, either in the sense of worst-case or average-case optimal. But why would this be relevant?

The variance of the posterior $Q_{\#}P_{x|a}$ is equal to $e_{\mathrm{WCE}}(M)^2$ where $M = (A, b)$.

(This is a special case of the fact from Bayesian decision theory that (for equaliser rules) minimax $\leftrightarrow$ Bayes.)

For the $\mathcal{X} = H^1(D)$ example, with $D = [0, 1]$ the kernel $k(t, t') = \min(t, t')$, we will prove later that optimal information (i.e. the points $\{t_i\}_{i=1}^n$ that minimise the posterior variance) are a uniform grid over $[0, 1]$.

The mean $b(a)$ of the posterior $Q_{\#}P_{x|a}$ can be considered as a classical numerical method and we can ask about *optimal information* for b, either in the sense of worst-case or average-case optimal. But why would this be relevant?

The variance of the posterior $Q_{\#}P_{x|a}$ is equal to $e_{\text{WCE}}(M)^2$ where $M = (A, b)$.

(This is a special case of the fact from Bayesian decision theory that (for equaliser rules) minimax $\leftrightarrow$ Bayes.)

For the $\mathcal{X} = H^1(D)$ example, with $D = [0, 1]$ the kernel $k(t, t') = \min(t, t')$, we will prove later that optimal information (i.e. the points $\{t_i\}_{i=1}^n$ that minimise the posterior variance) are a uniform grid over $[0, 1]$.

The mean $b(a)$ of the posterior $Q_{\#}P_{x|a}$ can be considered as a classical numerical method and we can ask about *optimal information* for b, either in the sense of worst-case or average-case optimal. But why would this be relevant?

The variance of the posterior $Q_{\#}P_{x|a}$ is equal to $e_{\text{WCE}}(M)^2$ where $M = (A, b)$.

(This is a special case of the fact from Bayesian decision theory that (for equaliser rules) minimax $\leftrightarrow$ Bayes.)

For the $\mathcal{X} = H^1(D)$ example, with $D = [0, 1]$ the kernel $k(t, t') = \min(t, t')$, we will prove later that optimal information (i.e. the points $\{t_i\}_{i=1}^n$ that minimise the posterior variance) are a uniform grid over $[0, 1]$.

The mean $b(a)$ of the posterior $Q_{\#}P_{x|a}$ can be considered as a classical numerical method and we can ask about *optimal information* for b, either in the sense of worst-case or average-case optimal. But why would this be relevant?

The variance of the posterior $Q_{\#}P_{x|a}$ is equal to $e_{\text{WCE}}(M)^2$ where $M = (A, b)$.

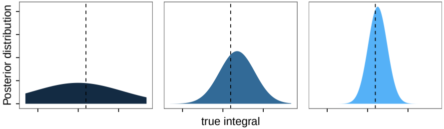
(This is a special case of the fact from Bayesian decision theory that (for equaliser rules) minimax $\leftrightarrow$ Bayes.)

For the $\mathcal{X} = H^1(D)$ example, with $D = [0, 1]$ the kernel $k(t, t') = \min(t, t')$, we will prove later that optimal information (i.e. the points $\{t_i\}_{i=1}^n$ that minimise the posterior variance) are a uniform grid over $[0, 1]$.

Posterior Contraction

Of course, we are not interested in just the mean of $Q_{\#}P_{x|a}$ but the full distribution $Q_{\#}P_{x|a}$ itself.

A basic question is "does this probability mass contract to the true value $Q(x)$?"



For Bayesian Quadrature, where P_x is Gaussian, this can be answered through the properties of Gaussians:

For Bayesian Quadrature, if the true integrand satisfies $\|x\|_{\mathcal{X}} < \infty$, then for all $\epsilon > 0$ there exists C_ϵ such that:

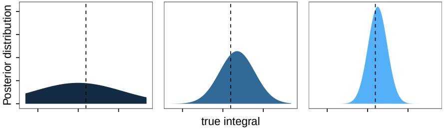
$$Q_{\#}P_{x|a}(I_{\text{true}} - \epsilon, I_{\text{true}} + \epsilon) = 1 - o\left(\exp(-C_\epsilon / e_{\text{WCE}}(M)^2)\right)$$

where I_{true} is the true value of the integral and $M = (A, B)$, $B = Q_{\#}P_{x|a}$.

Posterior Contraction

Of course, we are not interested in just the mean of $Q_{\#}P_{x|a}$ but the full distribution $Q_{\#}P_{x|a}$ itself.

A basic question is "does this probability mass contract to the true value $Q(x)$?"



For Bayesian Quadrature, where P_x is Gaussian, this can be answered through the properties of Gaussians:

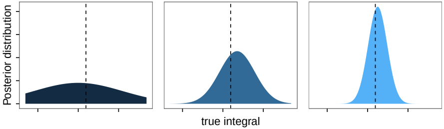
For Bayesian Quadrature, if the true integrand satisfies $\|x\|_{\mathcal{X}} < \infty$, then for all $\epsilon > 0$ there exists C_ϵ such that:

$$Q_{\#}P_{x|a}(I_{\text{true}} - \epsilon, I_{\text{true}} + \epsilon) = 1 - o\big(\exp(-C_\epsilon / e_{\text{WCE}}(M)^2)\big)$$

where I_{true} is the true value of the integral and $M = (A, B)$, $B = Q_{\#}P_{x|a}$.

Of course, we are not interested in just the mean of $Q_{\#}P_{x|a}$ but the full distribution $Q_{\#}P_{x|a}$ itself.

A basic question is "does this probability mass contract to the true value $Q(x)$?"



For Bayesian Quadrature, where P_x is Gaussian, this can be answered through the properties of Gaussians:

For Bayesian Quadrature, if the true integrand satisfies $\|x\|_{\mathcal{X}} < \infty$, then for all $\epsilon > 0$ there exists C_ϵ such that:

$$Q_{\#}P_{x|a}(I_{\text{true}} - \epsilon, I_{\text{true}} + \epsilon) = 1 - o\left(\exp(-C_\epsilon / e_{\text{WCE}}(M)^2)\right)$$

where I_{true} is the true value of the integral and $M = (A, B)$, $B = Q_{\#}P_{x|a}$.

Calibration

THINGS GOT REALLY INTERESTING WHEN THE STATISTICIAN STARTED DOING WARD ROUNDS.

Given a specific kernel, e.g. Matérn kernel below:

$$k_\alpha(t,t';\sigma,\lambda) \;:=\; \lambda^2 \prod_{i=1}^{d} \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)^{\alpha} K_\alpha \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)$$

we need to specify hyper-parameters (λ, σ).

These hyper-parameters can greatly influence the posterior mean and variance. From a Bayesian perspective, these need to be set adequately to obtain good quantification of uncertainty.

In this Part, we consider *empirical Bayes*, which entails maximising the marginal likelihood of the data over the hyper-parameters:

$$\text{argmax}_{\sigma,\lambda}\; p\left(\{x(t_i)\}_{i=1}^{n} \middle| \sigma, \lambda, \{t_i\}_{i=1}^{n}\right)$$

Theoretically difficult to estimate α - see counterexamples in Szabó et al. [2015].

Given a specific kernel, e.g. Matérn kernel below:

$$k_\alpha(t, t'; \sigma, \lambda) \;:=\; \lambda^2 \prod_{i=1}^{d} \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)^{\alpha} K_\alpha \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)$$

we need to specify hyper-parameters (λ, σ).

These hyper-parameters can greatly influence the posterior mean and variance. From a Bayesian perspective, these need to be set adequately to obtain good quantification of uncertainty.

In this Part, we consider *empirical Bayes*, which entails maximising the marginal likelihood of the data over the hyper-parameters:

$$\text{argmax}_{\sigma,\lambda}\; p\left(\{x(t_i)\}_{i=1}^n \middle| \sigma, \lambda, \{t_i\}_{i=1}^n\right)$$

Theoretically difficult to estimate α - see counterexamples in Szabó et al. [2015].

Given a specific kernel, e.g. Matérn kernel below:

$$k_\alpha(t, t'; \sigma, \lambda) \; := \; \lambda^2 \prod_{i=1}^{d} \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)^\alpha K_\alpha \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)$$

we need to specify hyper-parameters (λ, σ).

These hyper-parameters can greatly influence the posterior mean and variance. From a Bayesian perspective, these need to be set adequately to obtain good quantification of uncertainty.

In this Part, we consider *empirical Bayes*, which entails maximising the marginal likelihood of the data over the hyper-parameters:

$$\text{argmax}_{\sigma,\lambda} \; p\big(\{x(t_i)\}_{i=1}^n \big| \sigma, \lambda, \{t_i\}_{i=1}^n\big)$$

Theoretically difficult to estimate α - see counterexamples in Szabó et al. [2015].

Given a specific kernel, e.g. Matérn kernel below:

$$k_\alpha(t, t'; \sigma, \lambda) \ := \ \lambda^2 \prod_{i=1}^{d} \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)^{\alpha} K_\alpha \left(\frac{\sqrt{2\alpha}|t_i - t_i'|}{\sigma} \right)$$

we need to specify hyper-parameters (λ, σ).

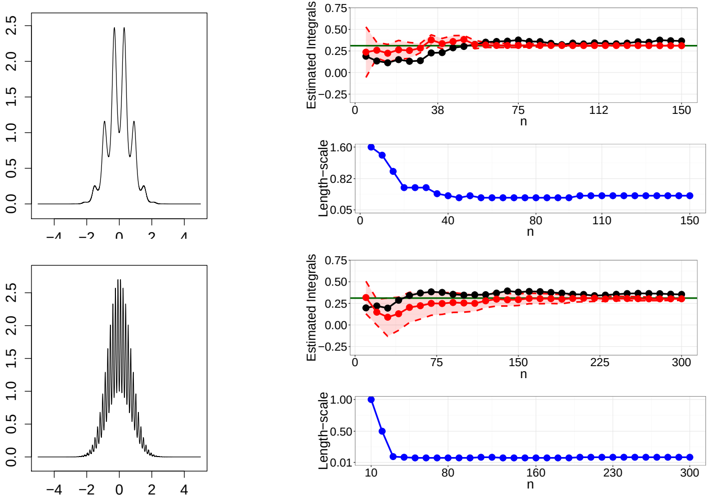
These hyper-parameters can greatly influence the posterior mean and variance. From a Bayesian perspective, these need to be set adequately to obtain good quantification of uncertainty.

In this Part, we consider *empirical Bayes*, which entails maximising the marginal likelihood of the data over the hyper-parameters:

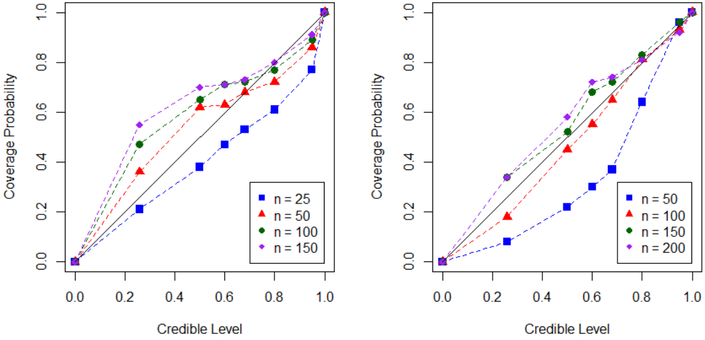
$$\text{argmax}_{\sigma,\lambda} \ p\left(\{x(t_i)\}_{i=1}^n \middle| \sigma, \lambda, \{t_i\}_{i=1}^n\right)$$

Theoretically difficult to estimate α - see counterexamples in Szabó et al. [2015].

Calibration on Test Functions



Calibration on Test functions



- Empirical Bayes can be over-confident when n is small.
- Alternative option would be marginalisation - but requires that a hyper-prior be specified.

Conclusion

In Part III it has been argued that:

- For Gaussian priors P_x, the theory of approximation in RKHS is important.
- For Bayesian Quadrature, the analysis of the full posterior $Q_{\#} P_{x|a}$ reduced to analysis of the posterior mean $b(a)$ and was classical.
- Calibration of uncertainty remains an important open problem.

END OF PART III

Conclusion

In Part III it has been argued that:

- For Gaussian priors P_x, the theory of approximation in RKHS is important.
- For Bayesian Quadrature, the analysis of the full posterior $Q_{\#}P_{x|a}$ reduced to analysis of the posterior mean $b(a)$ and was classical.
- Calibration of uncertainty remains an important open problem.

END OF PART III

In Part III it has been argued that:

- For Gaussian priors P_x, the theory of approximation in RKHS is important.
- For Bayesian Quadrature, the analysis of the full posterior $Q_{\#}P_{x|a}$ reduced to analysis of the posterior mean $b(a)$ and was classical.
- Calibration of uncertainty remains an important open problem.

END OF PART III

In Part III it has been argued that:

- For Gaussian priors P_x, the theory of approximation in RKHS is important.
- For Bayesian Quadrature, the analysis of the full posterior $Q_\# P_{x|a}$ reduced to analysis of the posterior mean $b(a)$ and was classical.
- Calibration of uncertainty remains an important open problem.

END OF PART III

In Part III it has been argued that:

- For Gaussian priors P_x, the theory of approximation in RKHS is important.
- For Bayesian Quadrature, the analysis of the full posterior $Q_\# P_{x|a}$ reduced to analysis of the posterior mean $b(a)$ and was classical.
- Calibration of uncertainty remains an important open problem.

END OF PART III