A Bayes–Sard Cubature Method

arXiv:1804.03016

Toni Karvonen[1], Chris Oates[2] and Simo Särkkä[1]

[1] *Aalto University, Finland*
[2] *Newcastle University and Alan Turing Institute, UK*

SAMSI-Lloyds-Turing Workshop on Probabilistic Numerical Methods
Alan Turing Institute, London

April 12, 2018

Motivation

Cubature: Compute approximations to the integral of a deterministic function $f^\dagger \colon D \to \mathbb{R}$, $D \subset \mathbb{R}^d$ by the means of a cubature rule:

$$\sum_{i=1}^{n} w_i f^\dagger(x_i) \approx I(f^\dagger) := \int_D f^\dagger \,\mathrm{d}\nu.$$

Bayesian cubature: Model $f^\dagger$ with a stochastic process consistent with the obtained evaluations $f^\dagger(x_i)$ at x_i and integrate this stochastic process.

Gaussian processes: The "standard" approach is to use a Gaussian process $f \sim \mathcal{GP}(0, k)$ with a covariance kernel k and condition on the data $\mathcal{D}_X := \{(x_i, f^\dagger(x_i)\}_{i=1}^n$:

$$\text{“}I(f) \mid \mathcal{D}_X \approx I(f^\dagger)\text{”}.$$

Origins

Oates (2017). Posterior Integration on an Embedded Riemannian Manifold. *arXiv:1712.01793v1*:

Bayesian cubature using the sum kernel

$$k_\sigma(x, x') = k(x, x') + \sigma^2$$

yields, at the "weakly informative prior limit" $\sigma \to \infty$, a valid Bayesian cubature rule whose weights sum to one.

What about using $k_\sigma(x, x') = k(x, x') + \sigma^2 \sum_{i=1}^{Q} p_i(x) p_i(x')$*?*

Contributions

We develop the ***Bayes–Sard cubature***, a generalisation of the conventional Bayesian cubature, by augmenting the GP with a parametric prior mean whose features span a finite-dimensional function space π and taking an improper limit.

If $\dim(\pi) < n$, the BSC is exact for every function in π. This makes it more stable than BC and less sensitive to, e.g., misspecification of the kernel length-scale.

If $\dim(\pi) = n$, a judicious choice of the space π allows for endowing *any* cubature rule with a meaningful probabilistic output. The variance coincides with the worst-case error in the RKHS of k.

Table of contents

Gaussian process prior

Prior

We assign f a Gaussian process prior with a parametric mean:

$$f(x) \mid \gamma \sim \mathcal{GP}\big(s(x), k(x, x')\big),$$

$$s(x) \mid \gamma = \sum_{j=1}^{Q} \gamma_j p_j(x),$$

$$\gamma \sim \mathcal{N}(0, \Sigma),$$

where the functions $p_1, \ldots, p_Q$ form a basis of a Q-dimensional ($Q \leq n$) function space π.

We are going to do regression at the *flat prior limit* $\Sigma^{-1} \to 0$; see O'Hagan (1978) and Rasmussen & Williams (2006), Section 2.7.

Flat prior limit

Given the data $\mathcal{D}_X = (X, f_X^\dagger)$, the flat prior $\Sigma^{-1} \to 0$ GP posterior is

$$f(x) \mid \mathcal{D}_X \sim \mathcal{GP}\big(s_X(f^\dagger)(x), k_X(x, x')\big)$$

with the mean and covariance functions

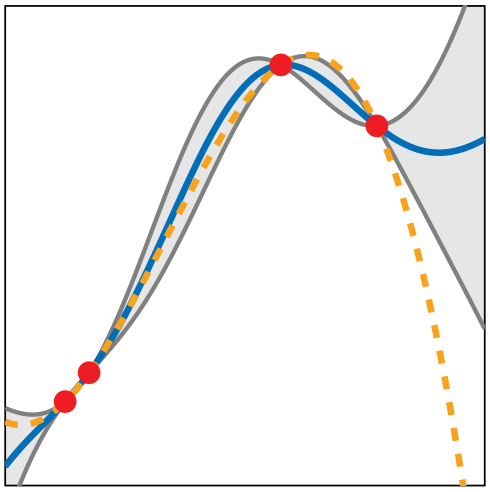
$$\begin{aligned} s_X(f^\dagger)(x) &= \alpha^\top k_X(x) + \beta^\top p_X(x), \\ k_X(x, x') &= k(x, x') - k_X(x)^\top K_X^{-1} k_X(x') + C_{k,\pi}(x, x'). \end{aligned}$$

The coefficient vectors $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^Q$ solve

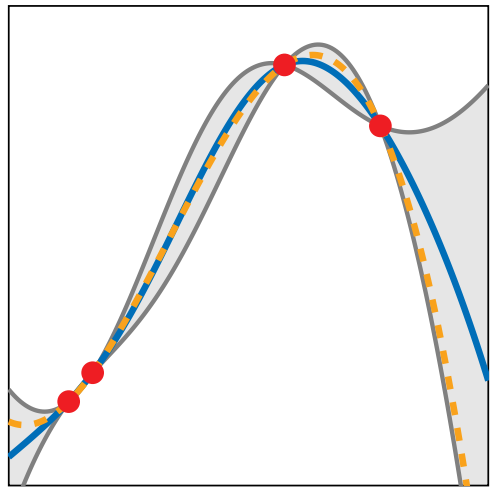
$$\begin{bmatrix} K_X & P_X \\ P_X^\top & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} f_X^\dagger \\ 0 \end{bmatrix},$$

where $[P_X]_{ij} = p_j(x_i)$ is the $n \times Q$ *Vandermonde matrix*.

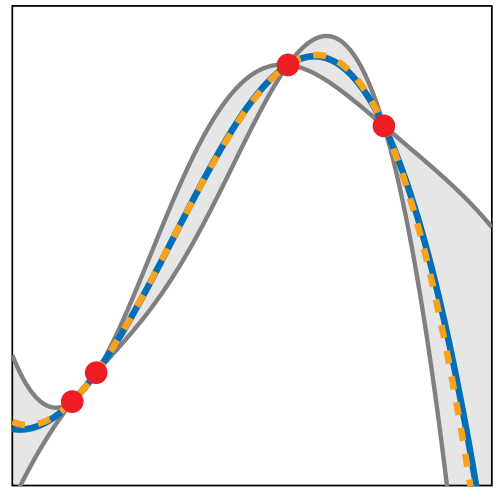
$\Sigma^{-1} = 10\,\mathrm{I}$



$\Sigma^{-1} = \mathrm{I}$



$\Sigma^{-1} = 0.1\,\mathrm{I}$



Some properties

Unisolvency: The point set $X \subset D$ needs to be π-*unisolvent* to guarantee that $P_X \in \mathbb{R}^{n \times Q}$ is of full rank. Not a problem usually.

Connection to interpolation: When π is a polynomial space, the posterior mean $s_X(f^\dagger)$ coincides with an interpolant for a conditionally positive-definite kernel defined in Wendland (2005), Section 8.5.

Reproduction of elements in π: If $f^\dagger \in \pi$, then $s_X(f^\dagger) = f^\dagger$. In particular, if $Q = \dim(\pi) = n$, the posterior mean $s_X(f^\dagger)$ is the *unique interpolant from* π to the data $\mathcal{D}_X$:

$$s_X(f^\dagger)(x) = p(x)^\top P_X^{-\top} f_X^\dagger.$$

Table of contents

Bayes–Sard cubature

At the limit $\Sigma^{-1} \to 0$, the integral $I(f) = \int_D f \mathrm{d}\nu$ has the Gaussian posterior distribution

$$I(f) \mid \mathcal{D}_X \sim \mathcal{N}\big(\mu_X(f^\dagger), \sigma_X^2\big)$$

with the mean and variance

$$\mu_X(f^\dagger) = \int_D s_X(f^\dagger)(x)\mathrm{d}\nu(x),$$
$$\sigma_X^2 = \int_D \int_D k_X(x, x')\mathrm{d}\nu(x)\mathrm{d}\nu(x').$$

The mean is used to approximate $I(f^\dagger)$ and σ_X^2 for quantification of epistemic uncertainty. We call this approximation the ***Bayes–Sard cubature***.

Bayes–Sard cubature: properties

Mean: The mean indeed takes the form of a cubature rule:

$$\mu_X(f^\dagger) = I\big(s_X(f^\dagger)\big) = \sum_{i=1}^{n} w_{k,i} f^\dagger(x_i)$$

for weights $w_k \in \mathbb{R}^n$ obtained from the solution of

$$\begin{bmatrix} K_X & P_X \\ P_X^\top & 0 \end{bmatrix} \begin{bmatrix} w_k \\ w_\pi \end{bmatrix} = \begin{bmatrix} k_{\nu,X} \\ p_\nu \end{bmatrix}.$$

Variance: The Bayes–Sard variance σ^2_X is *non-zero*.

A kernel perspective: Bayesian cubature

By a standard equivalence, the BC weights $w^{\mathsf{BC}} \in \mathbb{R}^n$ for the kernel k (i.e., $f \sim \mathcal{GP}(0, k)$) are *worst-case optimal* in the RKHS $H(k)$ of k:

$$w^{\mathsf{BC}} = \operatorname*{arg\,min}_{w \in \mathbb{R}^n} e_{H(k)}(X, w)$$

for the WCE

$$e_{H(k)}(X, w) := \sup_{\|h\|_{H(k)} \leq 1} \left| \int_D h \, \mathrm{d}\nu - \sum_{i=1}^{n} w_i h(x_i) \right|.$$

The associated variance is precisely the squared WCE:

$$\mathrm{Var}\left[I(f) \mid \mathcal{D}_X \right] = e_{H(k)}(X, w^{\mathsf{BC}})^2.$$

A kernel perspective: Bayes–Sard cubature

Our prior model yields the marginal

$$f(x) \sim \mathcal{GP}\big(0, k(x,x') + p(x)^\top \Sigma p(x')\big)$$

that essentially corresponds to using the kernel

$$k_\sigma(x,x') := k(x,x') + \sigma^2 k_\pi(x,x'),$$

with $k_\pi(x,x') = \sum_{i=1}^{Q} p_i(x)p_i(x')$ in BC.

Consequently,

$$w_\sigma^{\mathsf{BC}} \to w_k \qquad \text{as} \qquad \sigma \to \infty.$$

A kernel perspective: implications

The RKHS of k_σ is $H(k_\sigma) = \{g + p : g \in H(k), p \in \pi\}$ equipped with the norm

$$\|h\|^2_{H(k_\sigma)} = \min_{g \in H(k),\, p \in \pi} \{\|g\|^2_{H(k)} + \sigma^2 \|p\|^2_{H(k_\pi)} : g + p = h\}.$$

The weights w_σ^{BC} are selected to minimise the WCE in $H(k_\sigma)$ and $H(k_\sigma)$ is dominated by the part from π. At the limit $\sigma \to \infty$:

- "$\dim(\pi)$ weights are spent to be exact on π and the rest for integrating functions from $H(k)$ well".
- If $\dim(\pi) = n$, "all weights are spent on π and nothing is done to integrate functions from $H(k)$ well".

If $\dim(\pi) = n$, the Bayes–Sard weights are independent of k and the BSC variance σ^2_X coincides with the worst-case error $e_{H(k)}(X, w_k)^2$.

Why Sard?

The Bayes–Sard cubature is reminiscent of the method of Sard (1949)[1] finding out the "best" quadrature formula for given n nodes:

1. Select $m \leq n$ and require that the rule is exact whenever the integrand is a polynomial of degree at most $m - 1$.
2. Dispose of the remaining $n - m$ degrees of freedom by minimising a suitable error functional.

Indeed, the Bayes–Sard weights w_k satisfy

$$w_k = \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \|k_\nu - w^\top k_X\|_{H(k)} \quad \text{subject to} \quad P_X^\top w = p_\nu.$$

[1] Sard, A. (1949). Best approximate integration formulas; best approximation formulas. *American Journal of Mathematics*, 71(1):80–91.

All cubature rules have probabilistic counterparts

If $\dim(\pi) = n$, the BSC weights are the solution to

$$\begin{bmatrix} p_1(x_1) & \cdots & p_1(x_n) \\ \vdots & \ddots & \vdots \\ p_n(x_1) & \cdots & p_n(x_n) \end{bmatrix} \begin{bmatrix} w_{k,1} \\ \vdots \\ w_{k,n} \end{bmatrix} = \begin{bmatrix} I(p_1) \\ \vdots \\ I(p_n) \end{bmatrix}.$$

Selecting, for example, a partition $D = \cup_{i=1}^{n} D_i$ for disjoint D_i such that $x_i \in D_i$ and $\nu(D_i) = 1/n$ yields $w_{k,i} = 1/n$. *We get (quasi) Monte Carlo.*

Theorem

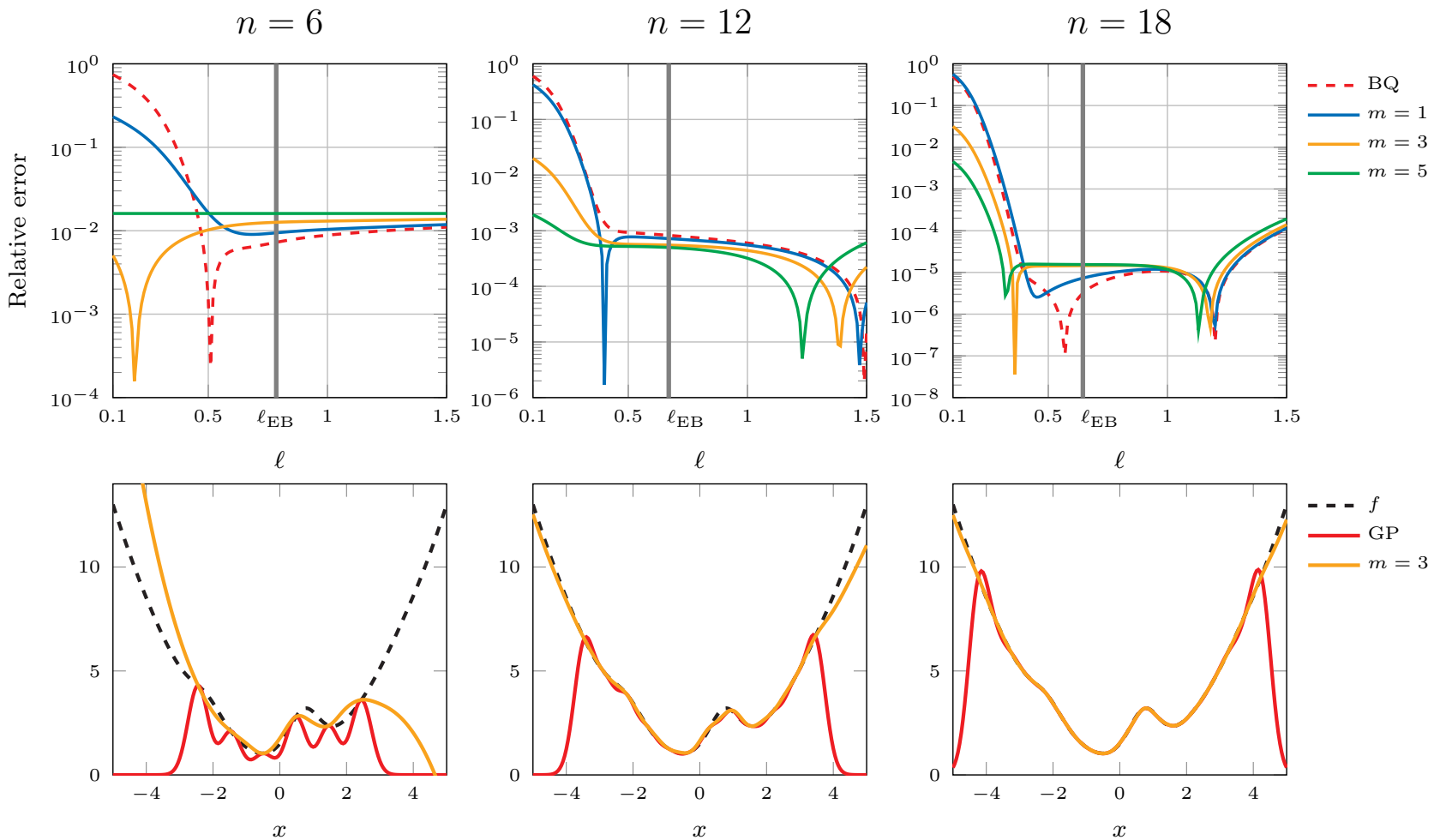
Consider a cubature rule with point set X of size n and non-zero weights $w \in \mathbb{R}^n$. Then there exists a function space π of dimension n, such that the Bayes–Sard method recovers $w_k = w$ with $\sigma_X^2 = e_k(X, w)^2$.

Why (maybe) use Bayes–Sard?

- The posterior mean $\mu_X(f^\dagger)$, used to approximate $I(f^\dagger)$, is less dependent on the kernel and its parameters.
- Potentially easier to select the kernel and its hyperparameters.
- Linear constraints, such as $\sum_{i=1}^n w_{k,i} = 1$, can be easily encoded.
- If you desire to use a cubature rule of your choice, this can be endowed with a meaningful probabilistic output.

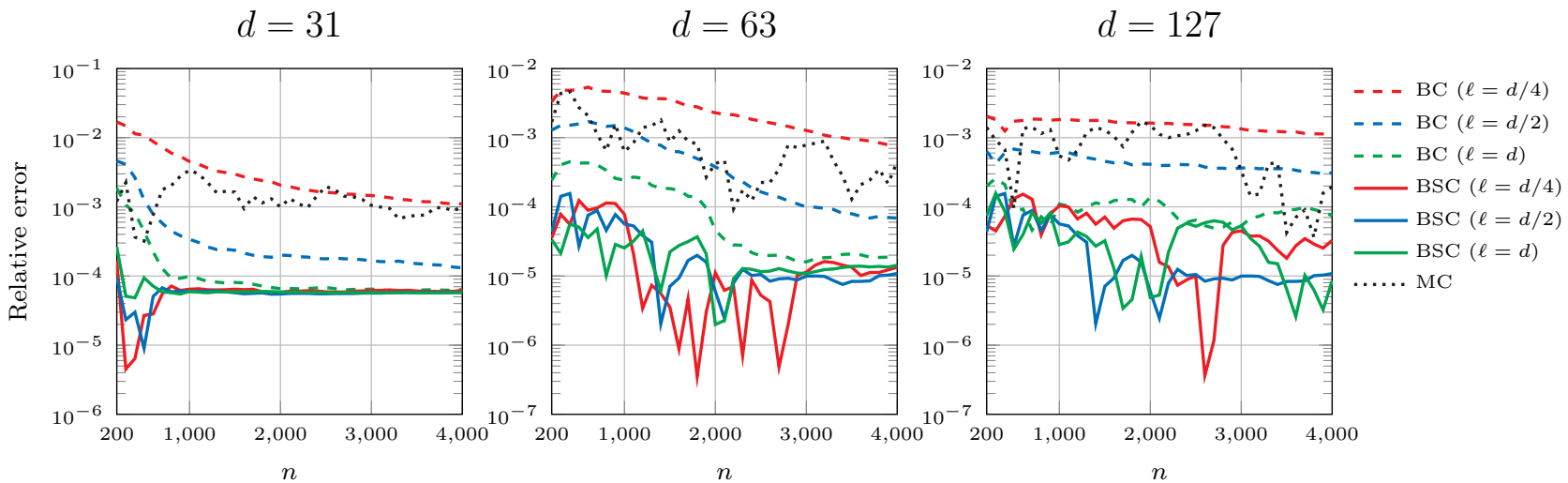
Table of contents

Sensitivity to length-scale: $d = 1$



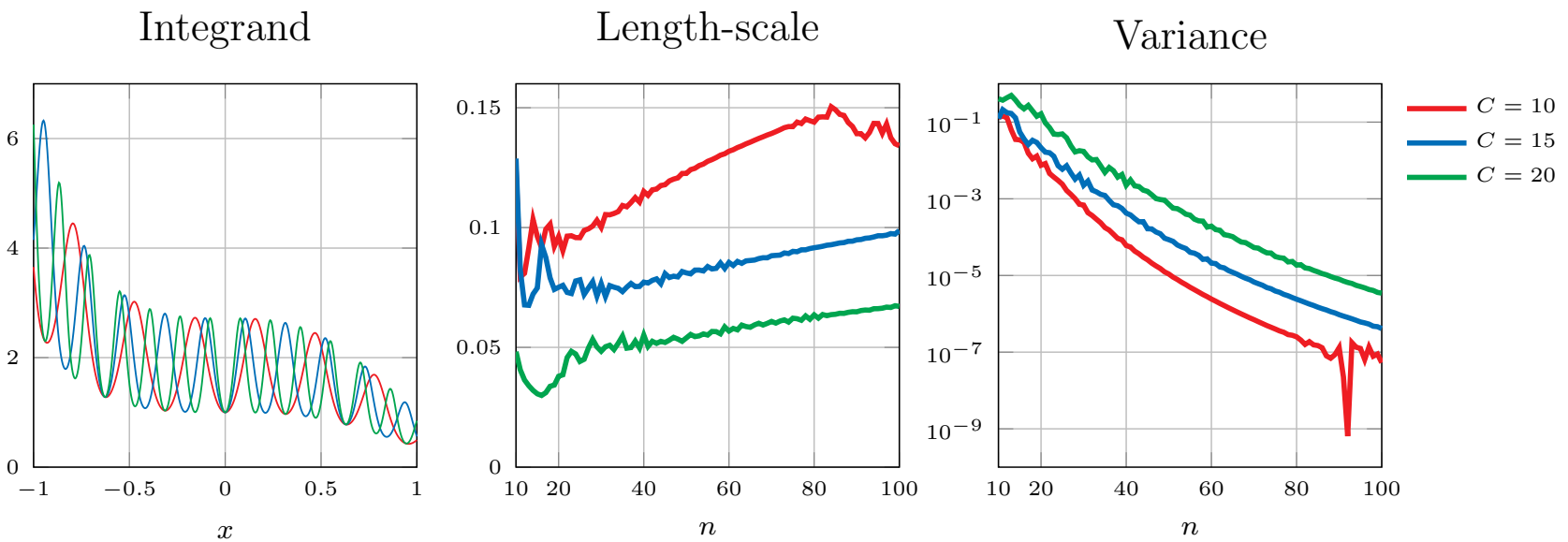
$$f^{\dagger}(x) = \exp\left(\sin(2x) - \frac{x^2}{5}\right) + \frac{x^2}{2}$$

Sensitivity to length-scale: $d > 1$



A zero coupon bond integrand arising from $d + 1$ step Euler–Maruyama discretisation.

UQ for Clenshaw–Curtis might be meaningful



$$f^{\dagger}(x) = \exp\left(\sin(Cx)^2 - x^3\right)$$

Concluding remarks

- The integral approximations appear in a non-probabilistic context already in Bezhaev (1991).
- Convergence rates not affected if π is a polynomial space with fixed dimension.
- Use a fast approximate GP method to fit the kernel parameters. Little effect on the integral estimate, but is UQ going to be meaningful?
- Compare UQ for QMC to that obtained using Fred's method?