Numerical integration as a statistical inference problem

Toni Karvonen

Dept. of Electrical Engineering and Automation, Aalto University

The 38th Finnish Summer School on Probability and Statistics
Lammi, Finland

May 29, 2018
Table of contents

Bayesian cubature

Uncertainty quantification

Computational approaches
Numerical integration

In *numerical integration*, the aim is to approximate the intractable integral of a *deterministic* $f : \Omega \to \mathbb{R}$ using a *cubature rule*:

$$Q(f) = \sum_{i=1}^{n} w_i f(x_i) \approx \int_{\Omega} f \, d\mu,$$

where $w_i \in \mathbb{R}$ are the *weights* and $x_i \in \Omega$ the *nodes* or *points*.

Examples:
- *Gaussian cubature*: approximate $f$ with a polynomial and integrate the polynomial.
- *(Quasi)-Monte Carlo*: select nodes that fill the space well and set $w_i = 1/n$.

We take a *probabilistic approach*: place a prior on $f$ and condition the process on the “data” $\mathcal{D} = \{(x_i, f(x_i))\}_{i=1}^{n}$. 
Bayesian cubature

1. Place a Gaussian process prior with a positive-definite covariance kernel $k : \Omega \times \Omega \to \mathbb{R}$ on the integrand:

$$
\begin{bmatrix}
  f(x_1) \\
  \vdots \\
  f(x_n)
\end{bmatrix}
\sim \mathcal{N}
\left(0,
\begin{bmatrix}
  k(x_1, x_1) & \cdots & k(x_1, x_n) \\
  \vdots & \ddots & \vdots \\
  k(x_n, 1) & \cdots & k(x_n, x_n)
\end{bmatrix}
\right).
$$

2. Condition on the function evaluations $f_X = [f(x_1), \ldots, f(x_n)]$ at $X = \{x_1, \ldots, x_n\}$:

$$
f(x) \mid \mathcal{D} \sim \mathcal{GP}(s_{f,X}(x), k_X(x, x')).
$$

3. Integrate the posterior:

$$
\int_{\Omega} (f \mid \mathcal{D}) \, d\mu \sim \mathcal{N}(Q_X^{BC}(f), \sigma_X^2).
$$
Bayesian cubature: equations

The integral posterior $\int_{\Omega} (f \mid D) \, d\mu \sim \mathcal{N}(Q_X^{BC}(f), \sigma_X^2)$ has

$$Q_X^{BC}(f) = f_X^T K_X^{-1} k_{\mu,X} \quad \text{and} \quad \sigma_X^2 = \int_{\Omega} k_{\mu} \, d\mu - Q_X^{BC}(k_{\mu}),$$

where

$$[K_X]_{ij} = k(x_i, x_j), \quad k_{\mu}(x) = \int_{\Omega} k(x, x') \, d\mu(x'), \quad [k_{\mu,X}]_i = k_{\mu}(x_i).$$

The posterior mean $Q_X^{BC}(f)$ takes the form of a cubature rule $Q_X^{BC}(f) = \sum_{i=1}^{n} w_i^{BC} f(x_i)$ with the weights $w^{BC} = K_X^{-1} k_{\mu,X} \in \mathbb{R}^n$.

The posterior variance $\sigma_X^2$ can be used to quantify epistemic uncertainty inherent to this approximation of $\int_{\Omega} f \, d\mu$. 
Use integral of the **GP posterior mean function** to approximate that of the **integrand function** $f$. 
Use integral of the GP posterior mean function to approximate that of the integrand function \( f \).
Use integral of the GP posterior mean function to approximate that of the integrand function $f$. 
Use integral of the **GP posterior mean function** to approximate that of the **integrand function** $f$. 
Bayesian cubature: illustration and interpolation

\[ f \mid \mathcal{D} \]

\[ \int_{\Omega} (f \mid \mathcal{D}) \, d\mu \]

Use integral of the \textbf{GP posterior mean function} to approximate that of the \textbf{integrand function} \( f \).

\[ x_1 \ x_2 \ x_3 \ x_4 \ x_5 \]
Bayesian cubature: illustration and interpolation

\[ f \mid D \]

\[ \int_{\Omega} (f \mid D) \, d\mu \]

Use integral of the **GP posterior mean function** to approximate that of the **integrand function** \( f \).
Reproducing kernel Hilbert spaces

Every positive-definite kernel $k$ defines a Hilbert space $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ of functions. The worst-case error $e_{\mathcal{H}_k}(Q)$ of a cubature rule $Q$ on $\mathcal{H}_k$ is

$$e_{\mathcal{H}_k}(Q) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \int_{\Omega} f \, d\mu - \sum_{i=1}^{n} w_i f(x_i) \right|.$$ 

Bayesian cubature rules are worst-case optimal in reproducing kernel Hilbert spaces:

$$w^{\text{BC}} = \arg \min_{w \in \mathbb{R}^n} e_{\mathcal{H}_k}(Q) = \arg \min_{w \in \mathbb{R}^n} \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left| \int_{\Omega} f \, d\mu - \sum_{i=1}^{n} w_i f(x_i) \right|.$$ 

Furthermore, the posterior std $\sigma_X$ coincides with the WCE:

$$\sigma_X = e_{\mathcal{H}_k}(Q^{\text{BC}}_X).$$
Table of contents

Bayesian cubature

Uncertainty quantification

Computational approaches
For uncertainty quantification to be meaningful, one needs to incorporate (i) prior information about the integrand and (ii) information obtained from the function evaluations.

This is done by selecting the covariance kernel $k$ and its hyperparameters:

1. **Smoothness** (i.e., number of derivatives) of the kernel tells how smooth the function is expected to be.
2. **Length-scale** $\ell > 0$ tells how fast the function is expected to vary.
3. **Amplitude** $\lambda > 0$ scales the variance.

Example: $k(x, x') = \lambda \exp \left( - \frac{\|x - x'\|^2}{2\ell^2} \right)$
A potential specification procedure:

1. **Smoothness** is difficult to extract from the data. Usually heuristically selected by the user.

2. **Length-scale** could be marginalised, but this is intractable. Maximal likelihood is typically used:

   \[
   \ell_{\text{ML}} = \arg\max_{\ell > 0} \left[ -\frac{1}{2} f_x^T K_{\ell, X}^{-1} f_X - \frac{1}{2} \log \det(K_{\ell, X}) - \frac{n}{2} \log(2\pi) \right].
   \]

3. **Amplitude** can be determined in closed form for maximal likelihood or marginalised using an improper prior \( p(\lambda) \propto \frac{1}{\lambda} \).

Selecting the length-scale parameter well is usually imperative and creates a computational bottleneck.
\[ f(x) = \exp \left( \sin(2x) - \frac{x^2}{5} \right) + \frac{x^2}{2} \] and \[ \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-x^2/2} \, dx \approx 2.0693 \]
Table of contents

Bayesian cubature

Uncertainty quantification

Computational approaches
Computational and practical issues

**Computational cost:** Recall that Bayesian cubature requires computing

\[ Q_{X}^{BC}(f) = \sum_{i=1}^{n} w_{i}^{BC} f(x_i) \quad \text{and} \quad \sigma_{X}^{2} = \int_{\Omega} k_{\mu} \, d\mu - \sum_{i=1}^{n} w_{i}^{BC} k_{\mu}(x_i), \]

where the weights \( w^{BC} \) are solved from

\[ K_{X} w^{BC} = k_{\mu,X} \quad \text{with} \quad [K_{X}]_{ij} = k(x_i, x_j), \quad [k_{\mu,X}]_{i} = k_{\mu}(x_i). \]

This introduces prohibitive cubic computational cost.

**Robustness and reliability:** Bayesian cubature can be inaccurate if the nodes do not cover the space well and the kernel hyperparameters (in particular, the length-scale) are not properly set.
Exploiting symmetry

If the measure $\mu$ and kernel $k$ share a symmetric structure, it makes sense to expect that weights assigned to points that form a symmetric set are equal.

**Result:** If $X$ is a union of $J$ symmetric sets, then there are only $J$ distinct weights that can be computed in time $O(Jn + J^3)$. Typically, $J \ll n$. 
Robustness against parameter misspecification

Bayesian cubature can be made more robust by inclusion of specific *exactness conditions* on a pre-determined finite function space.

Given a function family \( \pi = \{p_1, \ldots, p_Q\} \subset C(\Omega) \), the constraints \( Q_X^{BC}(p_i) = \int_\Omega p_i \, d\mu \) can be encoded in a probabilistic way that yields the weights

\[
\begin{bmatrix}
K_X & P_X \\
P_T & 0
\end{bmatrix}
\begin{bmatrix}
w^{BC} \\
0
\end{bmatrix} =
\begin{bmatrix}
k_{\mu,X} \\
p_{\mu,X}
\end{bmatrix}
\]

with \([P_X]_{ij} = p_i(x_j)\) and \([p_{\mu,X}]_i = \int_\Omega p_i \, d\mu\).

For example, a “hybrid” between standard Bayesian cubature and Monte Carlo is based on \( \pi = \{1\} \) and has

\[
\begin{bmatrix}
K_X & 1 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
w^{BC} \\
a
\end{bmatrix} =
\begin{bmatrix}
k_{\mu,X} \\
1
\end{bmatrix}.
\]
Standard GP interpolant and one with additional polynomial exactness constraints to an integrand function.
References

Foundational:


Computational approaches:


Thank you for your attention!