Theoretical Guarantees for Probabilistic Integration

François-Xavier Briol[†]

Collaborators: Chris. J. Oates[‡], Mark Girolami[†⋆], Michael A. Osborne[∗] and Dino Sejdinovic[⁕]

†University of Warwick, Statistics Department
‡U.T. Sydney, School of Mathematical and Physical Sciences
⋆The Alan Turing Institute for Data Science
∗University of Oxford, Department of Engineering Science
⁕University of Oxford, Statistics Department

Neural Information Processing Systems (NIPS) 2015
Workshop on Probabilistic Integration

The Problem

- Let f be continuous and square-integrable, let π be a probability density function and $\mathcal{X} \subseteq \mathbb{R}^d$ be our integration domain. We want to compute (numerically):

$$\Pi[f] = \int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x} \approx \sum_{i=1}^{n} w_i f(\mathbf{x}_i) = \hat{\Pi}[f] \tag{1}$$

- Main Problem: Most methods are very slow! (i.e. $\mathcal{O}_P(n^{-1/2})$).

The Problem

- Let f be continuous and square-integrable, let π be a probability density function and $\mathcal{X} \subseteq \mathbb{R}^d$ be our integration domain. We want to compute (numerically):

$$\Pi[f] = \int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x} \approx \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i) = \hat{\Pi}[f] \tag{1}$$

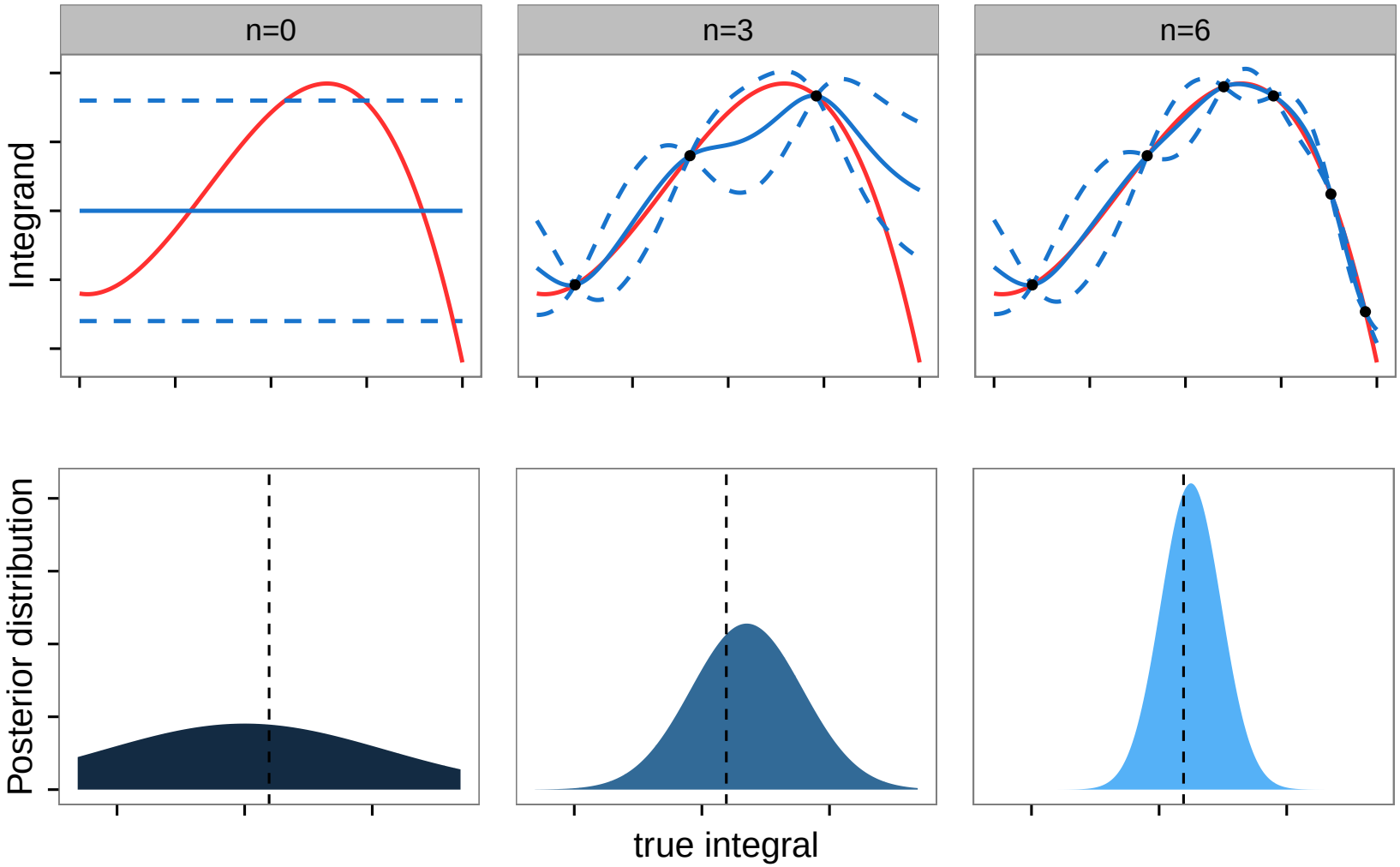
- Main Problem: Most methods are very slow! (i.e. $\mathcal{O}_P(n^{-1/2})$).
- Solution: Bayesian Quadrature (BQ) makes use of *prior information* about f to guide our choice of $\{\boldsymbol{x}_i, w_i\}_{i=1}^n$.
- This leads to another problem: Numerical results are good, but we would like some guarantees of faster convergence...

Why it matters in practice...

Let's consider the problem of Global Illumination in Computer Vision:

Each pixel requires you to compute three spherical integrals...
No time to evaluate the integrands on thousands of i.i.d. samples!
(see Marques et al., 2013)

Sketch of Bayesian Quadrature



Bayesian Quadrature

Let $\boldsymbol{z}_i = \mu_\pi(\boldsymbol{x}_i) = \int_{\mathcal{X}} k(\boldsymbol{x}_i, \boldsymbol{y})\pi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$. The posterior Gaussian distribution over the solution of the integral has the following mean and variance:

- **mean:** $\mathbb{E}\big[\Pi[f]\big] = \boldsymbol{z}^T \boldsymbol{K}^{-1} \boldsymbol{f} = \sum_{i=1}^{n} w_i^{\mathrm{BQ}} f(x_i) = \Pi_{\mathrm{BQ}}[f]$
- **variance:** $\mathbb{V}\big[\Pi[f]\big] = \Pi[\mu_\pi] - \boldsymbol{z}^T \boldsymbol{K}^{-1} \boldsymbol{z}$

How to get some theory...(I)

- In an RKHS the standard quantity to consider is the worst-case error:

$$\|\Pi - \hat{\Pi}_{\mathrm{BQ}}\|_{\mathrm{op}} = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \Pi[f] - \hat{\Pi}_{\mathrm{BQ}}[f] \right| \tag{2}$$

How to get some theory...(I)

- In an RKHS the standard quantity to consider is the worst-case error:

$$\|\Pi - \hat{\Pi}_{\mathrm{BQ}}\|_{\mathrm{op}} = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \Pi[f] - \hat{\Pi}_{\mathrm{BQ}}[f] \right| \tag{2}$$

- We call **convergence rate** the rate σ_n in the number of states n at which the worst-case error decreases.

How to get some theory...(I)

- In an RKHS the standard quantity to consider is the worst-case error:

$$\|\Pi - \hat{\Pi}_{\mathrm{BQ}}\|_{\mathrm{op}} = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \Pi[f] - \hat{\Pi}_{\mathrm{BQ}}[f] \right| \tag{2}$$

- We call **convergence rate** the rate σ_n in the number of states n at which the worst-case error decreases.

- Link between posterior variance and worst-case error:

$$\mathbb{V}\big[\Pi[f]\big] = \mathrm{MMD}^2(\pi, \hat{\pi}_{\mathrm{BQ}}) = \|\Pi - \hat{\Pi}_{\mathrm{BQ}}\|_{\mathrm{op}}^2 \tag{3}$$

- The BQ weights are optimal w.r.t. the worst-case error. (see Huszar & Duvenaud, 2012).

How to get some theory...(II)

Theorem (Bayesian re-weighting)

Consider $\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)$ *and* $\hat{\Pi}_{BQ}[f] = \sum_{i=1}^{n} w_i^{BQ} f(\boldsymbol{x}_i)$. *Suppose we have a convergence rate* σ_n *for* $\hat{\Pi}[f]$ *(i.e.* $\|\hat{\Pi} - \Pi\|_{op} \leq \sigma_n$*). Then:*

$$\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \sigma_n$$

How to get some theory...(II)

Theorem (Bayesian re-weighting)

Consider $\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)$ *and* $\hat{\Pi}_{BQ}[f] = \sum_{i=1}^{n} w_i^{BQ} f(\boldsymbol{x}_i)$. *Suppose we have a convergence rate* σ_n *for* $\hat{\Pi}[f]$ *(i.e.* $\|\hat{\Pi} - \Pi\|_{op} \leq \sigma_n$*). Then:*

$$\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \sigma_n$$

Theorem (Non-parametric Regression)

Fix states $X = \{\boldsymbol{x}_i\}_{i=1}^{n}$. *Then we have* $\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \|f - \mathbb{E}[f]\|_2$, *where* $\hat{\Pi}_{BQ}$ *is the BQ rule based on* X *and* $\mathbb{E}[f]$ *is our non-parametric approximation of* f.

How to get some theory...(II)

Theorem (Bayesian re-weighting)

Consider $\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(\boldsymbol{x}_i)$ *and* $\hat{\Pi}_{BQ}[f] = \sum_{i=1}^{n} w_i^{BQ} f(\boldsymbol{x}_i)$. *Suppose we have a convergence rate* σ_n *for* $\hat{\Pi}[f]$ *(i.e.* $\|\hat{\Pi} - \Pi\|_{op} \leq \sigma_n$*). Then:*

$$\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \sigma_n$$

Theorem (Non-parametric Regression)

Fix states $X = \{\boldsymbol{x}_i\}_{i=1}^{n}$. *Then we have* $\|\hat{\Pi}_{BQ} - \Pi\|_{op} \leq \|f - \mathbb{E}[f]\|_2$, *where* $\hat{\Pi}_{BQ}$ *is the BQ rule based on* X *and* $\mathbb{E}[f]$ *is our non-parametric approximation of* f.

We can also show rates of contraction!

Example 1: Frank-Wolfe Bayesian Quadrature

The algorithm:

- Use a convex optimization algorithm called the Frank-Wolfe algorithm to select design points $\left\{\boldsymbol{x}_i\right\}_{i=1}^{n}$. (see Bach, Lacoste-Julien & Obozinski, 2012)
- Weight the observation with the corresponding Bayesian Quadrature weights $\left\{\boldsymbol{w}_i^{\mathrm{BQ}}\right\}_{i=1}^{n}$.

Example 1: Frank-Wolfe Bayesian Quadrature

The algorithm:

- Use a convex optimization algorithm called the Frank-Wolfe algorithm to select design points $\{\boldsymbol{x}_i\}_{i=1}^n$. (see Bach, Lacoste-Julien & Obozinski, 2012)
- Weight the observation with the corresponding Bayesian Quadrature weights $\{\boldsymbol{w}_i^{\text{BQ}}\}_{i=1}^n$.

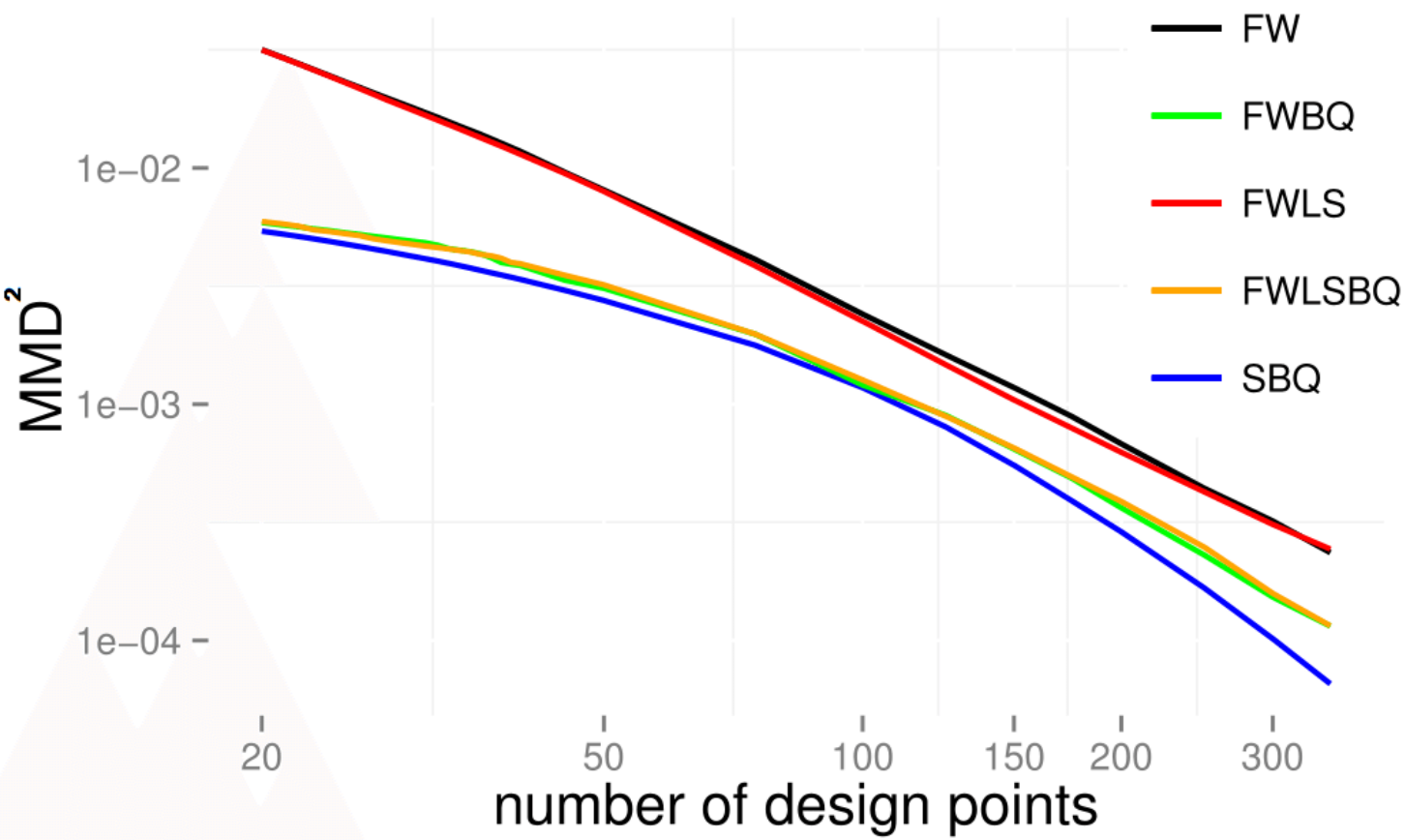
Theorem (Convergence/Consistency of FWBQ)

Suppose $\mathcal{H}$ is finite dim. Frank-Wolfe Bayesian Quadrature converges to the true integral $\int_{\mathcal{X}} f(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ at the following rates (for some $0 < C < \infty$):

$$\|\hat{\Pi}_{FWBQ} - \Pi\|_{op} \leq \begin{cases} \mathcal{O}_P(n^{-1}) & \text{for FWBQ, } \rho_i = 1/(i+1) \\ \mathcal{O}_P(\exp(-Cn)) & \text{for FWLSBQ} \end{cases}$$

This follows easily from the "Bayesian re-weighting" theorem!

FWBQ Simulations I



Example 2: Bayesian Monte Carlo & Quasi-Monte Carlo

- We can also obtain convergence rates for Bayesian Monte Carlo (BMC) and Bayesian Quasi-Monte Carlo (BQMC)!
- We consider Sobolev spaces $\mathcal{H}_\alpha$, which are functions of smoothness α. The RKHS obtained from Matérn kernels are Sobolev spaces.

Example 2: Bayesian Monte Carlo & Quasi-Monte Carlo

- We can also obtain convergence rates for Bayesian Monte Carlo (BMC) and Bayesian Quasi-Monte Carlo (BQMC)!
- We consider Sobolev spaces $\mathcal{H}_\alpha$, which are functions of smoothness α. The RKHS obtained from Matérn kernels are Sobolev spaces.

Theorem (Convergence/Consistency of BQMC in Sobolev spaces)

Let $\mathcal{X} = [0,1]^d$ and Π is Unif$(\mathcal{X})$. Consider $\hat{\Pi}_{BQMC}$ whose states $\{\boldsymbol{x}_i\}_{i=1}^n$ are a higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ net over $\mathbb{Z}_b$. Let $\mathcal{H}$ be an RKHS that is norm-equivalent to $\mathcal{H}_\alpha$. Then we have

$$\|\hat{\Pi}_{BQMC} - \Pi\|_{op} = \mathcal{O}(n^{-\alpha/d+\epsilon}), \tag{4}$$

where $\epsilon > 0$ can be arbitrarily small.

This follows from "Bayesian re-weighting" theorem! We know it is optimal for Sobolev Spaces!

Example 2: Bayesian Monte Carlo & Quasi-Monte Carlo

- i.i.d samples from π re-weighted using Bayesian Quadrature weights (introduced by Rasmussen & Ghahramani, 2003).

Theorem (Convergence/Consistency of BMC in Sobolev Spaces)

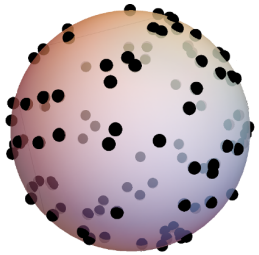
Suppose (i) $\mathcal{X}$ is bounded with Lipschitz boundary and satisfies an interior cone condition, (ii) π is bounded away from 0 and ∞ on $\mathcal{X}$, and (iii) $\mathcal{H}$ is norm-equivalent to $\mathcal{H}_\alpha$ where $\alpha > d/2$. Then

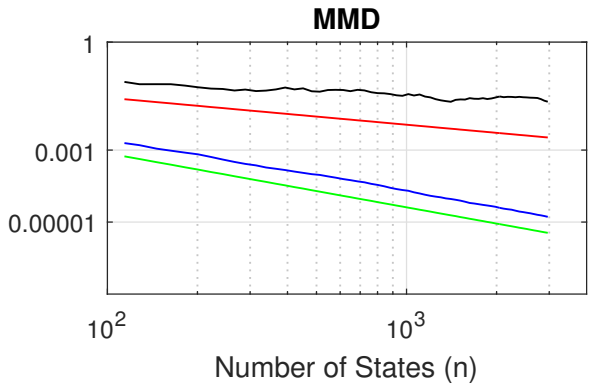
$$\|\hat{\Pi}_{BMC} - \Pi\|_{op} \quad = \quad \mathcal{O}_P(n^{-\alpha/d+\epsilon}) \tag{5}$$

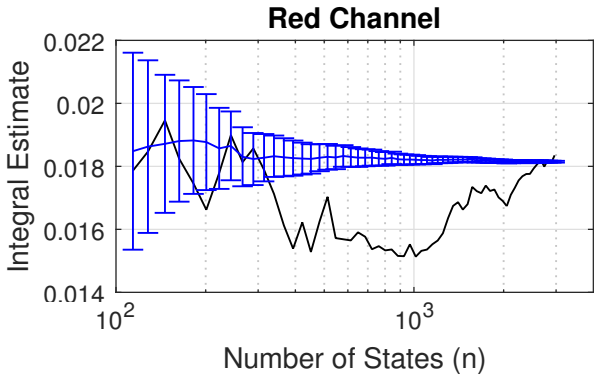
where $\epsilon > 0$ can be arbitrarily small.

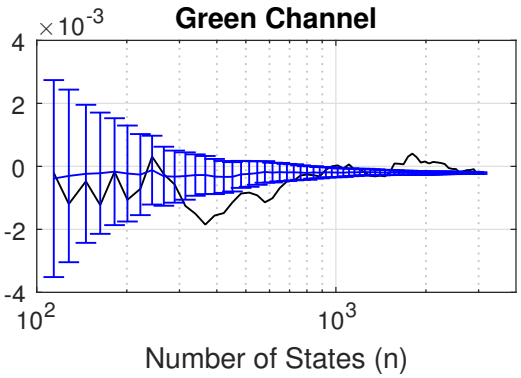
Follows from "non-parametric regression" theorem!

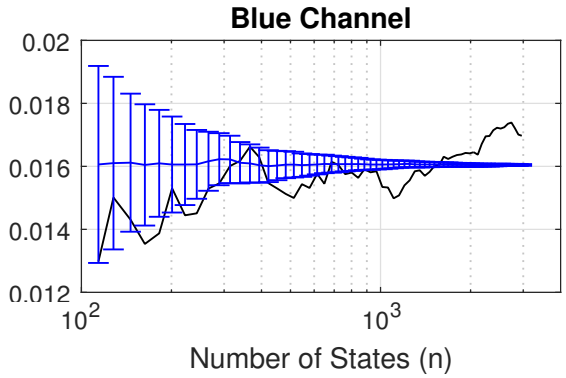
Back to our example: Global Illumination











We provide rates of $\mathcal{O}_P(n^{-\frac{3}{4}})$ which is optimal for $\mathcal{H}^{\frac{3}{2}}(\mathbb{S}^2)$!

References (I)


Briol, F-X., Oates, C. J., Girolami, M. & Osborne, M. A. (2015). *Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees*. In Advances in Neural Information Processing Systems (NIPS 2015).


Briol, F-X., Oates, C. J., Girolami, M., Osborne, M. A. & Sejdinovic, D. (2015). *Probabilistic Integration*. arXiv:1512.00933.

There's much more:

- Theory for noisy case (i.e. we observe $y_i = f(\boldsymbol{x}_i) + \epsilon_i$).
- Probabilistic Integration with intractable kernel mean μ_π using control functional kernels.
- Probabilistic Integration in high dimensions ($d = 50$) using weighted RKHS.

`www.warwick.ac.uk/fxbriol/probabilistic_integration/`

References (II)


Bach, F. (2015). *On the Equivalence between Quadrature Rules and Random Features.* arXiv:1502.06800.


F. Bach, S. Lacoste-Julien & G. Obozinski. On the Equivalence between Herding and Conditional Gradient Algorithms. Proceedings of the International Conference on Machine Learning (ICML), 2012.


Chen, Y., Welling, M., Smola, A.(2010) *Super-Samples from Kernel Herding.* In Proceedings of the Conference of Uncertainty in Artificial Intelligence (UAI), pages 109-116.

References (III)


Rasmussen, C. E. & Ghahramani, Z. (2003). Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems* (NIPS), pages 489-496.


Huszar, F. & Duvenaud, D. (2012). Optimally-weighted Herding is Bayesian Quadrature. In *Uncertainty in Artificial Intelligence* (UAI), pages 377-385.


Marques, R., Bouville, C., Ribardiere, M., Santos, P. & Bouatouch, K. (2013) *A Spherical Gaussian framework for Bayesian Monte Carlo Rendering of Glossy Surfaces.* IEEE Transactions on Visualization and Computer Graphics, 19(10):1619-1632.


Sriperumbudur, S., Gretton, A. Fukumizu, K., Schölkopf, B. & Lanckriet, G. *Hilbert Space Embeddings and Metrics on Probability Measures.* In Journal of Machine Learning Research, 11:1517-1561.