Monte Carlo methods for expensive models based on repulsive point processes

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France

Joint work with

Figure: Adrien Hardy, Ayoub Belhadji, Pierre Chainais

Numerical integration and DPPs

Tight interpolation rates in RKHSs

Numerical integration and DPPs

Tight interpolation rates in RKHSs

- The goal is to approximate

$$\int f \mathrm{d}\mu = \int f(x)\omega(x)\mathrm{d}x \approx \sum_{i=1}^{N} w_i f(x_i),$$

with nodes x_i and weights w_i.

Monte Carlo integration (importance sampling, MCMC, etc.)

- Choose the nodes randomly, and the weights $w_i = w(x_i, x_{-i})$.
- Typical error is

$$\forall f \in L^2, \quad \sqrt{\mathbb{E}\left[\int f \mathrm{d}\mu - \sum_{i=1}^{N} w_i f(x_i)\right]^2} \sim \frac{1}{\sqrt{N}}.$$

- The goal is to approximate

$$\int f \mathrm{d}\mu = \int f(x)\omega(x)\mathrm{d}x \approx \sum_{i=1}^{N} w_i f(x_i),$$

with nodes x_i and weights w_i.

Quasi Monte Carlo integration (Dick and Pilichshammer, 2010)

- Choose the nodes deterministically, and the weights $w_i = w(x_i, x_{-i})$.
- Typical error is

$$\sup_{f\in\mathcal{F}} \left| \int f \mathrm{d}\mu - \sum_{i=1}^{N} w_i f(x_i) \right| \sim \frac{\log^{s-1} N}{N}.$$

Plan

Numerical integration and DPPs

Tight interpolation rates in RKHSs

RKHSs are spaces of smooth functions with a kernel

- Consider the RKHS $\mathcal{F}$ with kernel κ, i.e. the completion of

$$\left\{ \sum_{i=1}^{M} \alpha_i \kappa(x_i, \cdot), M \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_M \in \mathbb{R}^d \right\}.$$

for the inner product defined by $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{F}} := \kappa(x, y)$.

- Under general assumptions, $\mathcal{F} \subset L^2(\mathrm{d}\mu)$, is dense, there is an ON basis (e_n) of $L^2(\mathrm{d}\mu)$ and $\sigma_n \to 0$ such that, pointwise,

$$\kappa(x, y) = \sum_{n \geqslant 1} \sigma_n e_n(x) e_n(y).$$

- In that case, $f \in \mathcal{F}$ if and only if $\sum_n \sigma_n^{-1} |\langle f, e_n \rangle|^2$ converges.

RKHSs are spaces of smooth functions with a kernel

- Consider the RKHS $\mathcal{F}$ with kernel κ, i.e. the completion of

$$\left\{ \sum_{i=1}^{M} \alpha_i \kappa(x_i, \cdot), M \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_M \in \mathbb{R}^d \right\}.$$

for the inner product defined by $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{F}} := \kappa(x, y)$.

- Under general assumptions, $\mathcal{F} \subset L^2(\mathrm{d}\mu)$, is dense, there is an ON basis (e_n) of $L^2(\mathrm{d}\mu)$ and $\sigma_n \to 0$ such that, pointwise,

$$\kappa(x, y) = \sum_{n \geqslant 1} \sigma_n e_n(x) e_n(y).$$

- In that case, $f \in \mathcal{F}$ if and only if $\sum_n \sigma_n^{-1} |\langle f, e_n \rangle|^2$ converges.

RKHSs are spaces of smooth functions with a kernel

- Consider the RKHS $\mathcal{F}$ with kernel κ, i.e. the completion of

$$\left\{ \sum_{i=1}^{M} \alpha_i \kappa(x_i, \cdot), M \in \mathbb{N}, \alpha_1, \ldots, \alpha_n \in \mathbb{R}, x_1, \ldots, x_M \in \mathbb{R}^d \right\}.$$

  for the inner product defined by $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{F}} := \kappa(x, y)$.

- Under general assumptions, $\mathcal{F} \subset L^2(\mathrm{d}\mu)$, is dense, there is an ON basis (e_n) of $L^2(\mathrm{d}\mu)$ and $\sigma_n \to 0$ such that, pointwise,

$$\kappa(x, y) = \sum_{n \geqslant 1} \sigma_n e_n(x) e_n(y).$$

- In that case, $f \in \mathcal{F}$ if and only if $\sum_n \sigma_n^{-1} |\langle f, e_n \rangle|^2$ converges.

Quadrature in an RKHS

- Let $f \in \mathcal{F}$, $g \in L^2(\mathrm{d}\mu)$ then

$$\left|\int fg\mathrm{d}\mu - \sum_{i=1}^{N} w_i f(x_i)\right| \leqslant \|f\|_{\mathcal{F}} \left\|\mu_g - \sum_{i=1}^{N} w_i \kappa(x_i, .)\right\|_{\mathcal{F}}, \qquad (1)$$

  where

$$\mu_g = \int g(x)\kappa(x, .)\mathrm{d}\mu(x)$$

  is the mean element of g.

- Once the nodes $x_1, \dots, x_N$ are known, minimizing the RHS of (1) in w boils down to inverting an $N \times N$ matrix.

- $\sum_{i=1}^{N} \hat{w}_i \kappa(x_i, \cdot)$ is the orthogonal projection of μ_g onto

$$S = \mathrm{Span}\left(\kappa(x_1, \cdot), \dots, \kappa(x_N, \cdot)\right).$$

- The squared volume in S of the parallelotope spanned by these functions is $\det[\kappa(x_i, x_j)]$.

- Let $f \in \mathcal{F}$, $g \in L^2(\mathrm{d}\mu)$ then

$$\left| \int fg \mathrm{d}\mu - \sum_{i=1}^{N} w_i f(x_i) \right| \leqslant \|f\|_{\mathcal{F}} \left\| \mu_g - \sum_{i=1}^{N} w_i \kappa(x_i, .) \right\|_{\mathcal{F}}, \qquad (1)$$

  where

$$\mu_g = \int g(x) \kappa(x, .) \mathrm{d}\mu(x)$$

  is the mean element of g.

- Once the nodes $x_1, \dots, x_N$ are known, minimizing the RHS of (1) in w boils down to inverting an $N \times N$ matrix.

- $\sum_{i=1}^{N} \hat{w}_i \kappa(x_i, \cdot)$ is the orthogonal projection of μ_g onto

$$S = \mathrm{Span}\left(\kappa(x_1, \cdot), \dots, \kappa(x_N, \cdot)\right).$$

- The squared volume in S of the parallelotope spanned by these functions is $\det[\kappa(x_i, x_j)]$.

Quadrature in an RKHS

- Let $f \in \mathcal{F}$, $g \in L^2(\mathrm{d}\mu)$ then

$$\left| \int fg\mathrm{d}\mu - \sum_{i=1}^{N} w_i f(x_i) \right| \leqslant \|f\|_{\mathcal{F}} \left\| \mu_g - \sum_{i=1}^{N} w_i \kappa(x_i, .) \right\|_{\mathcal{F}}, \qquad (1)$$

where

$$\mu_g = \int g(x)\kappa(x, .)\mathrm{d}\mu(x)$$

is the mean element of g.

- Once the nodes $x_1, \ldots, x_N$ are known, minimizing the RHS of (1) in w boils down to inverting an $N \times N$ matrix.
- $\sum_{i=1}^{N} \hat{w}_i \kappa(x_i, \cdot)$ is the orthogonal projection of μ_g onto

$$S = \mathrm{Span}\left(\kappa(x_1, \cdot), \ldots, \kappa(x_N, \cdot)\right).$$

- The squared volume in S of the parallelotope spanned by these functions is $\det[\kappa(x_i, x_j)]$.

- Let $f \in \mathcal{F}$, $g \in L^2(\mathrm{d}\mu)$ then

$$\left| \int fg \mathrm{d}\mu - \sum_{i=1}^{N} w_i f(x_i) \right| \leqslant \|f\|_{\mathcal{F}} \left\| \mu_g - \sum_{i=1}^{N} w_i \kappa(x_i, .) \right\|_{\mathcal{F}}, \qquad (1)$$

  where

$$\mu_g = \int g(x)\kappa(x, .)\mathrm{d}\mu(x)$$

  is the mean element of g.

- Once the nodes $x_1, \dots, x_N$ are known, minimizing the RHS of (1) in w boils down to inverting an $N \times N$ matrix.

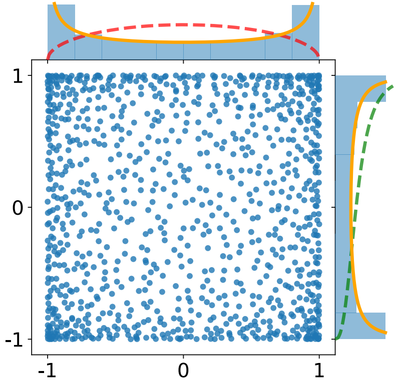
- $\sum_{i=1}^{N} \hat{w}_i \kappa(x_i, \cdot)$ is the orthogonal projection of μ_g onto

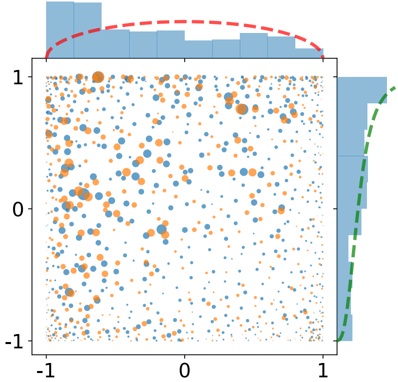
$$S = \mathrm{Span}\left(\kappa(x_1, \cdot), \dots, \kappa(x_N, \cdot)\right).$$

- The squared volume in S of the parallelotope spanned by these functions is $\det[\kappa(x_i, x_j)]$.

Volume sampling

- Let $x_1, \ldots, x_N \sim Z^{-1} \det[\kappa(x_i, x_j)] \, \mathrm{d}\mu(x_1) \ldots \mathrm{d}\mu(x_N)$
- Solve the linear program for the weights $w_1, \ldots, w_N$.





Volume sampling

- Let $x_1, \dots, x_N \sim Z^{-1} \det[\kappa(x_i, x_j)] \, \mathrm{d}\mu(x_1) \dots \mathrm{d}\mu(x_N)$
- Solve the linear program for the weights $w_1, \dots, w_N$.

Theorem (Belhadji, Bardenet, and Chainais, 2020)

Assume again $\sum_{n=1}^{N} |\langle g, e_n \rangle|^2 \leqslant 1$. *Then*

$$\mathbb{E} \left\| \mu_g - \sum_{i=1}^{N} w_i \kappa(x_i, \cdot) \right\|_{\mathcal{F}}^2 \leqslant \sigma_N \left(1 + \beta_N\right),$$

where $\beta_N = \min_{M \in [2:N]} [(N - M + 1)\sigma_N]^{-1} \sum_{m \geqslant M} \sigma_m$.

- It is known[1] that $\inf_{\substack{Y \subset \mathcal{F} \\ \dim Y = N}} \sup_{\|g\|_{\mathrm{d}\omega} \leqslant 1} \inf_{y \in Y} \|\mu_g - y\|_{\mathcal{F}}^2 = \sigma_{N+1}$.

[1] Pinkus, 2012.

Take-home message

- Without suprise, more smoothness means faster convergence.
- Random quadratures can be as fast as possible in RKHSs.

Many open problems

- Practical relevance of RKHS hypothesis in applications?
- What should $g \in L^2(\mu)$ be in $\int fg\,\mathrm{d}\mu$?
- Robustness to kernel choice.
- How do we efficiently sample from continuous volume sampling without spectral knowledge? See e.g. Rezaei and Gharan, 2019.
- Can faster-than-Monte Carlo error come with limited smoothness? Yes, see e.g. Bardenet and Hardy, 2020.

References I

Bardenet, R. and A. Hardy (2020). "Monte Carlo with Determinantal Point Processes". In: *Annals of Applied Probability.*

Belhadji, A., R. Bardenet, and P. Chainais (2020). "Kernel interpolation with continuous volume sampling". In: *International Conference on Machine Learning (ICML).*

Dick, J. and F. Pilichshammer (2010). *Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press.

Pinkus, A. (2012). *N-widths in Approximation Theory*. Vol. 7. Springer Science & Business Media.

Rezaei, A. and S. O. Gharan (2019). "A Polynomial Time MCMC Method for Sampling from Continuous Determinantal Point Processes". In: *International Conference on Machine Learning*, pp. 5438–5447.