Positively Weighted Kernel Quadrature via Subsampling

Satoshi Hayakawa, Harald Oberhauser, Terry Lyons
Mathematical Institute, University of Oxford

NeurIPS 2022

Numerical quadrature

♣ For a probability distribution μ over a space $\mathcal{X}$

$$\sum_{i=1}^{n} w_i f(x_i) \approx \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x)$$

is called a quadrature rule,
where $w_i \in \mathbb{R}$ are weights and $x_i \in \mathcal{X}$ are sample points

Roughly speaking, this research is about...

When an integrand f is in a space so-called RKHS,

- can we find a configuration of x_i with small integral error?
- what is the convergence guarantee then?

♣ Kernel quadrature

Let k be a positive definite kernel and μ be a Borel probability measure on $\mathcal{X}$

For a quadrature rule Q_n for μ

$$Q_n(f) = \sum_{i=1}^{n} w_i f(x_i) \left(\approx \int_{\mathcal{X}} f(x) \, \mathrm{d}\mu(x) =: \mu(f) \right),$$

the worst-case error is defined as

$$\mathrm{wce}(Q_n) := \sup_{\|f\|_{\mathcal{H}} \le 1} |Q_n(f) - \mu(f)|,$$

where $\mathcal{H}$ is the RKHS associated with k

▷ We want to minimize this $\mathrm{wce}(Q_n)$

Why kernel quadrature?

♣ What is the benefit of kernel quadrature?

- Includes classical examples such as Sobolev spaces
- Can compute the worst-case error (theoretical guarantee!)

$$\begin{aligned} \mathrm{wce}(Q_n)^2 &= \sup_{\|f\|\le 1} \left\langle f,\ \boldsymbol{w}^\top k(\cdot, X) - k_\mu \right\rangle_{\mathcal{H}}^2 \\ &= \boldsymbol{w}^\top k(X,X)\boldsymbol{w} - 2\boldsymbol{w}^\top k_\mu(X) + \mu(k_\mu) \end{aligned}$$

where $\boldsymbol{w} = (w_i)_{i=1}^n$, $X = (x_i)_{i=1}^n$, $k_\mu := \int_{\mathcal{X}} k(\cdot, x)\,\mathrm{d}\mu(x)$
... if k_μ is known, we can convex-optimize the weights

- Application to GP regression / Bayesian quadrature
  "Fast Bayesian Inference with Batch Bayesian Quadrature via Kernel Recombination" [Adachi et al., 2022]

Mercer decomposition

♣ Consider the spectral decomposition of an integral operator

$$\mathcal{K} : L^2(\mu) \to L^2(\mu); \quad f \mapsto \int_{\mathcal{X}} k(\cdot, x) f(x) \, \mathrm{d}\mu(x),$$

we have the following Mercer decomposition under a mild condition:

$$k(x, y) = \sum_{m=1}^{\infty} \sigma_m e_m(x) e_m(y),$$

where

- (σ_m, e_m) are eigenpairs of $\mathcal{K}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$
- $(e_m)_{m=1}^{\infty} \subset L^2(\mu)$, $(\sqrt{\sigma_m} e_m)_{m=1}^{\infty} \subset \mathcal{H}$ are orthonormal
- σ_n typically decays polynopmially (Sobolev kernel) or exponentially (Gaussian kernel)

▷ Empirically, we approximately have $\min \mathrm{wce}(Q_n)^2 \sim \sigma_n$

Kernel quadrature: a table

♣ Comparison with other methods

	Method	Bound of squared wce	Computational complexity	C	M	E
gradient method	Herding [10, 4]	$1/n$	$n \cdot (n$: global optimization$)$	✓	✓	
	SBQ [25]	Not found	$n \cdot (n^2$: global optimization$)$		✓	
sampling method	Leveraged [3]	$\sigma_m, m = \mathcal{O}(n \log n)$	Unavailable			
	DPP [7, 6]	r_{n+1}	$n^3 \cdot$ (rejection sampling)			
	CVS [8]	σ_{n+1}	Unavailable		✓	
sub-sampling a larger sample	KT++ [14, 15, 56]	$(1/n^2 + 1/N)\,\mathrm{polylog}(N)$	$N \log^3 N$	✓	✓	✓
	Ours:					
	Mercer†	r_n	$nN_\varphi + C(n, N_\varphi)$	✓		
	M. + empirical‡	$r_n + \frac{1}{N}$	$nN + n^3 \log(N/n)$	✓		✓
	Nyström†	$n\sigma_n + r_{n+1} + \frac{n}{\sqrt{\ell}}$	$n\ell N_\varphi + n\ell^2 + C(n, N_\varphi)$	✓	✓	
	N. + empirical‡	$n\sigma_n + r_{n+1} + \frac{n}{\sqrt{\ell}} + \frac{1}{N}$	$n\ell N + n\ell^2 + n^3 \log(N/n)$	✓	✓	✓

- C = convex, M= *not* using Mercer decomposition, E = *not* using expectations
- $r_n = \sum_{m=n}^{\infty} \sigma_m$

References: **Herding** [Chen et al., 2010; Bach et al., 2012], **SBQ** [Huszár and Duvenaud, 2012], **Leveraged** [Bach, 2017], **DPP** [Belhadji et al., 2019; Belhadji, 2021], **CVS** [Belhadji et al., 2020], **KT++** [Dwivedi and Mackey, 2021, 2022; Shetty et al., 2022]

Main theoretical result

We call Q_n a convex quadrature when $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$
Let $k_0 = \sum_{i=1}^{n-1} \varphi_i(x)\varphi_j(y)$ be another kernel with $k_1 := k - k_0$ being positive definite ($k_{1,\mathrm{diag}}(x) := k_1(x,x)$)

Theorem (Theorem 1 in the paper)

For an empirical measure $\tilde{\mu}_N = \frac{1}{N}\sum_{j=1}^N \delta_{y_j}$ with $y_j \sim_{iid} \mu$ we can construct an n-point convex quadrature Q_n with

$$Q_n(\varphi_i) = \tilde{\mu}_N(\varphi_i), \qquad Q_n(k_{1,\mathrm{diag}}) \leq \tilde{\mu}_N(k_{1,\mathrm{diag}})$$

in $\mathcal{O}\big(nN + n^3 \log(N/n)\big)$ computational steps
Resulting Q_n satisfies

$$\mathbb{E}\left[\mathrm{wce}(Q_n)^2\right] \leq 8 \int_{\mathcal{X}} k_1(x,x)\,\mathrm{d}\mu(x) + \frac{1}{N}\int_{\mathcal{X}} k(x,x)\,\mathrm{d}\mu(x)$$

Remarks

♣ We should use a big $N \gg n$ to obtain a good quadrature

♣ Reduction of a big discrete measure is known as recombination [Litterer and Lyons, 2012; Tchernychova, 2015]

♣ If we know the Mercer decomposition, we can use $k_0(x, y) = \sum_{m=1}^{n-1} \sigma_m e_m(x) e_m(y)$, then the guarantee becomes

$$\mathrm{wce}(Q_n)^2 = \mathcal{O}\left(\sum_{m=n}^{\infty} \sigma_m + \frac{1}{N}\right)$$

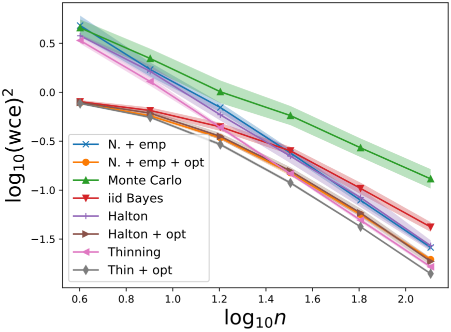
(Corollary 2 in the paper)

♣ Otherwise, we can use the Nyström approximation: theoretical bound (Theorem 3, Corollary 4 in the paper) is loose but empirically performs very well

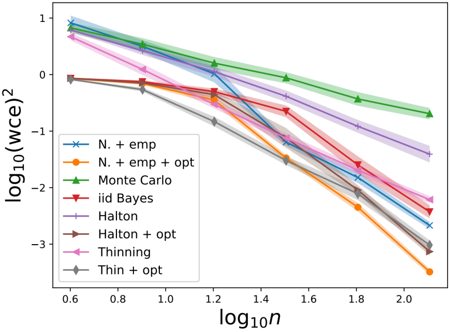
Numerical experiments

♣ Periodic sobolev spaces (d: dimension, r: smoothness)

- Used $N = n^2$ for subsampling-based methods
- $d \geq 2$: product RKHS, no known optimal
- '+ opt' is additionally optimizing the weights $\boldsymbol{w}$



(c) $d = 2, r = 1$

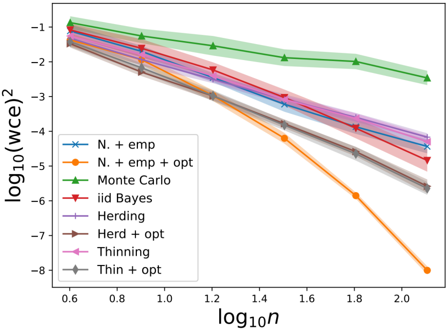


(d) $d = 3, r = 3$

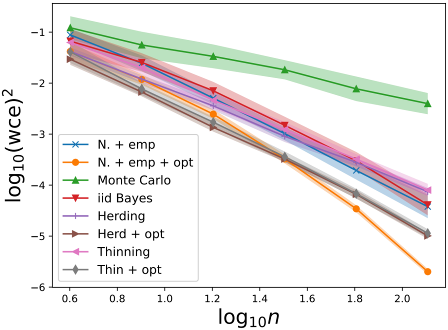
Numerical experiments

♣ From UCI Machine Learning Repository, we used '3D Road Network' and 'Combined Cycle Power Plant'

- Regarded the data as an equal-weight discrete measure
- Gaussian kernel with median heuristics



(a) 3D Road Network data



(b) Power Plant data

Conclusion

♣ Summary

- Kernel quadrature: a numerical integration in RKHS
- We have given a practical algorithm for constructing kernel quadrature with theoretical guarantee
  - Outperforms others by exploiting spectral decay

♣ Future work

- Can we improve guarantees for the Nyström version?
- Why does '+opt' work so well with our methods?

References

Adachi, M., Hayakawa, S., Jørgensen, M., Oberhauser, H., and Osborne, M. A. (2022). Fast Bayesian inference with batch Bayesian quadrature via kernel recombination. In *Advances in Neural Information Processing Systems*.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751.

Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012). On the equivalence between herding and conditional gradient algorithms. In *International Conference on Machine Learning*, pages 1355–1362.

Belhadji, A. (2021). An analysis of Ermakov–Zolotukhin quadrature using kernels. In *Advances in Neural Information Processing Systems*, volume 34.

Belhadji, A., Bardenet, R., and Chainais, P. (2019). Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems*, volume 32, pages 12907–12917.

Belhadji, A., Bardenet, R., and Chainais, P. (2020). Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning*, pages 725–735. PMLR.

Chen, Y., Welling, M., and Smola, A. (2010). Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence*, pages 109–116.

Dwivedi, R. and Mackey, L. (2021). Kernel thinning. In *Conference on Learning Theory*, pages 1753–1753. PMLR.

Dwivedi, R. and Mackey, L. (2022). Generalized kernel thinning. In *International Conference on Learning Representations*.

Huszár, F. and Duvenaud, D. (2012). Optimally-weighted herding is Bayesian quadrature. In *Conference on Uncertainty in Artificial Intelligence*, pages 377–386.

Litterer, C. and Lyons, T. (2012). High order recombination and an application to cubature on Wiener space. *The Annals of Applied Probability*, 22(4):1301–1327.

Shetty, A., Dwivedi, R., and Mackey, L. (2022). Distribution compression in near-linear time. In *International Conference on Learning Representations*.

Tchernychova, M. (2015). *Carathéodory cubature measures*. PhD thesis, University of Oxford.